Contents

Original Papers

Developing a Framework to Infer Opioid Use Disorder Severity From Clinical Notes to Inform Natural Language Processing Methods: Characterization Study (e53366)  
Melissa Poulsen, Philip Freda, Vanessa Troiani, Danielle Mowery. 4

Feasibility, Acceptability, and Potential Efficacy of a Self-Guided Internet-Delivered Dialectical Behavior Therapy Intervention for Substance Use Disorders: Randomized Controlled Trial (e50399)  
Alexander Daros, Timothy Guimond, Christina Yager, Emma Palermo, Chelsey Wilks, Lena Quilty. 21

Identification of Predictors of Mood Disorder Misdiagnosis and Subsequent Help-Seeking Behavior in Individuals With Depressive Symptoms: Gradient-Boosted Tree Machine Learning Approach (e50738)  
Jiri Benacek, Nimotalai Lawal, Tommy Ong, Jakub Tomasik, Nayra Martin-Key, Erin Funnell, Giles Barton-Owen, Tony Olmert, Dan Cowell, Sabine Bahn. 41

Speech Features as Predictors of Momentary Depression Severity in Patients With Depressive Disorder Undergoing Sleep Deprivation Therapy: Ambulatory Assessment Pilot Study (e49222)  
Lisa-Marie Wadle, Ulrich Ebner-Priemer, Jerome Foo, Yoshiharu Yamamoto, Fabian Streit, Stephanie Witt, Josef Frank, Lea Zillich, Matthias Limberger, Ayimnisagul Ablimit, Tanja Schultz, Maria Gilles, Marcella Rietschel, Lea Sirignano. 54

eHealth in the Management of Depressive Episodes in Catalonia's Primary Care From 2017 to 2022: Retrospective Observational Study (e52816)  
Aina Fuster-Casanovas, Queralt Miró Catalina, Josep Vidal-Alaball, Anna Esclafé-Besa, Carme Carrión. 68

Incorporating a Stepped Care Approach Into Internet-Based Cognitive Behavioral Therapy for Depression: Randomized Controlled Trial (e51704)  
Jasleen Jagayat, Anchan Kumar, Yiia Shao, Amrita Pannu, Charmy Patel, Amirhossein Shirazi, Mohsen Omrani, Nazanin Alavi. 81

HealthySMS Text Messaging System Adjunct to Adolescent Group Cognitive Behavioral Therapy in the Context of COVID-19 (Let's Text!): Pilot Feasibility and Acceptability Study (e49317)  
Lauren Haack, Courtney Armstrong, Kate Travis, Adrian Aguilera, Sabrina Darrow. 106

The Implementation of Recommender Systems for Mental Health Recovery Narratives: Evaluation of Use and Performance (e45754)  
Emily Slade, Stefan Rennick-Egglesstone, Fiona Ng, Yasuhiro Kotera, Joy Llewellyn-Beardsley, Chris Newby, Tony Glover, Jeroen Keppens, Mike Slade. 149

A Comparison of ChatGPT and Fine-Tuned Open Pre-Trained Transformers (OPT) Against Widely Used Sentiment Analysis Tools: Sentiment Analysis of COVID-19 Survey Data (e50150)  
Juan Lossio-Ventura, Rachel Weger, Angela Lee, Emily Guinee, Joyce Chung, Lauren Atlas, Eleni Linos, Francisco Pereira. 175
Capacity of Generative AI to Interpret Human Emotions From Visual and Textual Data: Pilot Evaluation Study (e54369)
Zohar Elyoseph, Elad Refoua, Kfir Asraf, Maya Lvovsky, Yoav Shimoni, Dorit Hadar-Shoval. .......................................................... 476

Reviews
Translating Suicide Safety Planning Components Into the Design of mHealth App Features: Systematic Review (e52763)
Kim Gryglewicz, Victoria Orr, Marissa McNeil, Lindsay Taliaferro, Serenea Hines, Taylor Duffy, Pamela Wisniewski. .................................................. 122

Health Care Professionals’ Views on the Use of Passive Sensing, AI, and Machine Learning in Mental Health Care: Systematic Review With Meta-Synthesis (e49577)
Jessica Rogan, Sandra Bucci, Joseph Firth. .................................................................................................................. 226

Effectiveness of Online and Remote Interventions for Mental Health in Children, Adolescents, and Young Adults After the Onset of the COVID-19 Pandemic: Systematic Review and Meta-Analysis (e46637)
Linda Fischer-Grote, Vera Fössing, Martin Aigner, Elisabeth Fehrmann, Markus Boeckle. ................................................................. 242

Effectiveness of Virtual Reality–Based Well-Being Interventions for Stress Reduction in Young Adults: Systematic Review (e52186)
Joy Xu, Areej Khanotia, Shmuel Juni, Josephine Ku, Hana Sami, Vallen Lin, Roberta Walterson, Evelyn Payne, Helen Jo, Parmin Rahimpoor-Marani. ................................................................. 264

Effectiveness and User Experience of Virtual Reality for Social Anxiety Disorder: Systematic Review (e48916)
Simon Shahid, Joshua Kelson, Anthony Saliba. ................................................................................................................. 385

Research Letter
The Frequency of Design Studies Targeting People With Psychotic Symptoms and Features in Mental Health Care Innovation: Secondary Analysis of a Systematic Review (e54202)
Lars Veldmeijer, Gijs Terlouw, Jim Van Os, Job Van ’t Veer, Nynke Boonstra. ................................................................. 458
Developing a Framework to Infer Opioid Use Disorder Severity From Clinical Notes to Inform Natural Language Processing Methods: Characterization Study

Melissa N Poulsen¹*, MPH, PhD; Philip J Freda²*, PhD; Vanessa Troiani³, PhD; Danielle L Mowery⁴, PhD

¹Department of Population Health Sciences, Geisinger, Danville, PA, United States
²Department of Computational Biomedicine, Cedars-Sinai Medical Center, West Hollywood, CA, United States
³Department of Autism and Developmental Medicine, Geisinger, Danville, PA, United States
⁴Department of Biostatistics, Epidemiology and Informatics, Institute of Biomedical Informatics, University of Pennsylvania, Philadelphia, PA, United States

*these authors contributed equally

Abstract

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

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

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

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

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

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

https://mental.jmir.org/2024/1/e53366
Furthermore, most algorithms have been developed in the challenging to isolate within a patient’s medical record [14,15]. Thus, OUD can be disjointed care across hospitals, lack of specialty diagnostic missing from EHRs due to a variety of factors, including charts. Information about substance use disorders may be limited when diagnostic codes are minimally applied to patient characterization of problematic opioid use [12], and overdose opioid use detection [9,10], identification of OUD [11], prescription opioid misuse [7], OUD prediction [8], nonmedical opioid use, including identification of patients at risk for misuse or chronic pain. Patients with chronic pain with opioid prescriptions are at increased risk for opioid misuse [16,17], but only focusing on such populations may exclude patient populations with illicit opioid use or otherwise outside of chronic pain treatment. Previous research indicates clinical notes provide a source of rich information that could improve efforts to identify and characterize OUD [12,18,19].

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

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

- How accurately can OUD severity be inferred from text from a small number of targeted clinical notes per patient?
- Where and how do clinical teams document OUD-related information, in terms of clinical note types and encounter settings?
• How is OUD-related information documented over time relative to an opioid- or OUD-specific health care encounter?
• What is the frequency of OUD-related concepts and their co-occurrence in clinical notes?

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

**Methods**

**Study Design**

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

**Ethical Considerations**

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

**Study Population**

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

<table>
<thead>
<tr>
<th>Study Group</th>
<th>Inclusion Criteria</th>
<th>Exclusion Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CP-RX</strong></td>
<td>Chronic pain: chronic pain was defined as having at least two opioid analgesic prescriptions for nonprogressive musculoskeletal pain.</td>
<td>Non-European ancestry: individuals with non-European ancestry were excluded because this study's sample was originally assembled for a genetic study.</td>
</tr>
<tr>
<td></td>
<td>Opioid use disorder (OUD; mild, moderate, or severe) confirmed through chart review.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Opioid use began with opioid analgesic prescription. Based on self-report in a survey question.</td>
<td></td>
</tr>
<tr>
<td><strong>CP-nonRX</strong></td>
<td>Chronic pain.</td>
<td>Non-European ancestry: individuals with non-European ancestry were excluded because this study's sample was originally assembled for a genetic study.</td>
</tr>
<tr>
<td></td>
<td>OUD (mild, moderate, or severe) confirmed through chart review.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Opioid use did not begin with opioid analgesic prescription. Based on self-report in a survey question.</td>
<td></td>
</tr>
<tr>
<td><strong>OUD-DX</strong></td>
<td>At least one diagnosis code for OUD. International Classification of Disease codes used to define OUD were based on Jennings et al [29].</td>
<td>Chronic pain.</td>
</tr>
<tr>
<td></td>
<td>Order for medications for OUD including buprenorphine, buprenorphine-naloxone, and naltrexone. Geisinger providers did not prescribe methadone for OUD treatment.</td>
<td></td>
</tr>
<tr>
<td><strong>OUD-TX</strong></td>
<td>At least one diagnosis code for OUD.</td>
<td>Chronic pain.</td>
</tr>
<tr>
<td></td>
<td>Order for medications for OUD.</td>
<td></td>
</tr>
<tr>
<td><strong>Control</strong></td>
<td>In total, 1 opioid analgesic order.</td>
<td>Chronic pain.</td>
</tr>
<tr>
<td></td>
<td>Diagnosis code for OUD.</td>
<td>Order for medications for OUD.</td>
</tr>
</tbody>
</table>
Data Collection
We obtained notes from inpatient, outpatient, and emergency department (ED) encounters. In total, 11 note types were obtained, selected based on clinician input: outpatient clinic notes, progress notes, ancillary progress notes, history and progress (H&P) notes, discharge summaries, ED notes, ED provider notes, ED triage notes, ED support staff notes, communication notes, and lactation notes. We obtained notes for 3 encounter dates per patient: an index date representing either the first observed OUD diagnosis (for OUD-DX and OUD-TX groups and some patients in the CP-RX and CP non-RX groups) or the most recent opioid analgesic order (for some patients in the CP-RX and CP non-RX groups and all patients in the control group) and the encounters immediately prior to and following the index date. Multiple notes per patient were obtained for some encounter dates. Notes were deidentified using Philter [30].

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

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

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

Results
Annotation of Clinical Notes
The annotation schema contained 27 classes, 12 of which included attributes (Figure 1; Multimedia Appendix 2). Interannotator agreement (IAA) for the 2 reviewers ranged from 20% to 100% across the 15 batches of reviewed notes but was above 70% for all except 4 batches (Figure 2). Lower IAA occurred with less frequently annotated classes and for classes with an attribute denoting a historic concept (eg, psychiatric condition current vs historic; IAA results not shown by class).
From the 100 sampled patients and 320 associated notes, we annotated 1436 sentences within 186 notes from 82 patients over 15 batches. The remaining notes did not yield any annotations (ie, they did not contain text relevant to the classification schema). Most patients without annotations were controls (11/18, 61%) and most had an index encounter based on an opioid analgesic order (16/18, 89%). They also had fewer notes available (mean 1.8, SD 1.1) and no notes were from the inpatient setting.

Among the 4 noncontrol groups, 73 patients had annotations (Table 1). OUD-TX was the only group in which all 20 sampled patients had at least one note with text relevant for annotation and the group accounted for the largest proportion of annotations.
Can Severity of Problematic Opioid Use Be Inferred From a Limited Number of Clinical Notes?

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

Table 2. Average OUD\(^b\) scores among 82 patients with annotated sentences.

<table>
<thead>
<tr>
<th>Study group or index encounter reason(^a)</th>
<th>Mean (SD) OUD score</th>
<th>Count of patients by OUD severity</th>
<th>PPVs(^c) for moderate or severe OUD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>None, n (%)</td>
<td>Mild, n (%)</td>
<td>Moderate, n (%)</td>
</tr>
<tr>
<td>CP-RX</td>
<td>3.4 (3.4)</td>
<td>8 (50)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>CP-nonRX</td>
<td>4.5 (3.9)</td>
<td>7 (39)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>OUD-DX</td>
<td>5.2 (2.5)</td>
<td>3 (16)</td>
<td>1 (5)</td>
</tr>
<tr>
<td>OUD-TX</td>
<td>6.8 (2.2)</td>
<td>1 (5)</td>
<td>1 (5)</td>
</tr>
<tr>
<td>Control</td>
<td>0.0 (0.0)</td>
<td>9 (100)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>OUD diagnosis</td>
<td>6.0 (2.7)</td>
<td>8 (13)</td>
<td>2 (3)</td>
</tr>
<tr>
<td>Opioid analgesic order</td>
<td>0.6 (1.7)</td>
<td>11 (92)</td>
<td>0 (0)</td>
</tr>
</tbody>
</table>

\(^a\)OUD: opioid use disorder.
\(^b\)Reason for selecting index encounter, either an OUD diagnosis or an order for an opioid analgesic.
\(^c\)PPV: positive predictive value.
\(^d\)N/A: not applicable.

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

The mean number of notes per patient that were reviewed and annotated differed slightly by degree of severity (Table 3), but these differences were not statistically significant (1-way ANOVA for mean number of notes reviewed: $F_{1,38}=0.60; P=.62$; for mean number of notes annotated: $F_{1,36}=1.46; P=.24$). We observed no consistent patterns in the frequency of note types or encounter types by severity.

History of substance misuse was among the most prevalent classes across severity groups and was present in 56-126 (17%-38%) of the 332 notes (Table 3). For those with “no OUD,” the classes psychiatric condition and daily tobacco use were also common. The class OUD treatment automatically led to a classification of “severe OUD” and was present in 36 (16%) of the 226 notes among patients scored as “severe OUD.”
Table 3. Note characteristics by OUD severity among 100 sampled study patients with 332 notes.

<table>
<thead>
<tr>
<th>Note characteristics</th>
<th>Not annotated</th>
<th>OUD severity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>None</td>
</tr>
<tr>
<td><strong>Patients, n</strong></td>
<td>18</td>
<td>28</td>
</tr>
<tr>
<td><strong>Notes reviewed, n</strong></td>
<td>32</td>
<td>98</td>
</tr>
<tr>
<td><strong>Number of notes reviewed per patient, mean (SD)</strong></td>
<td>1.8 (1.2)</td>
<td>3.3 (2.3)</td>
</tr>
<tr>
<td><strong>Total number of notes annotated, n</strong></td>
<td>N/A&lt;sup&gt;b&lt;/sup&gt;</td>
<td>90</td>
</tr>
<tr>
<td><strong>Number of notes with annotations per patient, mean (SD)</strong></td>
<td>N/A</td>
<td>2.3 (2.0)</td>
</tr>
<tr>
<td><strong>Note type, n (%)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Progress notes</td>
<td>17 (53)</td>
<td>24 (24)</td>
</tr>
<tr>
<td>H&amp;P&lt;sup&gt;c&lt;/sup&gt; notes</td>
<td>0 (0)</td>
<td>21 (21)</td>
</tr>
<tr>
<td>ED&lt;sup&gt;d&lt;/sup&gt; notes</td>
<td>4 (13)</td>
<td>9 (9)</td>
</tr>
<tr>
<td>ED provider notes</td>
<td>3 (9)</td>
<td>22 (22)</td>
</tr>
<tr>
<td>Discharge summaries</td>
<td>0 (0)</td>
<td>16 (16)</td>
</tr>
<tr>
<td>ED triage notes</td>
<td>2 (6)</td>
<td>1 (1)</td>
</tr>
<tr>
<td>Ancillary progress notes</td>
<td>0 (0)</td>
<td>1 (1)</td>
</tr>
<tr>
<td>Communication notes</td>
<td>0 (0)</td>
<td>1 (1)</td>
</tr>
<tr>
<td>OPT&lt;sup&gt;e&lt;/sup&gt; clinic notes</td>
<td>1 (3)</td>
<td>2 (1)</td>
</tr>
<tr>
<td>ED support staff notes</td>
<td>5 (16)</td>
<td>1 (1)</td>
</tr>
<tr>
<td>Lactation notes</td>
<td>0 (0)</td>
<td>0 (0)</td>
</tr>
<tr>
<td><strong>Encounter type, n (%)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OPT</td>
<td>18 (56)</td>
<td>25 (26)</td>
</tr>
<tr>
<td>ED</td>
<td>14 (44)</td>
<td>33 (34)</td>
</tr>
<tr>
<td>ED to IPT&lt;sup&gt;f&lt;/sup&gt;</td>
<td>0 (0)</td>
<td>21 (21)</td>
</tr>
<tr>
<td>IPT</td>
<td>0 (0)</td>
<td>18 (18)</td>
</tr>
<tr>
<td><strong>Most frequent classes (percentage of notes with class)</strong>&lt;sup&gt;a&lt;/sup&gt;</td>
<td>N/A</td>
<td>History of substance misuse (38%)</td>
</tr>
<tr>
<td></td>
<td>N/A</td>
<td>Psychiatric condition (31%)</td>
</tr>
</tbody>
</table>

<sup>a</sup>OUD: opioid use disorder.
<sup>b</sup>N/A: not applicable.
<sup>c</sup>H&P: history and progress.
<sup>d</sup>ED: emergency department.
<sup>e</sup>OPT: outpatient.
<sup>f</sup>IPT: inpatient.
Where Is OUD-Relevant Information Found in Clinical Notes and How Is it Documented Over Time?

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

**Figure 3.** Heatmap of class frequencies by note type. ED: emergency department; H&P: history and progress; OUD: opioid use disorder.
<table>
<thead>
<tr>
<th>Note type</th>
<th>Number of notes</th>
<th>Number of notes with annotated sentences, n (%)</th>
<th>Counts of notes with annotated sentences by encounter setting (percentage of total)</th>
<th>Number of annotated sentences, n (%)</th>
<th>Mean (SD) number of annotated sentences per note</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>ED(^a), n (%)</td>
<td>ED to IPT(^b), n (%)</td>
<td>OPT(^c), n (%)</td>
</tr>
<tr>
<td>Progress notes</td>
<td>92</td>
<td>62 (67)</td>
<td>1 (2)</td>
<td>7 (11)</td>
<td>3 (5)</td>
</tr>
<tr>
<td>H&amp;P(^d) notes</td>
<td>30</td>
<td>27 (90)</td>
<td>8 (30)</td>
<td>13 (48)</td>
<td>6 (22)</td>
</tr>
<tr>
<td>ED notes</td>
<td>32</td>
<td>25 (78)</td>
<td>21 (84)</td>
<td>4 (16)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>ED provider notes</td>
<td>28</td>
<td>23 (82)</td>
<td>18 (78)</td>
<td>5 (22)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Discharge summaries</td>
<td>26</td>
<td>23 (88)</td>
<td>5 (22)</td>
<td>10 (43)</td>
<td>8 (35)</td>
</tr>
<tr>
<td>ED triage notes</td>
<td>26</td>
<td>9 (35)</td>
<td>9 (100)</td>
<td>0 (0)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Ancillary progress notes</td>
<td>30</td>
<td>9 (30)</td>
<td>2 (22)</td>
<td>4 (44)</td>
<td>3 (33)</td>
</tr>
<tr>
<td>Communication notes</td>
<td>12</td>
<td>3 (25)</td>
<td>0 (0)</td>
<td>2 (67)</td>
<td>1 (33)</td>
</tr>
<tr>
<td>OPT clinic notes</td>
<td>4</td>
<td>2 (50)</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>ED support staff notes</td>
<td>38</td>
<td>2 (5)</td>
<td>1 (50)</td>
<td>1 (50)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Lactation notes</td>
<td>2</td>
<td>1 (50)</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>1 (100)</td>
</tr>
</tbody>
</table>

\(^a\)ED: emergency department.  
\(^b\)IPT: inpatient.  
\(^c\)OPT: outpatient.  
\(^d\)H&P: history and progress.  
\(^e\)N/A: not applicable.

Regarding encounter type, compared to the inpatient setting, notes from ED and outpatient settings had the highest proportions of notes with annotations and a high diversity of classes represented (Figure 4).
Some classes appeared across varied note types, such as *OUD treatment*, *daily tobacco use*, *history of substance misuse*, and *psychiatric condition*, whereas others tended to only appear in a particular note type (eg, *OUD, naloxone, opioid craving, and overdose history* in progress notes; *drug seeking* in ED notes).

The largest proportion of annotations was observed at the index encounter for all classes except *intoxication* (Figure 5). The classes *opioid craving*, *opioid tolerance*, *polysubstance misuse*, and *vocational consequences* were only observed at the index encounter. Annotations were more common in “new” versus “historic” encounters, and several classes were only observed in “new” encounters and not “historic” encounters.
Which OUD-Related Concepts Are Most Common in Clinical Notes?

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

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

Several class pairs had highly correlated frequency distributions across note types, indicating similar documentation frequency (Figure 6). *Psychiatric condition, history of substance misuse, opioid misuse-illicit*, and *opioid misuse-prescription* were highly correlated with many other classes. Conversely, classes with the lowest correlations in their frequency distribution with other classes included *opioid tolerance*, *vocational consequences*, and *polysubstance misuse*. 
When evaluated across patients, the classes *psychiatric condition* and *history of substance misuse* also had high frequency of co-occurrence with other classes, as did *other illicit drug use* (Figure 6). Classes with the lowest frequency of co-occurrence with other classes included *opioid tolerance*, *opioid craving*, *vocational consequences*, and *OUD*.

**Discussion**

**Principal Results**

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

To our knowledge, this is the first study to develop an annotation schema to characterize OUD severity. Most patients received a severity score indicating severe OUD or no OUD. The paucity of scores indicating mild or moderate OUD may be explained by our approach to mapping classes to DSM-5 criteria, the selection of study cohorts, and the limited number of notes reviewed. The “crosswalk” of classes and DSM-5 criteria was based on prior work that included clinician input [11,27], but down weighting specific classes could be justified in future uses and would yield lower severity scores. Additionally, the resulting PPVs for moderate or severe OUD aligned with expectations for this study’s cohorts. Control group individuals all had a severity score of 0, evidence of the specificity of our approach. The OUD-TX group’s PPV of 0.90 is consistent with successful EHR-based algorithms for other conditions [32].

This was expected since these individuals received medication treatment for OUD, which, if documented in the notes we reviewed, would yield a score indicative of severe OUD. The OUD-DX group also had a high PPV, which is unsurprising given their OUD diagnosis. OUD is often underdiagnosed [14,15]; thus, the existence of an OUD diagnosis would be expected to indicate a true disorder, with signs and symptoms likely documented. That said, prior studies have found International Classification of Diseases codes are insufficient for accurately identifying OUD [11,12,14], which may explain why the PPV for OUD-DX was somewhat lower than the OUD-TX group. Both chronic pain groups had moderate PPVs and individuals in these groups were more likely to have a score indicating no OUD. This is likely explained by the index encounter reason—half of the CP-RX group and nearly a quarter of the CP-nonRX group were selected based on an opioid analgesic order. Most individuals whose index encounter was based on an opioid analgesic order had a score indicative of “no OUD.” Although these 2 cohorts had OUD confirmed through a chart review process in prior studies [27,28], the confirmation was only for mild OUD. It is also probable that given the limited number of notes reviewed for this study, additional notes may have yielded information indicative of OUD, which may have occurred later than the 3 selected encounter dates. Based on these findings, we believe that using a limited number of notes is likely sufficient for characterizing OUD severity only in the presence of other confirming information, such as an OUD diagnosis or a medication order for OUD treatment.

Documentation Patterns of Information Relevant to OUD in Clinical Notes

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

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

Absence of certain concepts is also informative for NLP tool development. Consistent with previous work [18,27], some classes relevant to OUD severity were rarely observed, including those representing consequences of opioid misuse and lack of control of opioid use. Reviewing a larger set of notes in an expanded patient sample could yield more frequent documentation of some classes, but some classes may represent concepts not traditionally recorded by clinicians (eg, vocational consequences). Documentation by clinicians of additional concepts such as the vocational, social, legal, and medical consequences patients face due to opioid use could be useful for patient care, particularly given the potential utility of such information in characterizing OUD severity. Future work to develop NLP frameworks related to OUD should consider that information captured in clinical notes may change over time with secular changes that occur in the drug landscape, clinical practices and documentation, and patients’ care-seeking behaviors.
Limitations
This study used clinical notes from an integrated health system, allowing for review of multiple note types across encounter settings. Findings may be different in health systems in which information is not integrated across settings; for example, opioid analgesic orders may be missing and OUD treatment siloed. Although we evaluated cohorts with varying evidence of opioid misuse, inclusion of other patient cohorts may yield different findings. In particular, because the cohorts were originally assembled for a genomic study, the CP-RX and CP-nonRX study groups included individuals only of European ancestry, limiting the generalizability of findings to other racial groups. Replication studies are necessary to understand whether study findings generalize to other settings and patient populations.

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

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

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

Conflicts of Interest
None declared.

Multimedia Appendix 1
Crosswalk of class or attribute combinations and Diagnostic and Statistical Manual of Mental Disorders, 5th edition (DSM-5) criteria for opioid use disorder (OUD).
[DOCX File, 18 KB - mental_v11i1e53366_app1.docx ]

Multimedia Appendix 2
Classes with brief definitions, example annotated sentences, and the count of annotated sentences by class or attribute.
[DOCX File, 19 KB - mental_v11i1e53366_app2.docx ]

References


Abbreviations

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

©Melissa N Poulsen, Philip J Freda, Vanessa Troiani, Danielle L Mowery. Originally published in JMIR Mental Health (https://mental.jmir.org), 15.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Mental Health, is properly cited. The complete bibliographic information, a link to the original publication on https://mental.jmir.org/, as well as this copyright and license information must be included.
Feasibility, Acceptability, and Potential Efficacy of a Self-Guided Internet-Delivered Dialectical Behavior Therapy Intervention for Substance Use Disorders: Randomized Controlled Trial

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

\(^1\)Campbell Family Mental Health Research Institute, Centre for Addiction and Mental Health, Toronto, ON, Canada
\(^2\)Addictions Division, Centre for Addiction and Mental Health, Toronto, ON, Canada
\(^3\)Department of Psychiatry, University of Toronto, Toronto, ON, Canada
\(^4\)Penn Center for Mental Health, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, United States
\(^5\)Department of Psychological Science, University of Missouri-St Louis, St. Louis, MO, United States

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

Abstract

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

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

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

Results: iDBT was perceived as a credible and acceptable treatment. In terms of feasibility, 94% (68/72) of the participants started iDBT, 13% (9/68) were early dropouts, 35% (24/68) used it for the recommended 8 days in the first month, and 50% (34/68) were still active 4 weeks later. On average, the participants used iDBT for 2 hours and 24 minutes across 10 separate days. In the acute period, no greater benefit was found for the immediate group on substance dependence, although we did find lower depression (b=-2.46; \(P=.02\)) and anxiety (b=-2.22; \(P=.02\)). At follow-up, there were greater benefits in terms of reduced alcohol (b=-2.00; \(P=.02\)) and nonalcoholic substance (b=-3.74; \(P=.01\)) consumption in the immediate access group. Both groups demonstrated improvements in substance dependence in the acute (b=-1.73; \(P<.001\)) and follow-up period (b=-2.09; \(P<.001\)). At follow-up, both groups reported reduced depression, anxiety, suicidal behaviors, emotional dysregulation, and functional disability.
Conclusions: iDBT is a feasible and acceptable intervention for patients with SUDs, although methods for improving engagement are warranted. Although results did not support efficacy for the primary outcome at 4 weeks, findings support reductions in substance dependence and other mental health concerns at 12 weeks. Notwithstanding the limitations of this study, the results suggest the potential value of iDBT in the treatment of SUDs and other mental health conditions.

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

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

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

Introduction

Background

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

Psychological Treatments for SUDs

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

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

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

Internet-Delivered DBT

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

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

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

Current Study

This study aims to evaluate version 2.0 of Pocket Skills and advance the literature in several ways. First, the current investigation evaluates Pocket Skills 2.0, which includes some of the content from version 1.0, as well as revised and novel materials that have not yet been evaluated. Second, the delivery of iDBT in this study was predominantly self-guided, with limited therapist guidance compared with the previous trial that used iDBT intervention as a therapeutic adjunct [34]. Third, this investigation represented a more controlled study of Pocket Skills 2.0 as a stand-alone treatment in a sample of treatment-seeking adults with SUDs who were not receiving any other forms of psychological treatments. Finally, this investigation randomized participants to immediate versus delayed access to advance the previous single-arm study. A 12-week single-blinded parallel-arm waitlist-controlled RCT was initiated, with participants randomized to receive immediate access to the intervention or delayed access after 4 weeks. The 4-week intervention and follow-up periods are in line with previous implementations of self-guided digital mental health interventions [35-37]. These studies have found that attrition rates start to increase steadily after 4 weeks and especially after 7 to 8 weeks (eg, ×50%). Specifically, we hypothesized that greater than 50% of participants would start the intervention (H1a); not drop out early (H1b); engage with the intervention at a recommended dose of twice a week (or 8 d) in the first 4 weeks (H1c); and would still be using the intervention after 4 weeks (H1d). We also hypothesized that participants would rate the intervention as credible and acceptable on established measures (H1e).

Second, we hypothesized that (H2a) participants in the immediate versus delayed iDBT group would show significantly greater improvements in our primary outcome of substance use dependence at the acute (week 4) and follow-up periods (week 12) in the form of an interaction effect (group×time). In addition, we hypothesized (H2b) significantly greater improvements for the immediate versus delayed iDBT group for our secondary outcomes (ie, depression, anxiety, emotion dysregulation, suicidality, functional disability, dispositional mindfulness, DBT skills, risky behaviors, and frequency of alcohol and substance use). Third, we hypothesized that iDBT would (H3) produce significant main effect improvements in both groups in the acute (week 4) and follow-up phases (week 12) of the intervention for all outcome measures.

Methods

Study Design

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

Ethical Considerations

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

Participants and Recruitment

Enrollment ran from August 2022 to March 2023, and all follow-ups were completed by June 2023. Participants were recruited from psychiatric hospital clinician referrals, waitlists, and research registries and from the surrounding community through several methods of advertisement (eg, hospital and other websites, social media posts, private DBT clinics, and local community organizations). All advertisements sought individuals who wanted to reduce their alcohol or substance use and specifically stated that they would be offered an internet-delivered intervention. All prospective participants were initially informed about the study and were prescreened for eligibility over the phone.
Inclusion criteria were as follows: (1) aged 18 to 65 years; (2) fluent in English; (3) understanding and willingness to comply with study requirements; (4) referred to addictions programming at our hospital or seeking treatment from the community, but not currently receiving any CBT or DBT intervention (support groups and psychiatric services were allowed); (5) alcohol or SUD in the past year; (6) use of alcohol or substance in the past month; (7) access to the internet (assumed literacy); and (8) at least contemplation levels of wanting to reduce alcohol or substance use on the Contemplation Ladder measure [38]. Exclusion criteria included (1) any known practical factors that would preclude participation, (2) acute psychiatric (ie, suicidality, psychotic disorder) or medical condition (ie, acute intoxication or withdrawal) requiring medical attention, and (3) participation in another psychological intervention or treatment study. We did not exclude participants based on whether they were taking psychotropic medications or not.

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

Randomization and Blinding
Participants were randomly assigned to an immediate access group or a delayed access group. The randomization procedure was blinded to the participants and the experimenter who ran all baseline sessions (ARD). Thus, neither party knew which group the participant was randomized into (electronically), and baseline procedures were completed. None of the participants withdrew immediately following randomization. The experimenters were not blinded to the procedures following the baseline session, including the follow-up assessments and contact. All follow-up assessments were conducted remotely and consisted solely of self-report measures.

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

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

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

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

\(^{a}\) TIP: temperature, intense exercise, and paced breathing.  
\(^{b}\) ACCEPTS: activities, contributing, comparisons, emotions, pushing away, thoughts, and sensations.  
\(^{c}\) DEARMAN: describe, express, assert, reinforce, be mindful, appear confident, and negotiate.  
\(^{d}\) GIVE: be gentle, act interested, validate, and use an easy manner.  
\(^{e}\) FAST: be fair, no apologies, stick to values, and be truthful.

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

Diagnostic Interviews

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

Feasibility and Credibility Measures

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

Acceptability and Engagement Measures

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

Primary Outcome

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

Secondary Outcomes

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

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

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

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

The difficulties in emotion regulation scale, short form [49], is a 16-item self-report measure with excellent internal consistency, assessing emotion dysregulation based on a 6-facet model first described by Gratza and Roemer [50]. Higher scores suggest greater emotion dysregulation difficulties. The ω reliability coefficient was 0.83.

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

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

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

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

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

Statistical Analysis

Overview

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

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

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

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

Each primary and secondary outcome variable was assessed with models containing an interaction effect (group×time, as factor variables) and main effects only (group+time, as factor variables) along with a continuous covariate controlling for the baseline assessment of each outcome per person. The final model chosen for interpretation was the better fitting model based on lower Akaike information criterion and Bayes information criterion values. Therefore, if the model fit was improved by the inclusion of the interaction term, we report that model; otherwise, we removed the interaction term and report the model with the main effects only. All model comparisons were evaluated using maximum likelihood estimation with the lmerTest (version 3.1-3 [59]) package, which uses the Satterthwaite df method. To further reduce the number of statistical tests reported, we also opted to interpret only the week 4 and week 12 contrasts against baseline as these were the most pertinent time points to address our hypotheses. Our α significance level was P=0.05, and all statistical tests were 2 tailed. The outputs of each final statistical model are provided in Multimedia Appendix 2 for full transparency.

Power

To achieve at least a medium effect size reduction in substance dependence, as suggested by Wilks et al [30], we would require a minimum sample of 60 as per G*Power (version 3.1.9.7; Cohen f=0.15; 2 groups; 4 measurements over 12 wk; power=0.95; α=0.05; and correlation between measures of at least 0.70) [60]. With an expected attrition rate of approximately 20%, we aimed to recruit approximately 72 to 75 individuals in total.
Results

Hypothesis 1: Feasibility and Acceptability

Participant Enrollment and Demographic Characteristics

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

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

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Total ITT&lt;sup&gt;a&lt;/sup&gt; (N=72)</th>
<th>Immediate iDBT&lt;sup&gt;b&lt;/sup&gt; (n=38)</th>
<th>Delayed iDBT&lt;sup&gt;b&lt;/sup&gt; (n=34)</th>
<th>Group comparison statistical value</th>
<th>Group comparison P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (y), mean (SD)</td>
<td>34.1 (11.9)</td>
<td>33.4 (10.5)</td>
<td>34.8 (13.3)</td>
<td>t&lt;sub&gt;10&lt;/sub&gt;=0.50</td>
<td>.62</td>
</tr>
<tr>
<td><strong>Sex, n (%)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>47 (65)</td>
<td>24 (63)</td>
<td>23 (68)</td>
<td>χ²=0.2</td>
<td>.69</td>
</tr>
<tr>
<td>Male</td>
<td>25 (35)</td>
<td>14 (37)</td>
<td>11 (32)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Gender, n (%)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Woman</td>
<td>43 (60)</td>
<td>22 (58)</td>
<td>21 (62)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Man</td>
<td>24 (33)</td>
<td>13 (34)</td>
<td>11 (32)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other (nonbinary, transgender, gender-fluid, or other)</td>
<td>6 (8)</td>
<td>4 (10)</td>
<td>2 (6)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Sexual orientation, n (%)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Heterosexual</td>
<td>44 (61)</td>
<td>22 (58)</td>
<td>22 (65)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lesbian or gay</td>
<td>5 (7)</td>
<td>3 (8)</td>
<td>2 (6)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bisexual</td>
<td>11 (15)</td>
<td>5 (13)</td>
<td>6 (18)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other (pansexual, queer, asexual, questioning or not sure, or prefer not to answer)</td>
<td>12 (17)</td>
<td>8 (21)</td>
<td>4 (12)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Race or ethnicity, n (%)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Black (African, North American, and Caribbean)</td>
<td>7 (10)</td>
<td>2 (5)</td>
<td>5 (15)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>East or Southeast Asian</td>
<td>5 (7)</td>
<td>4 (10)</td>
<td>1 (3)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Latin American</td>
<td>5 (7)</td>
<td>3 (8)</td>
<td>2 (6)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>South Asian</td>
<td>9 (12)</td>
<td>5 (13)</td>
<td>4 (12)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>48 (67)</td>
<td>25 (66)</td>
<td>23 (68)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other (First Nations, Middle Eastern, mixed, or not listed)</td>
<td>8 (11)</td>
<td>6 (16)</td>
<td>2 (6)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Marital status, n (%)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Single</td>
<td>39 (54)</td>
<td>25 (66)</td>
<td>14 (41)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dating</td>
<td>18 (25)</td>
<td>8 (21)</td>
<td>10 (29)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Married</td>
<td>10 (14)</td>
<td>3 (8)</td>
<td>7 (21)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other (divorced, widowed, or separated)</td>
<td>5 (7)</td>
<td>2 (5)</td>
<td>3 (9)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Employment status, n (%)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Full-time employed</td>
<td>29 (40)</td>
<td>17 (45)</td>
<td>12 (35)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Part-time employed</td>
<td>19 (26)</td>
<td>10 (26)</td>
<td>9 (27)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unemployed</td>
<td>14 (19)</td>
<td>6 (16)</td>
<td>8 (23)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>On disability</td>
<td>7 (10)</td>
<td>3 (78)</td>
<td>4 (12)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prefer not to say</td>
<td>3 (4)</td>
<td>2 (5)</td>
<td>1 (3)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Current conditions, n (%)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Major depressive disorder</td>
<td>37 (51)</td>
<td>20 (53)</td>
<td>17 (50)</td>
<td>Fisher exact test</td>
<td>.99</td>
</tr>
<tr>
<td>Persistent depressive disorder</td>
<td>18 (25)</td>
<td>10 (26)</td>
<td>8 (23)</td>
<td>Fisher exact test</td>
<td>.99</td>
</tr>
<tr>
<td>Characteristics</td>
<td>Total ITT(^a) (N=72)</td>
<td>Immediate iDBT(^b) (n=38)</td>
<td>Delayed iDBT (n=34)</td>
<td>Group comparison statistical value</td>
<td>Group comparison (P) value</td>
</tr>
<tr>
<td>---------------------------------------------</td>
<td>------------------------</td>
<td>-----------------------------</td>
<td>---------------------</td>
<td>-----------------------------------</td>
<td>-----------------------------</td>
</tr>
<tr>
<td>Bipolar I or II disorder</td>
<td>6 (8)</td>
<td>4 (10)</td>
<td>2 (6)</td>
<td>Fisher exact test</td>
<td>.68</td>
</tr>
<tr>
<td>Generalized anxiety disorder</td>
<td>37 (51)</td>
<td>21 (55)</td>
<td>16 (47)</td>
<td>Fisher exact test</td>
<td>.64</td>
</tr>
<tr>
<td>Social anxiety disorder</td>
<td>22 (31)</td>
<td>10 (26)</td>
<td>12 (35)</td>
<td>Fisher exact test</td>
<td>.45</td>
</tr>
<tr>
<td>Posttraumatic stress disorder</td>
<td>18 (25)</td>
<td>10 (26)</td>
<td>8 (23)</td>
<td>Fisher exact test</td>
<td>.99</td>
</tr>
<tr>
<td>Other anxiety disorder</td>
<td>11 (15)</td>
<td>8 (21)</td>
<td>3 (9)</td>
<td>Fisher exact test</td>
<td>.20</td>
</tr>
<tr>
<td>Alcohol use disorder</td>
<td>47 (65)</td>
<td>23 (60)</td>
<td>24 (71)</td>
<td>Fisher exact test</td>
<td>.46</td>
</tr>
<tr>
<td>Depression use disorder</td>
<td>40 (56)</td>
<td>24 (63)</td>
<td>16 (47)</td>
<td>Fisher exact test</td>
<td>.24</td>
</tr>
<tr>
<td>Cannabis use disorder</td>
<td>24 (33)</td>
<td>12 (32)</td>
<td>12 (35)</td>
<td>Fisher exact test</td>
<td>.81</td>
</tr>
<tr>
<td>Nicotine use disorder</td>
<td>15 (21)</td>
<td>9 (24)</td>
<td>6 (18)</td>
<td>Fisher exact test</td>
<td>.57</td>
</tr>
<tr>
<td>Stimulant use disorder</td>
<td>9 (12)</td>
<td>7 (18)</td>
<td>2 (6)</td>
<td>Fisher exact test</td>
<td>.16</td>
</tr>
<tr>
<td>SH or A(^e) use disorder</td>
<td>1 (1)</td>
<td>1 (3)</td>
<td>0 (0)</td>
<td>Fisher exact test</td>
<td>.99</td>
</tr>
</tbody>
</table>

\(^a\)ITT: intent-to-treat (ie, completed baseline procedures and randomized to condition).
\(^b\)iDBT: internet-delivered dialectical behavioral therapy.
\(^c\)Some cells are empty because we report the group comparison statistic for the overall category above.
\(^d\)Participants could select multiple options.
\(^e\)SH or A: sedative, hypnotic, or anxiolytic.

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

Moreover, 94% (68/72) of the randomized participants started the intervention. Three participants in the delayed iDBT group did not attend the follow-up session and never connected to the intervention following attempts to reschedule and instructions provided via email. One additional participant withdrew from the delayed group owing to technical issues and being unable to sign in and therefore did not access the intervention. One participant asked to withdraw from the immediate group because another person. Participants also estimated a mean 58% (SD 21%; range 10%-90%) improvement in symptoms.

Acceptability and Engagement

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

Table 3. Pocket Skills engagement metrics across groups.

<table>
<thead>
<tr>
<th>Overall engagement</th>
<th>Immediate iDBT&lt;sup&gt;a&lt;/sup&gt; (n=38), mean (SD; range)</th>
<th>Delayed iDBT (n=30), mean (SD; range)</th>
<th>Continued intervention after the first day (n=59), mean (SD; range)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total time</td>
<td>2 h 24 min (2 h 45 min; 0 min to 16 h 30 min)</td>
<td>2 h 49 min (3 h 6 min; 0 min to 16 h 30 min)</td>
<td>1 h 51 min (2 h 12 min; 0 min to 8 h 56 min)</td>
</tr>
<tr>
<td>Interactions</td>
<td>542.78 (569.46; 0-2830)</td>
<td>627.47 (576.22; 0-2468)</td>
<td>429.17 (560.28; 2-2830)</td>
</tr>
<tr>
<td>Unique days</td>
<td>10.24 (10.81; 1-64)</td>
<td>11.66 (12.68; 1-64)</td>
<td>8.45 (7.80; 1-36)</td>
</tr>
<tr>
<td>Days spread</td>
<td>43.41 (32.65; 1-141)</td>
<td>50.95 (37.64; 1-141)</td>
<td>33.69 (22.42; 1-64)</td>
</tr>
<tr>
<td>Days in the first 4 wk</td>
<td>6.69 (5.51; 1-29)</td>
<td>7.08 (5.88; 1-29)</td>
<td>6.17 (5.15; 1-21)</td>
</tr>
<tr>
<td>Time in the first 4 wk</td>
<td>1 h 54 min (2 h 17 min; 0 min to 13 h 10 min)</td>
<td>2 h 9 min (2 h 30 min; 0 min to 13 h 10 min)</td>
<td>1 h 32 min (2 h 0 min; 1 min to 8 h 30 min)</td>
</tr>
<tr>
<td>Web interactions</td>
<td>422.82 (474.87; 0-2566)</td>
<td>463.74 (455.99; 0-2143)</td>
<td>362.83 (507.54; 2-2566)</td>
</tr>
</tbody>
</table>

<sup>a</sup>iDBT: internet-delivered dialectical behavioral therapy.

Module engagement

<table>
<thead>
<tr>
<th>General</th>
<th>38 min 19 s (1 h 21 min 35 s; 0 to 10 h 43 min)</th>
<th>48 min 50 s (1 h 43 min 5 s; 0 to 10 h 36 min)</th>
<th>25 min 24 s (39 min 11 s; 0 min 52 s to 3 h 26 min)</th>
<th>43 min 38 s (1 h 26 min 25 s; 2 min 25 s to 3 h 26 min)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mindfulness</td>
<td>46 min 17 s (41 min 43 s; 0 to 2 h 44 min)</td>
<td>53 min 46 s (43 min 35 s; 0 to 2 h 42 min)</td>
<td>36 min 58 s (38 min 29 s; 0 to 2 h 44 min)</td>
<td>53 min 21 s (40 min 21 s; 0 to 2 h 44 min)</td>
</tr>
<tr>
<td>Distress tolerance</td>
<td>11 min 0 s (20 min 38 s; 0 to 1 h 34 min)</td>
<td>13 min 06 s (20 min 54 s; 0 to 1 h 13 min)</td>
<td>8 min 34 s (20 min 40 s; 0 to 1 h 34 min)</td>
<td>12 min 41 s (21 min 41 s; 0 to 1 h 34 min)</td>
</tr>
<tr>
<td>Emotion regulation</td>
<td>32 min 38 s (46 min 56 s; 0 to 2 h 23 min)</td>
<td>36 min 31 s (51 min 31 s; 0 to 2 h 23 min)</td>
<td>28 min 39 s (41 min 8 s; 0 to 2 h 10 min)</td>
<td>37 min 37 s (48 min 31 s; 0 to 2 h 23 min)</td>
</tr>
<tr>
<td>Interpersonal effectiveness</td>
<td>6 min 35 s (12 min 53 s; 0 to 48 min 39 s)</td>
<td>8 min 05 s (14 min 24 s; 0-48 min 39 s)</td>
<td>3 min 49 s (9 min 41 s; 0 to 34 min 33 s)</td>
<td>7 min 35 s (13 min 34 s; 0 to 48 min 39 s)</td>
</tr>
<tr>
<td>Addiction</td>
<td>9 min 18 s (19 min 3 s; 0 to 1 h 19 min)</td>
<td>8 min 52 s (17 min 42 s; 0 to 1 h 12 min)</td>
<td>8 min 19 s (19 min 31 s; 0 to 1 h 19 min)</td>
<td>10 min 43 s (20 min 6 s; 1 h to 19 min)</td>
</tr>
</tbody>
</table>
Hypothesis 2: Were Improvements Greater in the Immediate Versus Delayed iDBT Group?

Overview

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

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

<table>
<thead>
<tr>
<th>Group and time</th>
<th>Continuous outcomesa</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SDSb, PHQ-9c, GAD-7d, SBQe, DERS-16f, MAASg, WHODASHb, DBT-WCCLi</td>
</tr>
<tr>
<td>Immediate iDBTj (n=38)</td>
<td></td>
</tr>
<tr>
<td>Week 0</td>
<td></td>
</tr>
<tr>
<td>Values, n (%)</td>
<td>38 (100) 38 (100) 38 (100) 38 (100) 38 (100) 38 (100) 38 (100) 38 (100)</td>
</tr>
<tr>
<td>Values, mean (SD)</td>
<td>8.7 (3.5) 12.9 (5.9) 11.9 (5.4) 8.6 (3.6) 51.8 (13.4) 3.6 (0.9) 16.5 (9.0) 1.8 (0.4)</td>
</tr>
<tr>
<td>Week 4</td>
<td></td>
</tr>
<tr>
<td>Values, n (%)</td>
<td>34 (89) 34 (89) 34 (89) 34 (89) 33 (87) 31 (82) 31 (82) 31 (82)</td>
</tr>
<tr>
<td>Values, mean (SD)</td>
<td>6.5 (4.3) 9.0 (4.9) 9.25 (4.5) 7.9 (3.3) 47.6 (13.5) 3.6 (0.8) 14.9 (8.6) 1.8 (0.5)</td>
</tr>
<tr>
<td>Week 8</td>
<td></td>
</tr>
<tr>
<td>Values, n (%)</td>
<td>28 (74) 28 (74) 28 (74) 28 (74) 28 (74) 27 (71) 27 (71) 27 (71)</td>
</tr>
<tr>
<td>Values, mean (SD)</td>
<td>6.3 (3.9) 9.5 (5.4) 8.7 (5.0) 7.8 (2.7) 44.1 (12.4) 4.0 (0.9) 12.4 (7.3) 1.9 (0.5)</td>
</tr>
<tr>
<td>Week 12</td>
<td></td>
</tr>
<tr>
<td>Values, n (%)</td>
<td>30 (79) 30 (79) 30 (79) 30 (79) 30 (79) 29 (76) 29 (76) 29 (76)</td>
</tr>
<tr>
<td>Values, mean (SD)</td>
<td>6.2 (4.1) 8.5 (4.6) 8.5 (4.3) 7.6 (3.1) 44.4 (10.8) 4.0 (0.8) 12.2 (6.9) 1.9 (0.5)</td>
</tr>
<tr>
<td>Delayed iDBTj (n=34)</td>
<td></td>
</tr>
<tr>
<td>Week 0</td>
<td></td>
</tr>
<tr>
<td>Values, n (%)</td>
<td>34 (100) 34 (100) 34 (100) 34 (100) 34 (100) 34 (100) 34 (100) 34 (100)</td>
</tr>
<tr>
<td>Values, mean (SD)</td>
<td>6.8 (3.5) 11.1 (6.7) 8.9 (6.1) 6.9 (3.5) 50.5 (12.1) 3.5 (1.0) 15.5 (9.2) 1.8 (0.4)</td>
</tr>
<tr>
<td>Week 4</td>
<td></td>
</tr>
<tr>
<td>Values, n (%)</td>
<td>34 (100) 34 (100) 34 (100) 34 (100) 34 (100) 33 (97) 33 (97) 33 (97)</td>
</tr>
<tr>
<td>Values, mean (SD)</td>
<td>5.6 (3.9) 10.1 (6.6) 8.6 (6.1) 6.9 (3.4) 49.9 (13.2) 3.7 (1.0) 14.3 (9.0) 1.7 (0.4)</td>
</tr>
<tr>
<td>Week 8</td>
<td></td>
</tr>
<tr>
<td>Values, n (%)</td>
<td>28 (82) 28 (82) 28 (82) 28 (82) 28 (82) 28 (82) 28 (82) 28 (82)</td>
</tr>
<tr>
<td>Values, mean (SD)</td>
<td>5.4 (3.9) 7.6 (6.1) 7.6 (5.4) 5.9 (3.3) 44.6 (13.5) 3.8 (0.9) 11.8 (9.3) 1.9 (0.5)</td>
</tr>
<tr>
<td>Week 12</td>
<td></td>
</tr>
<tr>
<td>Values, n (%)</td>
<td>28 (82) 28 (82) 28 (82) 28 (82) 28 (82) 28 (82) 28 (82) 28 (82)</td>
</tr>
<tr>
<td>Values, mean (SD)</td>
<td>5.3 (3.9) 7.3 (5.3) 7.1 (4.6) 6.1 (3.3) 43.1 (12.9) 4.1 (1.1) 11.1 (9.2) 2.0 (0.4)</td>
</tr>
</tbody>
</table>

aDescriptive statistics for variables with ordinal values (all secondary outcome variables) are presented in Multimedia Appendix 2.
bSDS: Substance Dependence Scale.
cPHQ-9: Patient Health Questionnaire-9.
dGAD-7: Generalized Anxiety Disorder-7.
eSBQ: Suicidal Behaviors Questionnaire.
fDERS-16: Difficulties in Emotion Regulation Scale-16 item.
gMAAS: Mindful Attention Awareness Scale.
hWHODAS: World Health Organization Disability Assessment Schedule.
iDBT-WCCL: Dialectical Behavior Therapy Ways of Coping Checklist.
jiddb: internet-delivered dialectical behavioral therapy.
k=number of participants contributing to the calculations.

https://mental.jmir.org/2024/1/e50399
Contrary to our hypotheses, we did not find any significant grouptime interactions for the severity of substance dependence at week 4 or week 12.

Secondary Outcomes

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

Hypothesis 3: Did iDBT Produce Improvements Regardless of Group?

Primary Outcome

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

Secondary Outcomes

There were several findings supporting the benefits of the intervention in the follow-up phase of the study (no other significant main effects emerged at week 4). At week 12, there were significant main effects of time for depression (b=−2.95; SE 0.79; 95% CI −4.50 to −1.39; P<.001), anxiety (b=−1.57; SE 0.73; 95% CI −2.99 to −0.14; P=.03), suicidality (b=−0.70; SE 0.24; 95% CI −1.12 to −0.28; P=.001), emotion dysregulation (b=−6.56; SE 1.20; 95% CI −8.90 to −4.21; P<.001), functional disability (b=−3.64; SE 0.77; 95% CI −5.15 to −2.14; P<.001), dispositional mindfulness (b=0.44; SE 0.09;
95% CI 0.27-0.62; \( P < .001 \), and DBT skill acquisition (b=0.14; SE=0.05; 95% CI 0.04-0.23; \( P = .005 \)), indicating that both groups saw significant improvements from baseline over the study duration, with no differences between groups (all \( P > .25 \)). There were no main effects of time for risky impulsive behaviors and no difference between groups (\( P > .23 \)).

**Discussion**

**Summary**

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

**Feasibility and Acceptability**

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

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

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

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

**Potential Efficacy**

With regard to our second hypothesis evaluating the waitlist control design of the study, we found little evidence that the immediate iDBT group benefited more in the acute period of 4 weeks. We did not find a difference between groups in our primary outcome at this time point, where we expected it. We found interactions at week 4 for 2 secondary outcomes (ie, depression and anxiety) in favor of the hypothesis. We found additional interactions during the follow-up period (week 12) for decreased standard alcoholic drinks per day and overall substance use frequency in favor of the immediate iDBT group. Notably, we originally planned a 16-week trial with an equal immediate and waitlist period of 8 weeks. However, in a pilot study, we found attrition and lack of engagement to be greater than in this study, which contributed to the revised study design.
Most studies on app- and internet-based interventions use one month as an acute test of the intervention with another month as a follow-up assessment period owing to increasing attrition after 7 to 8 weeks [35-37].

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

Limitations

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

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

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

Conclusions

Notwithstanding the limitations of this study, our iDBT intervention Pocket Skills 2.0 was supported as a feasible and acceptable intervention for those with SUIDs and other mental health concerns. However, methods to improve engagement should be further evaluated. The intervention not only showed potential effectiveness for substance dependence but also demonstrated positive effects across various mental health symptoms, affirming its clinical utility. These findings add to the sparse literature on internet-based DBT and internet-delivered psychological interventions for SUIDs. This format has the potential to increase accessibility and reduce the costs and resources required for in-person DBT. Several research priorities were identified to potentially improve engagement and optimize treatment outcomes as well as understand how our iDBT intervention can be integrated into the larger landscape of treatment options for SUIDs and other conditions.
Acknowledgments
This study was supported by a Canadian Institutes for Health Research Fellowship Award (#458772) awarded to ARD. ARD would like to thank Daniel J Tse for help in organizing and creating formulas to summarize the website analytic data and Megan Brndjar for help in organizing and tabulating the outcome data. The authors sincerely thank Geraldine “Geri” Rodriguez and Psychwire for permitting us to use their training videos featuring Dr Linehan.

Data Availability
Deidentified data may be requested from ARD.

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

Multimedia Appendix 1
CONSORT-eHEALTH checklist (V 1.6.1).
[PDF File (Adobe PDF File), 8678 KB - mental_v11i1e50399_appl1.pdf]

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

References
3. Schulte MT, Hser YI. Substance use and associated health conditions throughout the lifespan. Public Health Rev 2014;35(2) [FREE Full text] [doi: 10.1007/BF03391702] [Medline: 28366975]


45. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001 Sep;16(9):606-613 [FREE Full text] [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]


74. Hoeppner BB, Stout RL, Jackson KM, Barnett NP. How good is fine-grained timeline follow-back data? Comparing 30-day TLFB and repeated 7-day TLFB alcohol consumption reports on the person and daily level. Addict Behav 2010 Dec;35(12):1138-1143 [FREE Full text] [doi: 10.1016/j.addbeh.2010.08.013] [Medline: 20822852]


Abbreviations

CONSORT: Consolidated Standards of Reporting Trials
CONSORT-EHEALTH: Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth
DBT: dialectical behavior therapy
iDBT: internet-delivered dialectical behavior therapy
RCT: randomized controlled trial
REDCap: Research Electronic Data Capture
SUD: substance use disorder
Identification of Predictors of Mood Disorder Misdiagnosis and Subsequent Help-Seeking Behavior in Individuals With Depressive Symptoms: Gradient-Boosted Tree Machine Learning Approach

Jiri Benacek1, PgDip; Nimotalai Lawal1, MEng; Tommy Ong1, MEng; Jakub Tomasik1, PhD; Nayra A Martin-Key1, PhD; Erin L Funnell1,2, BSc; Giles Barton-Owen2, MA; Tony Olmert1, MD; Dan Cowell2, BSc; Sabine Bahn1,2, MD, PhD

1Department of Chemical Engineering and Biotechnology, Cambridge Centre for Neuropsychiatric Research, University of Cambridge, Cambridge, United Kingdom
2Psyomics Ltd, Cambridge, United Kingdom

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

Abstract

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

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

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

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

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

(JMIR Ment Health 2024;11:e50738) doi:10.2196/50738
KEYWORDS
misdiagnosis; help-seeking; gradient-boosted trees; machine learning; depression; bipolar disorder; diagnose; diagnosis; mood; mental health; mental disorder; mental disorders; depression; depressive; predict; predictive; prediction; depressed; algorithm; algorithms

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

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

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

Although previous studies have investigated factors contributing to the misdiagnosis, poor help-seeking behavior, and barriers to receiving a diagnosis, only a few have used machine learning methods to do so [19]. The use of machine learning in mental health research has increased in recent years, with many studies focusing on detection and diagnosis, treatment and support, public health, and research and clinical administration [19]. While not without limitations, the use of machine learning can offer data-driven insights into complex relationships between high-dimensional data [20,21]. Although other, mostly qualitative investigations have identified the predictors of help-seeking and misdiagnosis by considering factors individually, this study aims to take a more holistic approach. By developing machine learning models based on extensive self-reported patient data, we aim to identify and quantify interdependent predictive factors for the misdiagnosis of mental health disorders, specifically mood disorders, and help-seeking behavior in individuals who may have been misdiagnosed. Identifying such predictive factors could aid in avoiding preventable misdiagnosis, encourage help-seeking, and improve outcomes in patients presenting with depressive symptoms.

Methods
Data Acquisition
Overview
The data used in this report were collected as part of the Delta Study—a study aiming to facilitate a more accurate and earlier diagnosis of BD and MDD; carried out in the United Kingdom by the Cambridge Centre for Neuropsychiatric Research between 2018 and 2020 [22,23]. The study consisted of an adaptive digital questionnaire, the Composite International Diagnostic Interview (CIDI) [24], and 2 follow-up questionnaires at 6 and 12 months. The stages of the Delta Study are summarized in Figure 1. Participants were recruited nonrandomly through email, the Cambridge Centre for Neuropsychiatric Research (CCNCR) website, and paid Facebook advertisements. The eligibility criteria included at least mild depressive symptoms, indicated by a score of ≥5 on the Patient Health Questionnaire-9 (PHQ-9) [25] at the time of recruitment, aged between 18 and 45 years, and residency in the United Kingdom. Participants who indicated current suicidal ideation or intent, were pregnant, or breastfeeding, were excluded.
Adaptive Digital Questionnaire

In total, 3232 participants completed the adaptive digital questionnaire available on the Delta Study digital platform. The questionnaire consisted of 635 questions, divided into six sections: (1) demographic information and personal history; (2) manic and hypomanic symptoms; (3) depressive symptoms; (4) personality profiling; (5) treatment, medication, substance use, and family psychiatric history; and (6) other psychiatric conditions. As the questionnaire was adaptive to answers given by participants, the maximum number of questions an individual could answer was 382, with an average of 284. Within the questionnaire, participants reported their baseline diagnosis, and their current well-being (within the previous 14 days) was quantified using the Warwick-Edinburgh Mental Well-Being Scale (WEMWBS) [26].

Composite International Diagnostic Interview

Participants who completed the web-based mental health questionnaire were invited to complete the CIDI version 3.0 via telephone. The CIDI is a structured diagnostic interview for mental disorders created by the World Health Organization based on the International Classification of Diseases and Related Health Problems, 10th Revision (ICD-10). It was developed primarily for epidemiological studies and has been extensively validated, demonstrating high diagnostic reliability [27]. In this study, only sections pertaining to mood disorder diagnoses were applied, that is, the demographics, depression, and mania modules. Interviewers were trained by CIDI-certified instructors prior to conducting the interviews. In total, 924 participants completed the CIDI and received one of the following diagnoses in their results report: BDI, BDII, subthreshold BD, MDD with subthreshold BD, MDD, or no mood disorder diagnosis (referred to as “low mood”).

Follow-Up Questionnaires

Participants who completed the digital questionnaire were invited to fill out 2 follow-up questionnaires, 6 and 12 months after receiving their results report. The follow-up questionnaires aimed to determine the effects of participation in the Delta Study on participants’ quality of life and record subsequent changes in diagnosis and treatment. A total of 2064 participants...
completed at least 1 of the follow-up questionnaires, with 1780 respondents at 6 months and 1542 respondents at 12 months.

**Outcomes**

**Overview**

For the purposes of this study, 2 dependent variables were defined.

**Misdiagnosis**

For participants who completed the CIDI, the mood disorder diagnosis reported at baseline was compared to the diagnosis obtained from the CIDI, including patients with no mood disorder diagnosis at baseline who should have been diagnosed. CIDI diagnosis was used as the gold standard, and any mismatch with the baseline diagnosis was defined as misdiagnosis. This definition of misdiagnosis was consistent with previous studies investigating under- and misdiagnosis of mood disorders based on comparing patient-reported diagnoses to the outcomes of structured clinical interviews [28,29].

**Help-Seeking Behavior**

In the 6- and 12-month follow-up questionnaires, participants were asked: “Have you had an appointment with a GP or psychiatrist to talk about your mental health in the past 6 months?” A positive response to this question at either time point was defined as help-seeking. In order to examine help-seeking in misdiagnosed individuals, only those who were identified as misdiagnosed within outcome 1 were included in the analysis.

**Analysis**

**Overview**

Raw data processing and feature engineering were performed in R (version 3.6.3; R Core Team) [30]. Subsequent analyses and modeling were carried out using Python (version 3.9.7; Python Software Foundation) [31]. Main libraries used included Pandas (version 1.5.2; Pandas Development Team) [32] and NumPy version 1.23.5 [33] for data manipulation; scikit-learn version 1.0.2 [34], XGBoost (version 1.6.1; The XGBoost Contributors) [35], and SHAP version 0.41.0 [36] for modeling and interpretation; and Seaborn version 0.12.1 [37] and Matplotlib version 3.6.2 [38] for plotting.

**Data Preparation**

Prior to analysis, constant and duplicate variables were removed. Answers to questions examining the same symptom or construct were concatenated, and new features were created to represent these aggregated answers. Missing data were imputed where possible (for example, the answer to the question asking “Has anyone suggested you drink less?” was set to 0 for participants who had indicated they do not drink), and otherwise remained nonrandomly missing. Categorical variables were 1-hot encoded, that is, unique dummy variables were created where the presence of each category was denoted by “1,” and its absence was represented by “0.”

**Modeling and Interpretation**

This analysis aimed to develop predictive models to identify variables influencing (1) misdiagnosis and (2) help-seeking behavior in participants who were identified as potentially misdiagnosed. A decision tree–based machine learning algorithm Extreme Gradient Boosting (XGBoost) [35] was chosen to train the classification models due to being robust to outliers, agnostic to data distribution, having the ability to handle nonrandom missing data, offering good predictive power, and due to it allowing for good model interpretability. Repeated nested cross-validation (rNCV) was used for model training and evaluation to obtain accurate estimates of model performance in unseen data. rNCV relies on performing a k-fold cross-validation (CV) within each round of another CV. This allows for model-specific hyperparameter optimization in the inner loop, with the final model being trained using the best-performing set of parameters, and later evaluated in the outer loop of rNCV. For this analysis, a 4-fold stratified CV was used in both the inner and outer loops, where 3 of the folds acted as a training set and 1 as a test set. Tuned model parameters included the number of estimators (1 to 100), shrinkage rate (0.1 to 0.3) to prevent overfitting, and tree depth (1 or 2) to allow for first-order interactions between predictors. The training was repeated 100 times, generating a total of 400 models for each of the objectives. Generalized model performance was evaluated by calculating the area under the receiver operating characteristic curve (AUC). The classification cutoff was optimized for the Youden index [39] to balance the true positive and true negative rates and offset potential imbalances between classes. SHAP (Shapley additive explanations) analysis [36], which combines local interpretable model–agnostic explanations (LIME) [40] and Shapley sampling values [41] approaches, was used for model interpretation. Feature occurrence frequency was calculated as the percentage of the models that incorporated a given feature to generate predictions. Reported results represent mean and SD values across the rNCV models.

**Ethical Considerations**

The study protocol was approved by the University of Cambridge Human Biology Research Ethics Committee (HBREC 2017.11) and all enrolled participants signed a digital informed consent form.

**Results**

**Misdiagnosis**

The self-reported baseline diagnosis did not match the diagnosis assigned by CIDI for 471 (50.97%) of the 924 participants who completed the CIDI interview. These participants were therefore considered misdiagnosed. No between-group differences were observed in terms of age, sex, ethnicity, highest achieved education level, or relationship status between the correctly diagnosed and misdiagnosed groups (Table S1 in Multimedia Appendix 1). However, there were significant differences in employment status as well as well-being and PHQ-9 scores, with misdiagnosed individuals, on average, reporting lower well-being and more severe depressive symptoms.

On average, the models correctly classified 70% (SD 9%) of misdiagnosed participants and 71% (SD 9%) of correctly diagnosed participants, with a mean accuracy of 70% (SD 3%) and the out-of-fold AUC of 0.75 (SD 0.03; Figure 2 and Table

https://mental.jmir.org/2024/1/e50738

JMIR Ment Health 2024 | vol. 11 | e50738 | p.44

(page number not for citation purposes)
S2 in Multimedia Appendix 1). Among the 1045 variables evaluated, the strongest predictors of misdiagnosis were more severe composite depressive symptoms and unstable self-image (Figure 3). Unstable self-image was measured by a 4-level Likert scale question “Is your image and sense of yourself and what you believe in unstable and constantly changing?” The next strongest predictor was the diagnosing clinician, with those who were undiagnosed at baseline or reported a diagnosis by a psychiatrist more likely to be misdiagnosed. The top 10 predictors also included variables related to age at diagnosis of BD and MDD, with late (≥35 years of age) diagnosis or no diagnosis at all, increasing the likelihood of being misdiagnosed (Figure S1 in Multimedia Appendix 1). Misdiagnosed participants were also more likely to recklessly spend money, experienced more frequent intense mood swings or mania in general, had higher weight gain during low mood episodes, and were more sexually active than usual at the time of data collection.

Figure 2. Out-of-fold model performance in predicting misdiagnosis. Green lines represent predictive performance on unseen out-of-fold data for each of the 400 final models. The thick blue line represents the average of all ROC curves. The grey area represents 1 SD. AUC: area under the receiver operating characteristic curve; ROC: receiver operating characteristic.
Figure 3. Results for the top 10 variables in the misdiagnosis model. Shown is SHAP analysis of the factors associated with misdiagnosis. The features (y-axis) are ordered by their average feature importance, indicated by the value inside the brackets, across all models. Each colored dot represents a participant, where the color gradient shows the value of the answer (red if low, green if high, and grey if missing), and the corresponding value on the x-axis shows directionality and the impact on model output, as determined using SHAP analysis. Values below 0 show directionality toward being correctly diagnosed, whereas values above 0 show directionality toward misdiagnosis. SHAP: Shapley additive explanations.

Model performance was largely driven by the top 5 predictors, with a steady decline in SHAP scores for subsequent variables. Of the top 10 predictors, 9 were selected in more than 75% (n=300) of the models, suggesting a relatively stable model composition. The exception was a variable related to “being more sexually active than usual,” which was selected in 71% (n=284) of the models. More detailed information on feature selection frequency is provided in Figure S3 in Multimedia Appendix 1.

Help-Seeking Behavior
Help-seeking behavior was investigated in 379 participants who were misdiagnosed at the baseline and who had completed at least 1 of the follow-up questionnaires. Of those, 229 (60.42%) participants sought an appointment with a medical professional during the follow-up period to discuss their mental health and were therefore defined as “help-seekers.” The help-seeker and non–help-seeker groups differed significantly in the highest achieved education level, relationship status, well-being, and severity of depressive symptoms (Table S3 in Multimedia Appendix 1). Participants more likely to seek help were on average less formally educated, more likely single, reported higher mean severity of symptoms, and worse overall well-being.

The model achieved an AUC of 0.71 (SD 0.04; Figure 4), with a sensitivity of 65% (SD 13%), specificity of 72% (SD 13%), and average accuracy of 67% (SD 4%; Table S4 in Multimedia Appendix 1). The strongest predictor was the shorter time since patients last spoke to a general practitioner (GP) about their mental health at baseline (Figure 5). It was followed by sleep problems disrupting daily tasks and taking prescribed antidepressants, both associated with increased help-seeking. Consistent with this, lower help-seeking was observed in participants who had never been prescribed antidepressants, namely selective serotonin reuptake inhibitors (SSRIs). Furthermore, there was a lower likelihood of help-seeking with higher age at both the first episode of low mood and diagnosis of depression, which was similarly predictive to not having been previously diagnosed with depression. Finally, impaired ability to work, lower well-being scores, feeling worthless, and lower self-rated mental health were associated with help-seeking behavior.
Figure 4. Out-of-fold model performance in predicting help-seeking. Green lines represent predictive performance on unseen out-of-fold data of each of the 400 final models. The thick blue line represents an average of all ROC curves. The grey area represents 1 SD. AUC: area under the receiver operating characteristic curve; ROC: receiver operating characteristic.

Figure 5. Results for top 10 variables in the help-seeking model in misdiagnosed individuals. Shown is feature SHAP importance (in brackets) and feature SHAP values (data points). SHAP values below 0 show directionality toward low help-seeking (i.e., no appointment with GP or psychiatrist to discuss mental health), whereas values above 0 show directionality toward high help-seeking. GP: general practitioner; SHAP: Shapley additive explanations; SSRI: selective serotonin reuptake inhibitor.
The 3 variables, namely, time since last spoken to a GP, sleep problems disrupting daily tasks, and still taking prescribed antidepressants, were selected in nearly all models (Figure S4 in Multimedia Appendix 1), suggesting their high relevance for model predictions. Among other predictors, only age when diagnosed with depression was selected in more than 75% (n=300) of models, with the remaining features only selected in approximately 50% (n=200) of models, indicating their lower relevance.

Discussion

Principal Findings

This study aimed to develop machine learning models to explore factors potentially contributing to misdiagnosis and subsequent help-seeking in individuals experiencing low mood. For this purpose, we used data obtained through an extensive digital questionnaire concerning demographic, personality, and mental health data, as well as, the validated and standardized diagnostic interview, CIDI. Developed models achieved a fair level of predictive power, with AUCs of 0.75 and 0.71 for predicting misdiagnosis and help-seeking, respectively. Below, we discuss the main findings as well as the strengths and limitations of this analysis.

Misdiagnosis

The strongest predictor of misdiagnosis was the severity of depressed mood, with more severe depressive symptoms being associated with a greater risk of being misdiagnosed. This directionality was consistent with other top predictors of misdiagnosis, including unstable self-image, reckless spending, frequent intense mood swings, mania, weight gain during low mood, and being more sexually active than usual. Except for the instability of self-image, these predictors can be divided into either depression or mania or bipolar-related symptoms. Overall, the finding that individuals with more severe mental health symptoms are at a greater risk of being misdiagnosed is surprising, given the opposite could be expected as milder symptoms are harder to detect [42]. Several factors could contribute to this association, including the complexity of diagnosing mental health disorders [43], variability in symptom presentation [44,45], and the high degree of symptom overlap across different diagnoses [5]. A possible explanation for the increased risk of misdiagnosis among individuals with more severe symptoms is that they may present with prominent mood instability, such as that observed in patients with personality disorder, or rapidly cycling symptoms, making accurate diagnosis more challenging [9]. In addition, individuals with more severe symptoms often lack motivation to seek help, hence their symptoms may remain unrecognized for a longer time [46].

In the case of mood disorders, misdiagnosis of individuals with higher depressive symptom severity may result from the fact that patients with BD generally seek medical help during depressive episodes and often present with more severe depressive symptoms than patients with MDD, while underreporting manic phases [47,48]. In fact, less than a third of patients with BD report the presence of reckless behavior, excessive spending, and increased sexual interest or activity [49]. This contributes to approximately 40% of patients with BD receiving an incorrect initial diagnosis of unipolar depression [50]. Also, the association of frequent intense mood swings with mood disorder misdiagnoses may be related to incorrect treatment of depressive symptoms of BD with antidepressants, rather than mood stabilizer medication, which has the potential to induce mania and rapid cycling [51,52]. The second most predictive feature of misdiagnosis identified in this study was unstable self-image. Previous literature has shown that an unstable sense of self is associated with frequent changes in diagnosis, and often linked to complex and unstable personality characteristics [53]. The high ranking of self-image stability could, however, be a result of the high comorbidity rates between BD and other disorders featuring unstable self-image that were not evaluated by the diagnostic interview used in this study, such as borderline personality disorder [54]. This is especially important considering that such disorders may share a high number of similarities with BD, leading to frequent misdiagnoses [55,56]. The 2 additional symptoms that are ranked high in terms of predictive value for misdiagnosis in this analysis regard reckless spending and increased sexual activity, representing reckless or impulsive behavior, which are included in the diagnostic criteria of both BD and borderline personality disorder [17].

Finally, among the top predictors of misdiagnosis were 3 variables related to psychiatric history, including psychiatrist involvement in the diagnosis, age at depression diagnosis, and age at BD diagnosis. Interestingly, the models attributed a higher risk of misdiagnosis to individuals whose depression was diagnosed by a psychiatrist. This may be caused by the fact that patients at high risk of misdiagnosis, such as those with more complex symptom presentation or suspected comorbidities, are usually referred to secondary care, following the National Institute for Health and Care Excellence (NICE) guidelines [57]. However, this finding should be interpreted with caution, as diagnoses made by psychiatrists are generally more accurate than those derived from the CIDI. Also, participants who received a diagnosis of a mood disorder at an older age, or not at all, were more likely to be misdiagnosed. This finding is surprising, as previous literature suggests that the severity and impact of symptoms decline with age, with 86% of patients with BD diagnosed by the age of 25 [58]. However, it is possible that due to milder symptoms, patients who are older may remain undiagnosed for longer periods of time.

Help-Seeking

Analysis of participants with a mismatch between their self-reported formal diagnosis and the CIDI outcome revealed several predictors of help-seeking related to patients’ mental health history and symptoms. The most predictive feature was time since last spoken to a GP at baseline, with patients who had visited their GP more recently being more likely to seek help. Interestingly, that was not the case for the time since last spoken to a psychiatrist, likely due to most participants not being under secondary care and the long waiting times for psychiatric assessment [59]. In line with previous literature [60], these findings indicate that help-seeking was also associated with more severe psychiatric symptoms and

https://mental.jmir.org/2024/1/e50738

JMIR MENTAL HEALTH 2024 | vol. 11 | e50738

Benacek et al

(1 page number not for citation purposes)
having a previous diagnosis of mood disorder. Similarly, participants seeking help reported lower well-being, feeling more worthless, and more functional impairment in carrying out daily tasks and at work caused by symptoms and sleep problems.

Interestingly, while there was not a significant overall age difference between the help-seekers and non–help-seekers, further analyses showed a lower tendency to seek help in individuals who were over 35 years old at initial diagnosis of depression (Figure S2 in Multimedia Appendix 1). The pattern of people who are younger being more likely to seek help is in line with the published literature [61]. Together with the finding that the initial diagnosis at older age was a strong predictor of misdiagnosis [62], this result indicates that patients who are most likely to be misdiagnosed are also the least likely to seek help. Thus, older patients may require more support to tackle potential barriers to help-seeking and receiving a diagnosis, such as stigma and inadequate mental health education [63].

The final set of predictors of help-seeking was related to previous medication. Higher help-seeking was observed in misdiagnosed individuals who were still taking previously prescribed antidepressants, in particular SSRIs, as opposed to misdiagnosed individuals who either had never been prescribed SSRIs or other antidepressant medication or had stopped taking it. The association of antidepressant treatment with help-seeking indicates that the prescribed medication may have been ineffective, as is often the case when attempting to treat depressive episodes of BD with antidepressant monotherapy [64]. Compared with the patients with MDD, the patients with BD respond worse to antidepressant medication, with short-term nonresponse rates of 51.3% in BD versus 31.6% in MDD [65]. This difference is even more pronounced in the long-term, where the loss of response to antidepressants is 3.4 times more frequent in patients with BD, while withdrawal relapse into depression is 4.7 times less frequent in BD compared to patients with MDD [65]. Moreover, individuals with unrecognized BD who are treated with antidepressants sometimes develop symptoms of mania, which in turn may motivate patients or their relatives to seek consultation with a specialist [66].

Limitations

The main limitation of this study is the reliance on CIDI as the gold standard for mood disorder diagnosis. Although the CIDI demonstrates good agreement with structured diagnostic interviews conducted by clinicians [67], future studies should consider either retrospective or longitudinal study designs, and ideally access medical records for more accurate diagnoses, including those beyond mood disorders. Additionally, the study participants were recruited online following strict inclusion criteria and were predominantly White, necessitating further research in traditionally underrepresented ethnic minorities and more representative patient cohorts. Another limitation is the exclusion of individuals with current suicidal ideation, a characteristic that could be an important indicator of misdiagnosis. Finally, the observed associations do not necessarily imply causality, which can only be evaluated through prospective causal inference study designs.

Conclusions

This analysis leveraged comprehensive patient data, a robust machine learning algorithm, and an extensive validation framework, to identify predictors of mood disorder misdiagnosis in individuals experiencing depressive symptoms, and subsequent help-seeking. The results highlight the increased risk for misdiagnosis associated with incomplete symptom profiles, more severe or harder to detect symptoms, and older age. Therefore, comprehensive symptom monitoring outside of depressive episodes, mental health screening at earlier ages, and clinician knowledge of the influence of advanced age on misdiagnosis risk are important considerations for early and accurate diagnosis of mood disorders. Moreover, prior engagement with mental health services, functional impairment in performing daily tasks, and younger age were associated with a higher likelihood of help-seeking. Together, these results add to the growing application of machine learning techniques in examining existing barriers to accessing mental health services [19], and may ultimately lead to the development of novel screening tools or procedures for a comprehensive mental health risk assessment in individuals presenting with mood-related symptoms.

Acknowledgments

We are most grateful to all participants of the Delta Study for their time and efforts. We are also grateful to all members of the Delta Study Service User Advisory Group. We also thank all those involved in designing and conducting the Delta Study for their valuable input including Jason D Cooper, Sung Yeon Sarah Han, Lynn P Farrag, Emily Bell, Lauren V Friend, Sharmelee Thiahulan, Pawel Eljasz, Mark Agius, Neil Hunt, and the CIDI interviewers. This study was funded by the Stanley Medical Research Institute (07R-1888) and Psyomics Ltd. Stanley Medical Research Institute was not involved in any part of the research. Psyomics Ltd was involved in the design and conduct of the study; recruitment, data collection, and management; and review or approval of the study.

Authors’ Contributions

SB and DC conceived the Delta Study, conceptualized, and supervised the development of the web-based mental health questionnaire. SB, DC, GBO, and T Olmert contributed to the design of the study. GBO and T Olmert collected the web-based mental health questionnaire data. GBO and JT processed the web-based mental health questionnaire data. JB, NL, and T Ong analyzed the data. SB, JT, NAMK, and ELF advised the analysis. JB and JT wrote the first draft of the study, with contributions from NL, T Ong, NAMK, ELF, and SB. All authors contributed to the final version of the study.
Conflicts of Interest

SB is a director of Psynova Neurotech Ltd and Psynomics Ltd. SB, ELF, and DC have financial interests in Psynomics Ltd. GBO had financial interests in Psynomics Ltd. SB, JT, and T Olmert have received payments from the University of Cambridge for licensing of data from the Delta Study. SB and JT may benefit financially from patents arising from the Delta Study. ELF is a consultant for Psynomics Ltd. All other authors declare no competing interests.

Multimedia Appendix 1
Demographics, model performance metrics, dependence plots, and feature selection frequencies for all objectives.

References


15. Davies SC. Annual report of the chief medical officer 2013, public mental health priorities: investing in the evidence. Department of Health. 2014. URL: https://assets.publishing.service.gov.uk/media/5a8034e0e5274a2e87d87db7d/CMO_web_doc.pdf [accessed 2023-12-19]


Abbreviations

AUC: area under the receiver operating characteristic curve
BD: bipolar disorder
CCNR: Cambridge Centre for Neuropsychiatric Research
CIDI: Composite International Diagnostic Interview
Speech Features as Predictors of Momentary Depression Severity in Patients With Depressive Disorder Undergoing Sleep Deprivation Therapy: Ambulatory Assessment Pilot Study

Lisa-Marie Wadle, MSc; Ulrich W Ebner-Priemer, Prof Dr; Jerome C Foo, PhD; Yoshiharu Yamamoto, Prof Dr; Fabian Streit, PhD; Stephanie H Witt, PhD; Josef Frank, PhD; Lea Zillich, PhD; Matthias F Limberger, MA; Ayimnisagul Ablimit, MSc; Tanja Schultz, Prof Dr; Maria Gilles, MD; Marcella Rietschel, Prof Dr Med; Lea Sirignano, MSc

1Mental mHealth Lab, Institute of Sports and Sports Science, Karlsruhe Institute of Technology, Karlsruhe, Germany
2Department of Psychiatry and Psychotherapy, Central Institute of Mental Health, University of Heidelberg, Mannheim, Germany
3Department of Genetic Epidemiology in Psychiatry, Central Institute of Mental Health, University of Heidelberg, Mannheim, Germany
4Institute for Psychopharmacology, Central Institute of Mental Health, University of Heidelberg, Mannheim, Germany
5Department of Psychiatry, College of Health Sciences, University of Alberta, Edmonton, AB, Canada
6Educational Physiology Laboratory, Graduate School of Education, University of Tokyo, Tokyo, Japan
7Cognitive Systems Lab, University of Bremen, Bremen, Germany

Corresponding Author:
Lisa-Marie Wadle, MSc
Mental mHealth Lab
Institute of Sports and Sports Science
Karlsruhe Institute of Technology
Hertzstrasse 16, Bldg 06.31
Karlsruhe, 76187
Germany
Phone: 49 72160847543
Email: lisa.wadle@kit.edu

Abstract

Background: The use of mobile devices to continuously monitor objectively extracted parameters of depressive symptomatology is seen as an important step in the understanding and prevention of upcoming depressive episodes. Speech features such as pitch variability, speech pauses, and speech rate are promising indicators, but empirical evidence is limited, given the variability of study designs.

Objective: Previous research studies have found different speech patterns when comparing single speech recordings between patients and healthy controls, but only a few studies have used repeated assessments to compare depressive and nondepressive episodes within the same patient. To our knowledge, no study has used a series of measurements within patients with depression (eg, intensive longitudinal data) to model the dynamic ebb and flow of subjectively reported depression and concomitant speech samples. However, such data are indispensable for detecting and ultimately preventing upcoming episodes.

Methods: In this study, we captured voice samples and momentary affect ratings over the course of 3 weeks in a sample of patients (N=30) with an acute depressive episode receiving stationary care. Patients underwent sleep deprivation therapy, a chronotherapeutic intervention that can rapidly improve depression symptomatology. We hypothesized that within-person variability in depressive and affective momentary states would be reflected in the following 3 speech features: pitch variability, speech pauses, and speech rate. We parametrized them using the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) from open-source Speech and Music Interpretation by Large-Space Extraction (openSMILE; audEERING GmbH) and extracted them from a transcript. We analyzed the speech features along with self-reported momentary affect ratings, using multilevel linear regression analysis. We analyzed an average of 32 (SD 19.83) assessments per patient.

Results: Analyses revealed that pitch variability, speech pauses, and speech rate were associated with depression severity, positive affect, valence, and energetic arousal; furthermore, speech pauses and speech rate were associated with negative affect, and speech pauses were additionally associated with calmness. Specifically, pitch variability was negatively associated with
improved momentary states (ie, lower pitch variability was linked to lower depression severity as well as higher positive affect, valence, and energetic arousal). Speech pauses were negatively associated with improved momentary states, whereas speech rate was positively associated with improved momentary states.

Conclusions: Pitch variability, speech pauses, and speech rate are promising features for the development of clinical prediction technologies to improve patient care as well as timely diagnosis and monitoring of treatment response. Our research is a step forward on the path to developing an automated depression monitoring system, facilitating individually tailored treatments and increased patient empowerment.

(KEYWORDS

ambulatory assessment; experience sampling; ecological momentary assessment; speech features; speech pattern; depression; sleep deprivation therapy; mobile phone

Introduction

Background

Depression is one of the most prevalent health disorders worldwide [1,2]. The World Health Organization predicted that depression would be 1 of the 3 leading causes of disease burden by 2030 [3], even before its prevalence increased owing to the COVID-19 pandemic [4]. This disorder has symptoms that include depressed mood, loss of energy and interest, sleep problems, and diminished ability to concentrate [5]; thus, depression imposes a substantial burden on the patients as well as their surroundings, society, and the economy [6]. Most importantly, depression is a chronic disorder, characterized by multiple episodes over years or decades. However, strategies for secondary prevention or early detection of new episodes are missing.

The diagnosis and severity assessment of depression relies mostly on self- or caregiver reports, which are prone to retrospective and social desirability bias [7,8]. In addition, such assessments are time and resource intensive because clinical specialists are needed over the course of treatment and recovery [9]. Moreover, many new episodes remain undiagnosed or untreated, that is, secondary prevention is the main issue [10,11]. To reduce burden, the timely detection and diagnosis of (new) depressive episodes are critical.

In recent years, research has focused on the identification of mental health disorder indicators that can be derived automatically, driven by technological developments [12,13]. In particular, the innovation of the ambulatory assessment research technique has contributed strongly to this endeavor [14]. Different terms have been used for this kind of methodology: ambulatory assessment [15], ecological momentary assessment [16], experience sampling [17], and digital phenotyping [18]. Although the terms differ, all approaches use computer-assisted methodology to assess momentary self-reported symptoms (eg, via electronic diaries [ediaries]), behaviors, or physiological processes, or actively or passively collect smartphone and physical data or context information (eg, via wearables) while the participant performs normal daily activities in their natural environment [19]. The main advantages of ambulatory assessment are (1) the ability to collect real-life data in real time, thereby reducing retrospective recall bias and increasing ecological validity; and (2) the ability to collect data continuously (passively), which allows us to capture dynamic changes. Accordingly, ambulatory assessment is a promising tool for the timely detection of upcoming clinical episodes to prevent further clinical deterioration [20-22]. In particular, parameters captured objectively by wearables are useful because they can be assessed passively with a high frequency over prolonged time periods [23].

Promising markers that can be assessed objectively are speech and language, which are also metaphorically called “the mirror of the soul” [24]. Even before objective measurements with ambulatory assessment technology were feasible, clinical observations described the voice of patients with depression as low, slow, and hesitant, with these patients speaking in a monotonous and expressionless manner [24,25]. Voice and speech production may be affected by typical characteristics of the clinical nature of depression; for example, psychomotor retardation, energy loss, and cognitive difficulties also affect the vocal folds, leading to a lower intensity, rate, and loudness of speech, which manifest in a monotone and toneless voice [26-28]. Recent reviews have highlighted the potential of using speech markers to assess a variety of psychiatric disorders [29], especially depression [30]. The use of speech as a marker has several advantages because it can be recorded (1) casually; (2) in a noninvasive manner at people’s homes or in public places (with consent provided); and (3) at low cost because microphones are integrated in many devices such as smartphones, smartwatches, and hearing aids. With the availability of open-source speech analysis software (eg, open-source Speech and Music Interpretation by Large-Space Extraction [openSMILE; audEERING GmbH] and Praat) and advances in automatic speech processing technologies based on machine learning techniques, research and development on the use of acoustic and linguistic features to identify mood disorders in particular [29] have been made possible.

Prior Work

Many studies have successfully discriminated between healthy controls and patients with depression based on speech features [30]. However, understanding within-person (vs between-person) depression-related voice changes is essential in detecting new episodes, that is, the secondary prevention. To the best of our knowledge, only a few studies in samples with clinical (not subclinical) depression have examined the variability of speech features within persons [31-36]. In a
6-week treatment-monitoring study, weekly speech samples were obtained from 35 patients with depression using an interactive voice response system [31]. Patients with an improvement in depressive symptoms showed a significant increase in pitch and pitch variability, an increase in speech rate, and shorter speech pauses while speaking at their final assessment compared with their baseline assessment. Importantly, patients whose depressive symptoms did not improve did not show these changes.

The data set of Mundt et al [31] was reanalyzed multiple times [32, 34, 35]. Quatieri and Malyska [34] integrated additional speech features and identified that lower pitch variability, shimmer, and Jitter as well as an increased harmonics-to-noise ratio were correlated with lower depression severity. This is in contrast to the study by Mundt et al [31], who found that increased pitch variability was associated with lower depression severity, which Quatieri and Malyska [34] attributed to differences in the set of voice samples analyzed (read speech in the study by Mundt et al [31] and conversational speech in the study by Quatieri and Malyska [34] from the same patients).

Trevino et al [32] discussed speech rate extraction methods based on the data set of Mundt et al [31] and replicated results regarding speech rate in automatically derived phonologically based features. Speech rate was negatively correlated with depression scores and the psychomotor retardation item in particular. Moreover, the authors replicated the finding that speech pauses were positively correlated with depression severity.

Furthermore, Horwitz et al [35] reanalyzed a subset of data from the study by Mundt et al [31] with a focus on disentangling how speech features relate to the total assessment score and individual symptom items. The authors found a positive correlation between pitch variability and depression scores and a slower speech rate with increasing depression severity. Notably, they analyzed a different speech task and a different depression assessment in comparison with Mundt et al [31].

Mundt et al [33] replicated their results from Mundt et al [31] in a larger study. Here, 105 patients were observed in a 4-week randomized placebo-controlled study. Again, analyses entailed a comparison of the final and baseline assessments. For patients benefiting from the treatment, total pause time was lower, pitch was higher (pitch variability was not assessed), and speech rate was higher. For patients who did not benefit from the treatment, only speech rate increased; however, it increased significantly less than in patients benefiting from the treatment.

Yang et al [36] analyzed clinical interviews recorded in 7-week intervals. In contrast to Mundt et al [31], they did not find a change in pitch variability with a change in depression severity in the patients but rather in the interviewers. The authors also found shorter switching pauses between patient and interviewer (i.e., both interlocutors) with lower depression severity.

Although not completely consistent, these findings support the assumption that voice features change within individuals when depression severity changes. However, although data were collected at multiple time points during the study (e.g., weekly), except in the study by Yang et al [36], the analysis was limited to a comparison between the baseline and final assessments. However, given that the goal is to detect and ultimately prevent new depressive episodes and deterioration, it is essential to understand within-person trajectories of voice features and how they are associated with momentary states with increased granularity. In this study, we used a naturalistic data set where a rapidly acting antidepressant treatment (i.e., sleep deprivation therapy [SDT] [37]) was applied to patients experiencing a depressive episode. The antidepressant effect vanishes in most of the cases after recovery sleep. Baseline, the treatment effect of SDT, and relapse can be measured in a matter of 4 days, making it a preferable setting to study within-person fluctuations.

Aims and Hypotheses
To investigate the within-person relationship between fluctuations in depression severity and fluctuations in speech features, we used a longitudinal data set with an average of 32 (SD 19.83) assessments per patient. All patients had experienced an acute depressive episode and undergone SDT [37], a chronotherapeutic intervention that can rapidly improve depression symptomatology. The main advantage of this therapeutic is that we maximize the variance of affective states within the data set and ensure sufficient within-person fluctuations over time. As the amount of speech features is immense, resulting in alpha error inflation, we focused on 3 speech features with high face validity that have shown first hints in past research [31-36]. Specifically, we hypothesized that (1) changes in pitch variability, (2) shorter speech pauses, and (3) higher speech rate are associated with lower depression severity. In addition, we assessed the associations of these features with additional momentary affective states (i.e., positive affect, negative affect, valence, energetic arousal, and calmness). We hypothesized that the associations of speech features with negative affect are similar to those for depression severity and that the associations of speech features with the other momentary affective states listed follow the opposite pattern.

Methods
Sample
We used a data set that was collected as part of a pilot study (Sleep Deprivation and Gene Expression [SLEDGE II]; German Clinical Trials Register: DRKS00022025) gathering digital phenotypes and multomics data in a clinical sample undergoing SDT at the Central Institute of Mental Health in Mannheim, Germany. A total of 30 inpatients experiencing acute depressive episodes were enrolled in the study. The patients were diagnosed according to the International Classification of Diseases, Tenth Revision (ICD-10), codes by the senior clinician at admittance to the hospital. All patients received treatment as usual, which also included SDT (for a list of medications, refer to Textbox S1 in Multimedia Appendix 1). Exclusion criteria were comorbid substance use disorders or personality disorders. From this sample of 30 patients, the complete data sets of 8 (27%) patients were excluded from the final analyses (n=4, 50% did not record any videos; n=1, 13% did not say anything during the videos [23 videos]; n=2, 25% had no sound recorded in the videos owing to technical issues [30 recordings]; and n=1, 13%...
recorded only 2 videos); thus, the final sample consisted of 22 (73%) patients (n=12, 55% male) aged between 18 and 63 (mean 33.5, SD 12.4; median 29, IQR 23.25-42.75) years.

**Ethical Considerations**

The study was approved by the Ethics Committee II of the Medical Faculty Mannheim, University of Heidelberg (2013-563N-MA). All patients received detailed information about the aims and procedures of the study and provided informed consent. Patients could withdraw from the study at any time and did not receive any compensation for participation. Data was deidentified to ensure privacy.

**Study Procedure**

Patients were given a study smartphone (Nokia 4.2 or Samsung Galaxy J7) at the beginning of the study (day 0), instructed on how to use it, and (if necessary) performed test runs supervised by the study personnel. A telephone number for technical support and an information sheet regarding the ambulatory assessment procedure were handed out. Data were collected using movisensXS software (movisens GmbH) [38]. Patients underwent SDT as part of their depression treatment, which involves staying awake for approximately 36 hours. Treatment effect and relapse can be measured in a matter of 4 days [37], thus ensuring a maximum of within-person variance in the data set. After at least 1 day of baseline assessment (day 0), SDT was conducted on day 1. Patients stayed awake from 6 AM on day 1 to 6 PM on day 2. Recovery sleep was allowed from 6 PM on day 2 until 1 AM on day 3. Data were collected before, during, and after SDT for up to 26 days. In the first week of the study, smartphones sent prompts 3 times per day (morning, afternoon, and evening); in addition, self-initiated assessments were possible to report specific events or to catch up with missed assessments. To reduce the burden on patients, the sampling schema was altered to 2 prompts per day (morning and evening). With each prompt, patients were requested to fill out items concerning their affective state and to record a selfie video reporting how they felt currently. Patients returned the smartphone at the end of the study. The study personnel uploaded the data from the smartphones to the movisensXS platform [38] and then downloaded the data for analysis.

**Ambulatory Assessment: eDiary Ratings and Selfie Videos**

The data set contains 3 sets of momentary assessments in the form of eDiary ratings at each prompt (Textboxes S2-S4 in Multimedia Appendix 1): (1) the short version of the Allgemeine Depressionsskala (ADS-K) [39] adapted to momentary assessment with 14 items on depressive mood rated on a scale ranging from 0=rarely to 3=mostly (we left out the item regarding sleep from the original questionnaire because its inclusion was not reasonable in the momentary assessment design); (2) a total of 15 positive (cheerful, content, energetic, enthusiastic, relaxed, and happy) and negative (lonely, sad, insecure, anxious, depressed, low-spirited, guilty, distrustful, and irritable) affect items [40] rated on a 5-point Likert scale ranging from 1=not at all to 5=very much; and (3) a 6-item short version of the Multidimensional Mood Questionnaire (MDMQ) [41] capturing time-varying momentary fluctuations in daily life on the affect dimensions of valence (unwell to well and discontent to content), energetic arousal (without energy to full of energy and tired to awake), and calmness (tense to relaxed and agitated to calm). The items were presented on visual analog scales with 2 poles and a slider from 0 to 100. For each of the constructs, we computed mean values per scale, resulting in 6 outcome variables (depressive symptoms, positive affect, negative affect, valence, energetic arousal, and calmness). For the ADS-K, we also report sum scores as described in the tool’s manual; however, to increase comparability among outcomes, we used the mean value for analyses. If necessary, we recoded items such that higher values indicated a (1) higher intensity of depressive symptoms, (2) higher positive affect, (3) higher negative affect, (4) higher positive valence, (5) higher energetic arousal, and (6) higher calmness.

In addition to the aforementioned eDiary ratings, patients were requested to record selfie videos with the following instructions: “Please keep the camera stable during the recording and record your whole face. Please describe in 10-20 seconds how you currently feel.”

**Clinical Assessments**

The Montgomery–Åsberg Depression Rating Scale (MADRS) [42] was completed in the morning at 4 time points (baseline, morning before sleep deprivation, 1 week after sleep deprivation, and 2 weeks after sleep deprivation) and once at midday (the day after sleep deprivation night). The MADRS is a 10-item expert assessment of depressive symptom severity over the past week, with items rated on a 7-point scale ranging from 0 to 6; higher scores indicate higher severity.

**Data Preprocessing**

The data set contained 899 selfie videos in mp4 format. The full set of videos of 4 (13%) of the 30 patients had to be excluded owing to the reasons mentioned previously (55/899, 6.1%) and additional 2 videos had to be excluded because of technical damage (2/899, 0.02%). As our research questions focused on audio data (not visual data), we extracted the audio tracks of the remaining 842 (93.66%) from the original 899 selfie videos using ffmpeg package in Python and archived them as wav files (sampling rate: 48 kHz; mono=1 channel). We excluded test runs (14/842, 1.7%), accidental short recordings with no content (29/842, 3.4%), recordings during which the microphone was masked by the patient (27/842, 3.2%), and assessments in which 1 of the 2 corresponding assessments (speech or affective state) was missing (18/842, 2.1%). In addition, if 2 consecutive assessments were <15 minutes apart from each other, only the first assessment was kept unless its audio quality was insufficient or only the second assessment included assessments of affective states; in such cases, the second assessment was kept (21/842, 2.5%). We also excluded recordings with background noise that restricted speech intelligibility (9/842, 1.1%) or that included the speech of third parties (8/842, 1%). We filtered the remaining recordings (716/842, 85%) using DeepFilterNet2 [43] to remove background noise.
Acoustic Features

For our main analyses, we focused on the acoustic features pitch variability, speech pauses, and speech rate (Table 1). We restricted the number of features to limit error inflation and selected specifically these 3 features because they revealed sufficient empirical support to warrant an explicit hypothesis. We extracted acoustic features of the final recordings (n=716) using the functionals (v02) of the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [44] of the open-source toolkit openSMILE implemented in Python [45,46]. eGeMAPS is a minimalistic set of acoustic features recommended for clinical speech analysis; it helps to guarantee comparability between studies, given the proliferation of speech features. Features related to frequency, energy, spectrum, and tempo are included in the set. Pitch variability is represented by the SD of the logarithmic fundamental frequency (F0) on a semitone frequency scale starting at 27.5 Hz and measured in hertz. F0 is the lowest frequency of the speech signal and is perceived as pitch. Speech pauses are approximated as the mean length of unvoiced regions (F0=0) measured in seconds. With respect to speech rate, a transcription of the recordings is necessary, which we obtained using an automatic speech recognition system according to published procedures [47]. We corrected the transcripts manually. To determine speech rate, we calculated the ratio of words divided by the duration of the voice sample.

Table 1. Overview of extracted speech features.

<table>
<thead>
<tr>
<th>Speech feature</th>
<th>Technical feature</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pitch variability</td>
<td>F0semitoneFrom27.5Hz_sma3nz_stdddevNorm</td>
<td>SD of the F0 perceived as the extent to which a person’s pitch changes (in Hz)</td>
</tr>
<tr>
<td>Speech pauses</td>
<td>MeanUnvoicedSegmentLength</td>
<td>Mean of the length of unvoiced regions approximating silent parts of the speech sample (in seconds)</td>
</tr>
<tr>
<td>Speech rate</td>
<td>Words per second</td>
<td>Ratio of words counted on the basis of the automatically transcribed and manually corrected text divided by the duration of the speech sample</td>
</tr>
</tbody>
</table>

Besides our main analyses based on pitch variability, speech pauses, and speech rate, we decided to integrate further eGeMAPS features in an exploratory analysis. These features have been recommended in the context of affective states in particular because they contain additional cepstral and dynamic features [44]. We included the following features in the exploratory analyses: for voiced and unvoiced regions together, the mean and SD of the mel-frequency cepstral coefficients (MFCCs) 1 to 4 and spectral flux difference of the spectra of 2 consecutive frames; for voiced regions, the formant 2 to 3 bandwidths along with spectral flux and MFCCs 1 to 4; and for unvoiced regions, the mean and SD of the spectral flux [44].

Statistical Analysis

In addition to the mean, SD, and range, we present min and max as the mean of all patients’ minimum and maximum scores, respectively, of each parameter throughout the whole study. Moreover, following the recommendations by Snijders and Bosker [48], we computed Pearson correlation analyses with person-mean–centered variables to evaluate the relationship between affective scores and speech features. To generate person-mean–centered variables, we subtracted the individual’s mean from their score, which represents the variation around the individual’s mean.

To evaluate psychometric properties, we calculated McDonald \( \omega \) as the reliability coefficient using the multilevelTools package in R. For the MDMQ subscales, we used the misty package in R to calculate the Spearman-Brown corrected correlation coefficients because the subscales consist of only 2 items [49]. For the MADRS score at the time of inclusion, we calculated Cronbach \( \alpha \) using the psych package in R.

To analyze the within-person association of speech features and subjectively evaluated affective states, we used multilevel modeling [48] using the nlme package in R. Multilevel modeling offers two specific advantages for the given data: (1) separation of within-person effects from between-person effects and (2) allowing and considering different numbers of assessments per patient. Before the analyses, we centered time-variant level-1 predictors (pitch variability, speech pauses, and speech rate) at the person level and included the predictors time and time\(^2\) in minutes (each centered at 2 PM) as covariates. To facilitate the comparison of the magnitude of effects among different predictors, we report standardized beta coefficients (standardized \( \beta \)) according to the recommendations by Hox and van de Schoot [50] following the equation: standardized \( \beta = \frac{\beta \times (SD_{predictor} / SD_{outcome})}{\sigma_{null}} \). We further calculated Hox R\(^2\) values according to the recommendation by Hox and Maas [51] following the equation: \( R^2_{Hox} = (\sigma_{null} - \sigma_{model}) / \sigma_{null} \). We set the \( \alpha \) level at 5% and applied Bonferroni corrections for exploratory analyses (\( \alpha_{adj} = 0.002 \)). We performed all analyses in R (version 4.2.1, 2022-06-23).

Our analyses can be split into 4 parts: the calculation of intraclass correlation coefficients (ICCs); separate models with all speech features as predictors and all affective scores as outcomes; combined models with all speech features as simultaneous predictors; and exploratory analyses, including additional speech features. Specifically, we first descriptively investigated whether our study procedure resulted in sufficient within-person variance. For this purpose, we calculated ICCs, including all momentary affective ratings and speech recordings, regardless of whether they were assessed before, during, or after SDT. In general, the ICC indicates the amount of between-person variance in unconditional (null) models. The 2-level models analyzed contained repeated measures (level 1) that were nested within patients (level 2). The second step contained our main analysis: we calculated separate models for each speech feature (pitch variability [model set 1], speech pauses [model set 2], and speech rate [model set 3]) and each affective state (depression severity [ADS-K], positive affect, negative affect, valence, energetic arousal, and calmness),
Correlational analyses (Figure S1 in Multimedia Appendix 1) included between 698 and 716 observations depending upon the specific pairing. We found correlations among and between affective scores and speech features, except for pitch variability and speech rate, neither of which correlated with negative affect and calmness; in addition, there was no correlation between pitch variability and speech rate. Specifically, ADS-K scores correlated negatively with positive affect, all MDMQ subscales, and speech rate and correlated positively with negative affect, pitch variability, and speech pauses. Negative affect showed the same pattern, except for the pairings with pitch variability and speech rate, for which no correlations were found. Regarding positive affect, we found the opposite correlation pattern, that is, positive correlations with all MDMQ subscales and speech rate and negative correlations with pitch variability and speech pauses. The MDMQ subscales showed the same relationships as positive affect, except for the pairing between calmness and pitch variability and speech rate, for which no correlations were found. Within speech features, we found a negative correlation between pitch variability and speech pauses, no correlation between pitch variability and speech rate, and a negative correlation between speech pauses and speech rate. Overall, correlations among affective scores were strong (>0.5). Correlations among speech features as well as between affective scores and speech features were weak (<0.2), except for a strong negative correlation between speech pauses and speech rate.

The psychometric properties for momentary affective ratings were good to excellent. Specifically, McDonald ω values [52] were 0.87 (within-person) and 0.90 (between-person) for depressive symptoms (ADS-K), 0.87 (within-person) and 0.95 (between-person) for positive affect, and 0.87 (within-person) and 0.96 (between-person) for negative affect. The Spearman-Brown coefficients were 0.83 (within-person) and 0.94 (between-person) for valence, 0.74 (within-person) and 0.89 (between-person) for energetic arousal, and 0.74 (within-person) and 0.89 (between-person) for calmness. Cronbach α for the MADRS score at the time of inclusion was acceptable (.67).

**Association Between Speech Features and Momentary Affective Scores**

**Overview**

In Tables 2 and 3, we present the fixed effects of pitch variability, speech pauses, and speech rate separately for each affective state. Details, including the effects of time and time², are presented in Table S1 in Multimedia Appendix 1.
Table 2. Multilevel linear regression analysis to predict depression and positive and negative affect: fixed effects of pitch variability, speech pauses, and speech rate.

<table>
<thead>
<tr>
<th>Predictors</th>
<th>Outcome</th>
<th>Positive affect</th>
<th>Negative affect</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>β</td>
<td>Standardized β</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model set 1</td>
<td></td>
<td>1.27</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td>Intercept</td>
<td>.88</td>
<td>.14</td>
</tr>
<tr>
<td></td>
<td>Pitch variability</td>
<td>.52</td>
<td>.10</td>
</tr>
<tr>
<td></td>
<td></td>
<td>−.11</td>
<td>−.10</td>
</tr>
<tr>
<td>Model set 2</td>
<td></td>
<td>1.27</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td>Intercept</td>
<td>.52</td>
<td>.10</td>
</tr>
<tr>
<td></td>
<td>Speech pauses</td>
<td>−.11</td>
<td>−.10</td>
</tr>
<tr>
<td></td>
<td></td>
<td>−.43</td>
<td>−.19</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4.69</td>
<td>.17</td>
</tr>
</tbody>
</table>

aADS-K: Allgemeine Depressionsskala.
bN/A: not applicable.

Table 3. Multilevel linear regression analysis to predict valence, energetic arousal, and calmness: fixed effects of pitch variability, speech pauses, and speech rate.

<table>
<thead>
<tr>
<th>Predictors</th>
<th>Outcome</th>
<th>Valence</th>
<th>Energetic arousal</th>
<th>Calmness</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>β</td>
<td>Standardized β</td>
<td>SE</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model set 1</td>
<td></td>
<td>43.72</td>
<td>N/A</td>
<td>2.70</td>
</tr>
<tr>
<td></td>
<td>Intercept</td>
<td>−36.50</td>
<td>−.16</td>
<td>13.61</td>
</tr>
<tr>
<td></td>
<td>Pitch variability</td>
<td>43.26</td>
<td>N/A</td>
<td>2.69</td>
</tr>
<tr>
<td></td>
<td></td>
<td>−34.06</td>
<td>−.19</td>
<td>7.71</td>
</tr>
<tr>
<td></td>
<td>Speech rate</td>
<td>43.56</td>
<td>N/A</td>
<td>2.70</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6.49</td>
<td>.17</td>
<td>2.03</td>
</tr>
</tbody>
</table>

aN/A: not applicable.

**ADS-K Scores**

In the column entitled **ADS-K** (Table 2), we report the results of all models with ADS-K scores as the outcome. **Pitch variability** (standardized $\beta$=.14; $P$=.007), **speech pauses** (standardized $\beta$=.10; $P$=.005), and **speech rate** (standardized $\beta$=.10; $P$=.02) were significantly associated with the ADS-K score, indicating that higher pitch variability, longer speech pauses, and lower speech rate are associated with more severe depressive symptomatology.

**Positive and Negative Affect**

In the columns entitled **Positive affect and Negative affect** (Table 2), we show results for positive affect and negative affect, respectively, as outcomes. **Pitch variability** (standardized $\beta$=.18; $P$<.001), **speech pauses** (standardized $\beta$=.17; $P$<.001), and **speech rate** (standardized $\beta$=.18; $P$<.001) were significantly associated with positive affect, indicating that lower pitch variability, shorter speech pauses, and higher speech rate are associated with higher positive affect. The associations between negative affect and speech features were in the opposite direction.
of the associations between positive affect and the speech features just presented: speech pauses (standardized $\beta$=0.09; $P=0.02$) and speech rate (standardized $\beta$=−0.08; $P=0.04$) were significantly associated with negative affect, indicating that longer speech pauses and lower speech rate are associated with higher negative affect. We further found a trend with respect to the association between pitch variability and negative affect, but this result was not statistically significant (standardized $\beta$=0.08; $P=0.05$). In addition, we found trends with respect to the associations between negative affect and time and negative affect and time$^2$, specifically in the models that included pitch variability (time: standardized $\beta$=−0.04; $P=0.08$), speech pauses (time: standardized $\beta$=0.04; $P=0.08$; time$^2$: standardized $\beta$<0.01; $P=0.06$), and speech rate (time: standardized $\beta$=0.04; $P=0.09$), but these results were not statistically significant.

**MDMQ Results**

In the columns entitled Valence, Energetic arousal, and Calmness (Table 3), we present the results for the MDMQ. Pitch variability (standardized $\beta$=−16; $P=0.008$), speech pauses (standardized $\beta$=−19; $P<0.001$), and speech rate (standardized $\beta$=−17; $P=0.001$) were significantly associated with valence, indicating that lower pitch variability, shorter speech pauses, and higher speech rate are associated with higher (ie, positive) valence. In the model that included valence and speech pauses, we found a significant association between time$^2$ and valence (standardized $\beta$<0.01; $P=0.03$). In addition, we found trends with respect to the associations between valence and time$^2$, specifically in the models that included pitch variability (time: standardized $\beta$<0.01; $P=0.098$) and speech rate (time: standardized $\beta$<0.01; $P=0.07$), but these results were not statistically significant. Moreover, pitch variability (standardized $\beta$=−15; $P<0.001$), speech pauses (standardized $\beta$=−08; $P=0.049$), and speech rate (standardized $\beta$=11; $P=0.03$) were significantly associated with energetic arousal, indicating that lower pitch variability, shorter speech pauses, and higher speech rate are associated with higher energetic arousal. In all model combinations of energetic arousal and each speech feature, we found significant associations between time and energetic arousal (standardized $\beta$=−11; $P<0.001$) and time$^2$ and energetic arousal (standardized $\beta$<0.01; $P<0.001$). Furthermore, speech pauses (standardized $\beta$=−12; $P<0.001$) were significantly associated with calmness, indicating that shorter speech pauses are associated with greater calmness. In all model combinations of calmness and each speech feature, we found significant associations between time$^2$ and calmness (standardized $\beta$<0.01; $P=0.013$ for pitch variability, $P=0.003$ for speech pauses; $P=0.009$ for speech rate). In addition, we found a trend with respect to the association between speech rate and calmness (standardized $\beta$=0.09; $P=0.07$), but this result was not statistically significant.

**Combined Models**

In Tables 4 and 5, we display the results for the combined models that included all 3 speech features. In the model of ADS-K scores, associations with pitch variability (standardized $\beta$=17; $P<0.001$) and speech pauses (standardized $\beta$=−12; $P=0.01$) remained statistically significant. Regarding positive affect, associations with pitch variability (standardized $\beta$=−23; $P<0.001$) and speech pauses (standardized $\beta$=−19; $P<0.001$) remained statistically significant. We further found a trend regarding the association between positive affect and time (standardized $\beta$=−05; $P=0.09$), but this result was not statistically significant. Regarding negative affect, associations with pitch variability (standardized $\beta$=12; $P=0.008$), speech pauses (standardized $\beta$=−12; $P=0.005$), time (standardized $\beta$=−05; $P=0.03$), and time$^2$ (standardized $\beta$<0.01; $P=0.03$) remained statistically significant. In the model of valence, associations with pitch variability (standardized $\beta$=−22; $P<0.001$), speech pauses (standardized $\beta$=−22; $P<0.001$), and time$^2$ (standardized $\beta$<0.01; $P=0.01$) remained statistically significant. Regarding energetic arousal, associations with pitch variability (standardized $\beta$=−17; $P=0.003$), time (standardized $\beta$=12; $P<0.001$), and time$^2$ (standardized $\beta$<0.01; $P<0.001$) remained statistically significant. Regarding calmness, associations with speech pauses (standardized $\beta$=−17; $P=0.002$) and time$^2$ (standardized $\beta$=−01; $P=0.002$) remained statistically significant. We further found a trend for the association between calmness and pitch variability (standardized $\beta$=0.09; $P=0.097$), but this result was not statistically significant.

**Table 4.** Multilevel linear regression analysis to predict momentary depression, positive affect, and negative affect: fixed effects of the combined models that included pitch variability, speech pauses, speech rate, time, and time$^2$.

<table>
<thead>
<tr>
<th>Predictors</th>
<th>Outcome</th>
<th>Positive affect</th>
<th>Negative affect</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ADS-K$^b$</td>
<td>$\beta$</td>
<td>Standardized $\beta$</td>
</tr>
<tr>
<td>Intercept</td>
<td></td>
<td>1.28</td>
<td>N/A$^c$</td>
</tr>
<tr>
<td>Time</td>
<td></td>
<td>&lt;.01</td>
<td>0.02</td>
</tr>
<tr>
<td>Time$^2$</td>
<td></td>
<td>&lt;.01</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Pitch variability</td>
<td></td>
<td>1.11</td>
<td>.17</td>
</tr>
<tr>
<td>Speech pauses</td>
<td></td>
<td>.64</td>
<td>.12</td>
</tr>
<tr>
<td>Speech rate</td>
<td></td>
<td>&lt;.01</td>
<td>&lt;.001</td>
</tr>
</tbody>
</table>

$^a$For ADS-K=2%, for positive affect=6%, and for negative affect=2%.

$^b$ADS-K: Allgemeine Depressionsskala.

$^c$N/A: not applicable.
were linked to higher positive affect, reduced negative affect, negative affect (standardized $\beta_{MFCC 4}$ of voiced regions only) we revealed a significant association between the SD of the change in the spectrum in unvoiced regions was linked to lower energetic arousal (standardized $\beta_{valence}$), this indicates that a faster change in the spectrum in voiced regions was linked to better affective states (ADS-K: standardized $\beta_{valence}=.34$; negative affect: standardized $\beta_{valence}=.28$; valence: standardized $\beta_{valence}=.29$; energetic arousal: standardized $\beta_{valence}=.26$; and calmness: standardized $\beta_{valence}=.19$); with respect to the mean of spectral flux, this indicates that a faster change in the spectrum was linked to better affective states (ADS-K: standardized $\beta_{valence}=.22$, positive affect: standardized $\beta_{valence}=.28$, negative affect: standardized $\beta_{valence}=.28$, valence: standardized $\beta_{valence}=.28$, energetic arousal: standardized $\beta_{valence}=.26$; and calmness: standardized $\beta_{valence}=.19$); and with respect to the mean of spectral flux of voiced regions only, this indicates that a faster change in the spectrum in voiced regions was linked to better affective states (ADS-K: standardized $\beta_{valence}=.23$, positive affect: standardized $\beta_{valence}=.28$, negative affect: standardized $\beta_{valence}=.15$, valence: standardized $\beta_{valence}=.20$, energetic arousal: standardized $\beta_{valence}=.20$, and calmness: standardized $\beta_{valence}=.16$). Regarding the additional speech features, the following significant associations were found: the mean of spectral flux of unvoiced regions only was associated with positive affect, indicating that a faster change in the spectrum in unvoiced regions was linked to improved positive affect (standardized $\beta=.13$); and the mean of the MFCC 2 of voiced regions only was significantly associated with energetic arousal, indicating that a higher mean was linked to lower energetic arousal (standardized $\beta=.15$). Furthermore, we revealed a significant association between the SD of the MFCC 4 of voiced regions only ADS-K scores (standardized $\beta=.13$) as well as positive affect (standardized $\beta=.10$) and negative affect (standardized $\beta=.09$). Specifically, smaller SDs were linked to higher positive affect, reduced negative affect, and lower ADS-K scores.

### Exploratory Analysis

Analyzing additional speech features, we found significant associations of the equivalent sound level, the mean of spectral flux, and the mean of spectral flux of voiced regions only, individually, with all affective scores (Table S2 in Multimedia Appendix 1). With respect to equivalent sound level, this indicates that louder voice samples were linked to improved affective states (ADS-K: standardized $\beta=-.30$; positive affect: standardized $\beta=-.34$; negative affect: standardized $\beta=-.21$; valence: standardized $\beta=-.29$; energetic arousal: standardized $\beta=-.26$; and calmness: standardized $\beta=-.19$); with respect to the mean of spectral flux, this indicates that a faster change in the spectrum was linked to better affective states (ADS-K: standardized $\beta=-.22$, positive affect: standardized $\beta=-.28$, negative affect: standardized $\beta=-.28$, valence: standardized $\beta=-.28$, energetic arousal: standardized $\beta=-.26$; and calmness: standardized $\beta=-.19$); and with respect to the mean of spectral flux of voiced regions only, this indicates that a faster change in the spectrum in voiced regions was linked to better affective states (ADS-K: standardized $\beta=-.23$, positive affect: standardized $\beta=-.28$, negative affect: standardized $\beta=-.15$, valence: standardized $\beta=-.20$, energetic arousal: standardized $\beta=-.20$, and calmness: standardized $\beta=-.16$). Regarding the additional speech features, the following significant associations were found: the mean of spectral flux of unvoiced regions only was associated with positive affect, indicating that a faster change in the spectrum in unvoiced regions was linked to improved positive affect (standardized $\beta=.13$); and the mean of the MFCC 2 of voiced regions only was significantly associated with energetic arousal, indicating that a higher mean was linked to lower energetic arousal (standardized $\beta=.15$). Furthermore, we revealed a significant association between the SD of the MFCC 4 of voiced regions only ADS-K scores (standardized $\beta=.13$) as well as positive affect (standardized $\beta=.10$) and negative affect (standardized $\beta=.09$). Specifically, smaller SDs were linked to higher positive affect, reduced negative affect, and lower ADS-K scores.

### Discussion

#### Principal Findings

This is the first study to investigate whether speech features are associated with depression severity and momentary affective states in a longitudinal data set of patients with a depressive episode undergoing SDT. Our findings showed that lower pitch variability, higher speech rate, and shorter speech pauses were associated with better momentary states (ie, lower depression severity; higher positive affect and lower negative affect; and higher positive valence, energetic arousal, and calmness), supporting prior clinical observations with innovative methods applied to an intensive longitudinal data set.

Lower depression severity was accompanied by shorter speech pauses. This is in line with past research findings reporting that shorter speech pauses were associated with lower depression severity [31-33,36]. Our findings extend prior results because we also found an association between speech pauses and affective states more broadly, not limited to depressed mood. Regarding speech rate, we revealed associations with depression severity and all other affective state scales except for calmness. In particular, we found that higher speech rate was associated with lower depression symptomatology and lower negative affect, higher positive affect, higher positive valence, and higher energetic arousal. This is in line with prior research [31-33,35], in which a higher speech rate was found for patients who benefited from treatment.

Regarding pitch variability, we found support for our hypothesis that pitch variability changes with depression severity; more precisely, lower pitch variability was associated with lower depression symptomatology. This is in line with the studies by Quatieri and Malyska [34] and Horwitz et al [35], where a positive correlation between pitch variability and depression severity was found. However, the results reported in the studies by Mundt et al [31] and Yang et al [36] contrasted with ours and those found in the studies by Quatieri and Malyska [34] and Horwitz et al [35], that is, that higher pitch variability was associated with lower depression severity. A possible explanation for contradictory results in major depression are...
the heterogeneity of (1) the depression phenotype per se because diagnosis criteria include >400 possible symptom combinations [53,54]; and (2) the questionnaires, assessment approaches, statistical analyses, and speech feature extraction tools used in these studies. The within-person research design approach underlying our data set addressed the heterogeneity of the depression phenotype at least partially. Furthermore, we analyzed free speech collected naturally in a selfie task, whereas in the study by Mundt et al [31], read speech was used in the analyses. In line with what is suggested in the study by Quatieri and Malyska [34], this could also be a reason for the contradictory results. However, because assessing within-person fluctuations in daily life increases ecological validity, we regard our results as an important contribution.

Observing the full picture of associations, we note that the results for all 3 speech features are similar and do not provide evidence of specific associations (eg, association of 1 specific speech feature with 1 specific momentary affective state), showing no distinct patterns of momentary states for each speech feature. This is reasonable because the constructs overlap in content (eg, patients experiencing depression experience higher negative affect and lower positive affect).

In terms of the combined models evaluating the relative importance of the features, we found that in the 4 models (ADS-K, valence, positive affect, and negative affect) both pitch variability and speech pauses remained significant, whereas speech rate did not. Pitch variability remained the only significant parameter in the model of energetic arousal, and speech pauses remained the only significant parameter in the model of calmness. This suggests that pitch variability and speech pauses are speech features rather independent of each other, whereas the high correlation between speech pauses and speech rate might account for the fact that only 1 of these features (in this case, speech pauses) remained a significant predictor.

Limitations

First, this study examined a limited set of 3 speech features. Instead of applying brute force methods involving thousands of technical speech features, we selected speech features based on previous work and with high face validity, restricting the scope of our analysis. Although we did expand our scope of features in the exploratory analysis, it is very likely that other configurations and features (eg, the ComParE feature set containing 6373 features [55]) might also be predictive of affective states. Future work is needed to compare theory-driven approaches with brute force data-driven machine learning methods to find the best possible combination of speech features also considering aspects of computational power. However, selecting the features on a theoretical basis and restricting their pure number limits alpha error inflation and should increase replicability.

Second, although the sample size of our study was limited, this was a true within-person design with many data points per patient. In addition, we regard this study as a pilot study providing important indications regarding feasibility in a clinical context. As some patients dropped out of the study, and some recordings had to be excluded, in future studies, data collection needs to be integrated better into clinical routines. Moreover, the instructions for patients may need to be revised to reduce the likelihood of missing data and recording errors. However, the data set at hand is still unique in the relatively high number of assessments per patient and the applied SDT, which yielded meaningful variation in the depression severity within a short time period. From a theoretical perspective, it is crucial to emphasize that to uncover existing relations among variables, meaningful variance in both parameters is needed.

Third and last, selfie videos were recorded in a clinical environment, which may limit generalizability to other contexts. In future studies, ambulant patients could be integrated and other environments explored to evaluate the replicability of the results. However, our approach, which involved sampling free speech, offers higher ecological validity to reading standardized text paragraphs because it provides a closer representation of people’s everyday lives. The development of passive sensing will be helpful in this context (ie, the random assessment of audio bits in an ecological environment). To date, automated passive voice recordings in nonprotected environments have been restricted in 2-party consent states, such as Germany. However, in single-party consent states, a few speech-related applications can be used in the wild (eg, the Electronically Activated Recorder [56]). Although the development of technical devices is ongoing, future studies will have to consider ethical issues related to voice recording in natural settings (eg, ensuring that no third parties who did not give informed consent are recorded).

Conclusions

Our study provides evidence that fluctuations in the speech features pitch variability, speech pauses, and speech rate are associated with fluctuations in depression severity and other momentary affective states. Notably, the data were collected from clinically diagnosed patients (no subclinical sample or staged emotions) experiencing an acute depressive episode. A particularly important advantage is that our longitudinal ambulatory assessment data set ensured a maximum of within-person dynamics of depressive parameters within a short time period by applying a sleep deprivation intervention design. This is of great importance because future technology will try to predict upcoming depressive episodes on an individual level and will need information on within-person trajectories. For the development of such tailored precision medicine tools, pitch variability, speech pauses, and speech rate present promising features. Our research is a step forward on the path to developing an automated depression monitoring system, facilitating individually tailored treatments and increased patient empowerment.
Acknowledgments
This study was funded by the Deutsche Forschungsgemeinschaft (DFG; German Research Foundation; GRK2739/1, project 447089431) Research Training Group: KDPSchool—Designing Adaptive Systems for Economic Decisions; the Network of European Funding for Neuroscience Research (ERA-NET NEURON): “Impact of Early Life Metabolic and Psychosocial Stress on Susceptibility to Mental Disorders; From Converging Epigenetic Signatures to Novel Targets for Therapeutic Intervention (EMBED)” (01EW1904); and the Bundesministerium für Bildung und Forschung (BMBF; Federal Ministry of Education and Research)—funded e:Med program Target-OXY (FKZ: 031L0190A). The authors acknowledge support by the Karlsruhe Institute of Technology Publication Fund.

Authors' Contributions
MR, JCF, JF, SHW, LS, MG, YY, and FS planned the investigation and developed the sampling scheme. MG and LS were responsible for data collection. L-MW preprocessed the data, carried out statistical analysis, and interpreted the results. UWE-P and MFL contributed to the analysis and interpretation of data. TS and AA contributed to acoustic analysis. L-MW drafted the manuscript with contributions from UWE-P and LS. All authors revised and edited the manuscript critically and had final approval of the version to be published.

Conflicts of Interest
UWE-P reports consultancy for Boehringer Ingelheim and speaker honorarium from Angelini Pharma, both of which had no influence over the content of this paper. All other authors declare no other conflicts of interest.

Multimedia Appendix 1
Supplementary material.

References


Abbreviations

**ADS-K:** Allgemeine Depressionsskala  
**eGeMAPS:** extended Geneva Minimalistic Acoustic Parameter Set  
**F0:** fundamental frequency  
**ICC:** intraclass correlation coefficient  
**ICD-10:** International Classification of Diseases, Tenth Revision  
**MADRS:** Montgomery–Åsberg Depression Rating Scale  
**MDMQ:** Multidimensional Mood Questionnaire  
**MFCC:** mel-frequency cepstral coefficient  
**openSMILE:** open-source Speech and Music Interpretation by Large-Space Extraction  
**SDT:** sleep deprivation therapy  
**SLEDGE II:** Sleep Deprivation and Gene Expression
Speech Features as Predictors of Momentary Depression Severity in Patients With Depressive Disorder Undergoing Sleep Deprivation Therapy: Ambulatory Assessment Pilot Study

https://mental.jmir.org/2024/1/e49222

doi:10.2196/49222

©Lisa-Marie Wadle, Ulrich W Ebner-Priemer, Jerome C Foo, Yoshiharu Yamamoto, Fabian Streit, Stephanie H Witt, Josef Frank, Lea Zillich, Matthias F Limberger, Ayimnisagul Ablimit, Tanja Schultz, Maria Gilles, Marcella Rietschel, Lea Sirignano. Originally published in JMIR Mental Health (https://mental.jmir.org), 18.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Mental Health, is properly cited. The complete bibliographic information, a link to the original publication on https://mental.jmir.org/, as well as this copyright and license information must be included.
**Abstract**

**Background:** The reasons for mental health consultations are becoming increasingly relevant in primary care. The Catalan health care system is undergoing a process of digital transformation, where eHealth is becoming increasingly relevant in routine clinical practice.

**Objective:** This study aimed to analyze the approach to depressive episodes and the role of eHealth in the Catalan health care system from 2017 to 2022.

**Methods:** A retrospective observational study was conducted on diagnostic codes related to depressive episodes and mood disorders between 2017 and 2022 using data from the Catalan Institute of Health. The sociodemographic evolution and prevalence of depression and mood disorders in Catalonia were analyzed between 2017 and 2022. Sociodemographic variables were analyzed using absolute frequency and percentage. The prevalence of depressive episodes was calculated, highlighting the year-to-year changes. The use of eHealth for related consultations was assessed by comparing the percentages of eHealth and face-to-face consultations. A comparison of sociodemographic variables based on attendance type was conducted. Additionally, a logistic regression model was used to explore factors influencing face-to-face attendance. The analysis used R software (version 4.2.1), with all differences examined using 95% CIs.

**Results:** From 2017 to 2022, there was an 86.6% increase in the prevalence of depression and mood disorders, with women consistently more affected (20,950/31,197, 67.2% in 2017 and 22,078/33,169, 66.6% in 2022). In 2022, a significant rise in depression diagnoses was observed in rural areas (difference 0.71%, 95% CI 0.04%-1.43%), contrasting with a significant decrease in urban settings (difference –0.7%, 95% CI –1.35% to –0.05%). There was a significant increase in antidepressant use in 2022 compared to 2017 (difference 2.4%, 95% CI 1.87%-3.06%) and the proportion of eHealth visits rose from 4.34% (1240/28,561) in 2017 to 26.3% (8501/32,267) in 2022. Logistic regression analysis indicated that men (odds ratio [OR] 1.06, 95% CI 1.04-1.09) and younger individuals had a higher likelihood of eHealth consultations in 2022. Furthermore, individuals using eHealth consultations were more likely to use antidepressants (OR 1.54, 95% CI 1.50-1.57) and anxiolytics (OR 1.06, 95% CI 1.03-1.09).

**Conclusions:** The prevalence of depression in Catalonia has significantly increased in the last 6 years, likely influenced by the COVID-19 pandemic. Despite ongoing digital transformation since 2011, eHealth usage remained limited as of 2017. During the
lockdown period, eHealth accounted for nearly half of all health care consultations, representing a quarter of consultations by 2022. In the immediate aftermath of the COVID-19 pandemic, emerging evidence suggests a significant role of eHealth in managing depression-related consultations, along with an apparent likelihood of patients being prescribed antidepressants and anxiolytics. Further research is needed to understand the long-term impact of eHealth on diagnostic practices and medication use.

**KEYWORDS**
eHealth; depression; depressive disorder; primary health care; mental health patient; patient; patients; healthcare system; digital transformation; mental disorder; mental disorders; diagnostic; clinical practice; clinical practices; retrospective; observational; regression; digital tool; digital tools

**Introduction**

Depression is a mental disorder that is a major health problem due to its high prevalence, direct repercussions on people's quality of life, and strong social impact [1,2]. According to the World Health Organization, it is estimated that 5% of the world's population suffers from depression [3].

Catalonia, an autonomous community of Spain, had a total of 7,747,709 citizens at the beginning of 2022 [4]. The Catalan health care system, characterized by being public, universal, comprehensive, and equitable, provides health coverage to 7.6 million people [5]. At the level of the Catalan territory, in 2015, according to data from the Catalonia health survey, 16.6% of the population older than 15 years suffered from anxiety or depression, which affected more women (20.8%) than men (12.2%) [6]. In 2020, 5 years later and taking into account the outbreak of the SARS-CoV-2 pandemic, an increase in emotional distress and moderate and major depression was observed. Specifically, 27.19% of women and 17% of men suffered from emotional distress, and 12.2% of women and 5.7% of men had moderate or major depression in Catalonia. Therefore, a decrease in the emotional well-being of the Catalan population has been observed [7].

Primary care (PC), one of the gateways to the health care system, is key in the detection, management, and follow-up of illnesses such as depression. High-quality PC is the foundation of a leading health care system and fundamental to optimizing the performance of the health care system, fulfilling the following 5 dimensions of the quintuple aim: (1) enhancing the care experience, (2) improving population health, (3) reducing costs, (4) care team well-being, and (5) advancing health equity [8-10]. PC has, among other characteristics, longitudinality and universal accessibility that make it the only area designed to be used by all people throughout their lives. As early as 1994, Starfield [11] showed how people living in countries with high-quality PC had better health indices, a more equitable distribution of available resources, and a more efficient health care system.

The reasons for mental health consultations are becoming increasingly relevant in PC, especially among the younger population in the wake of the COVID-19 pandemic [12,13]. Since there is no standardized protocol for mental health care in Catalonia, the approach to these issues varies significantly. This variability could be attributed to differences in professionals’ training or their sensitivity toward handling such problems [14]. Based on the report by the Public Health Agency of Catalonia regarding drug consumption with potential misuse, it was found that antidepressants were the most commonly used group of drugs by the population in 2021. Among most antidepressant drugs, the consumption ratio by women is between 2 to 4 times higher than that by men [15].

Over the last decade, information and communication technology is being introduced into the health care systems of different industrialized countries [16]. Following this trend, the Catalan health care system has been immersed in a digital transformation process. Currently, eHealth within PC in Catalonia involves communication both between providers and patients and amongst providers themselves. These interactions occur via different modes, including telephone or video calls and electronic consultations. It is important to emphasize that although the Catalan health care system had already incorporated technology before, the pandemic served as a clear catalyst for the adoption of eHealth in Catalonia. Regarding the typology of eHealth visits, the majority are conducted via remote consultations (referred to as eConsulta in Catalonia) or phone calls, with video calls being relatively infrequent [17].

Since 2011, several cross-cutting projects of significant importance have been developed in Catalonia, incorporating digital health as a key component [18]. The design of the Catalonia shared medical record began with the aim of being able to share patient information among the different health providers, as well as a personal folder (La Meva Salut) to give citizens access to their personal information, such as current medication plan, diagnoses, vaccines administered, clinical reports, test results, and examinations. Also, within this personal folder, in 2015, eConsulta was developed as an asynchronous digital communication tool involving health care professionals and patients, allowing the population to send queries at any time to their PC doctor or nurse and receive a response within a maximum of 48 hours on working days [19,20]. In this sense, eConsulta represents a significant advancement in eHealth within the Catalan public health system. Prior to the pandemic, eConsulta usage had a monthly growth rate of 7%; from March 15 to May 2020, an exponential growth rate was observed [21]. A study focusing on the profile of health care professionals who utilized remote consultations like eConsulta before the pandemic revealed that physicians engaging with eConsulta were typically aged between 45 and 64 years, exceeded the 80th percentile in the Quality of Care Index, had a high level of accessibility to...
their patients, participated in educational activities, and operated within a health team framework in urban settings with a high socioeconomic status [22].

In the context of Catalan PC, eHealth has been widely implemented and adopted in routine clinical practice, even more so in the context of COVID-19 [23-25]. However, there are other digital health tools with great potential, such as mobile health or artificial intelligence, that have not yet been introduced into the public health care system in Catalonia.

In order to improve the management of depressive episodes in PC in the Catalan health care system and to obtain higher effectiveness rates close to the potential efficacy of the available treatments, the aim of this study was to analyze the evolution of the prevalence of depression and mood disorders in PC from 2017 to 2022, to examine the sociodemographic profile of the affected population, and to investigate the role played by eHealth in this context and assess its impact on consultations related to depression and mood disorders.

Methods

Study Design
This study was a retrospective observational study of diagnostic codes related to depressive episodes and mood disorders between 2017 to 2022 from the Catalan Institute of Health.

Sample
We analyzed the entire population of Catalonia that visited PC centers of the Catalan Institute of Health who had a face-to-face or eHealth consultations (eConsulta, telephone, or video consultations) associated with the selected diagnoses of depression and mood disorders in the period of January 2017 to December 2022. The Catalan Institute of Health is the main provider of health services in the public system. It manages approximately 80% of the PC teams in Catalonia and provides health care coverage to approximately 5.8 million people [16,21,26,27].

To obtain the study sample, the diagnostic codes were selected according to those typified in the International Statistical Classification of Diseases and Related Health Problems (ICD-10), which is the classification used by PC professionals when recording a diagnosis in the computerized PC clinic [28]. All diagnostic codes related to depression (mild, moderate, and major) and mood disorders (eg, cyclothymia and dysthymia) that are usually detected, managed, or followed up in Catalan PC were selected (Table S1 in Multimedia Appendix 1). Mood disorders are marked by persistent clinical manifestations. In the case of dysthymia, it necessitates at least 2 years of a consistently low mood, whereas cyclothymia involves oscillations between a depressed mood and euphoria, without fulfilling the criteria for major depression [29]. Therefore, these diagnoses were included because they are not reactive disorders of short duration. By including them, there is a reduced risk of missing cases of mild depressive disorders.

Finally, we excluded all consultations that did not have an associated diagnosis code among those selected. The part of the population of Catalonia whose public health coverage was not provided by the Catalan Institute of Health was also excluded. The database was obtained through the Information System for the Development of Research in Primary Care [30].

Variables
To study sociodemographic characteristics, we considered age, sex, drug use, recurrent depressive episode diagnoses, rurality, and the socioeconomic situation recorded through the MedeA index [31]. A variable was created to distinguish between individuals with recurrent depressive episode diagnoses (where depression had been diagnosed previously in their medical history) and those without such recurrent diagnoses. The MedeA is a deprivation index linked to each census tract of the population. This assessment focuses on the barriers to access employment, education, culture, and social development. The aim is to evaluate them at a level that is considered acceptable within the surrounding society or region. The assessment comprises subindicators for employment and education. It is only available for urban areas, which are defined as municipalities with more than 10,000 inhabitants and a population density of more than 150 inhabitants per km². Other areas were considered rural. The MedeA index is ranked in quintiles, from MedeA urban 1 indicating low deprivation to MedeA urban 5 indicating high deprivation [31]. The MedeA index was used to categorize rurality by grouping urban and rural groups.

Statistical Analysis
To observe the evolution of the profile of people diagnosed with depression and mood disorders, sociodemographic variables were described in 2017 and 2022. Categorical variables were described by absolute frequency and percentage. The difference between the years was calculated with percentage points.

To observe the evolution of the prevalence of depressive episodes and mood disorders over the years for the total sample, the prevalent cases for each year were divided with respect to the total population assigned to the Catalan Institute of Health PC centers throughout Catalonia. The percentage change was calculated to determine the year-to-year evolution of these prevalences.

To observe the use of eHealth in consultations related to depression and mood disorders over the years, the percentage of telematic and face-to-face consultations resulting in new diagnoses per year was calculated.

A comparison of the sociodemographic variables as a function of face-to-face attendance was performed. Categorical variables were described by absolute frequency and percentage. The difference was calculated as percentage points.

Lastly, a logistic regression model was applied to observe how the studied variables affected face-to-face attendance. The model incorporated the variables of sex, age, rurality, recurrence, antidepressants, and anxiolytics. The MedeA variable was not introduced into this model since the rurality variable was used.
Analyses were performed with R version 4.2.1 (R Foundation for Statistical Computing). All differences were examined using CIs, and a confidence level of 95% was established.

**Ethics Approval**

The study protocol was approved by the University Institute for Primary Care Research (IDIAP) Jordi Gol Health Care Ethics Committee (Code 23/013-P).

---

**Results**

**Description of Sociodemographic Profiles of New Cases From 2017 to 2022**

An analysis of the sociodemographic profile of new cases during 2017 (n=31,197) and 2022 (n=33,169) was performed to observe changes in the characteristics of this population group during the study period (Table 1).
Table 1. Comparison of demographic characteristics of new cases in 2017 and 2022.

<table>
<thead>
<tr>
<th>Age (years)</th>
<th>2017 (n=31,197), n (%)</th>
<th>2022 (n=33,169), n (%)</th>
<th>Absolute difference (%) from 2017 to 2022 (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-15</td>
<td>689 (2.21)</td>
<td>922 (2.78)</td>
<td>0.57 (0.33 to 0.81)</td>
</tr>
<tr>
<td>16-24</td>
<td>1437 (4.61)</td>
<td>2415 (7.28)</td>
<td>2.67 (2.31 to 3.04)</td>
</tr>
<tr>
<td>25-34</td>
<td>2727 (8.74)</td>
<td>3712 (11.2)</td>
<td>2.46 (1.98 to 2.91)</td>
</tr>
<tr>
<td>35-44</td>
<td>4924 (15.8)</td>
<td>4765 (14.4)</td>
<td>−1.4 (−1.97 to −0.86)</td>
</tr>
<tr>
<td>45-54</td>
<td>5653 (18.1)</td>
<td>6101 (18.4)</td>
<td>0.3 (−0.32 to 0.87)</td>
</tr>
<tr>
<td>55-64</td>
<td>5355 (17.2)</td>
<td>5754 (17.3)</td>
<td>0.1 (−0.41 to 0.76)</td>
</tr>
<tr>
<td>65-74</td>
<td>4094 (13.1)</td>
<td>3860 (11.6)</td>
<td>−1.5 (−1.99 to −0.87)</td>
</tr>
<tr>
<td>75-84</td>
<td>4171 (13.4)</td>
<td>3670 (11.1)</td>
<td>−2.3 (−2.81 to −1.79)</td>
</tr>
<tr>
<td>≥85</td>
<td>2147 (6.88)</td>
<td>1970 (5.94)</td>
<td>−0.94 (−1.32 to −0.56)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gender</th>
<th>2017 (n=31,197), n (%)</th>
<th>2022 (n=33,169), n (%)</th>
<th>Absolute difference (%) from 2017 to 2022 (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Women</td>
<td>20,950 (67.2)</td>
<td>22,078 (66.6)</td>
<td>−0.6 (−1.30 to 0.15)</td>
</tr>
<tr>
<td>Men</td>
<td>10,222 (32.8)</td>
<td>11,057 (33.4)</td>
<td>__</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>MedeA index</th>
<th>2017 (n=31,197), n (%)</th>
<th>2022 (n=33,169), n (%)</th>
<th>Absolute difference (%) from 2017 to 2022 (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rural</td>
<td>7312 (25.6)</td>
<td>8499 (26.3)</td>
<td>0.71 (0.04 to 1.43)</td>
</tr>
<tr>
<td>Urban 1</td>
<td>6036 (21.1)</td>
<td>6592 (20.4)</td>
<td>−0.7 (−1.35 to −0.05)</td>
</tr>
<tr>
<td>Urban 2</td>
<td>4269 (14.94)</td>
<td>4806 (14.9)</td>
<td>0 (−0.62 to 0.51)</td>
</tr>
<tr>
<td>Urban 3</td>
<td>5644 (19.76)</td>
<td>6419 (19.9)</td>
<td>0.1 (−0.50 to 0.77)</td>
</tr>
<tr>
<td>Urban 4</td>
<td>5300 (18.55)</td>
<td>5951 (18.4)</td>
<td>−0.2 (−0.73 to 0.50)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Recurrent</th>
<th>2017 (n=31,197), n (%)</th>
<th>2022 (n=33,169), n (%)</th>
<th>Absolute difference (%) from 2017 to 2022 (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>29,144 (93.4)</td>
<td>29,967 (90.3)</td>
<td>−3.07 (−3.49 to −2.65)</td>
</tr>
<tr>
<td>Yes</td>
<td>2053 (6.58)</td>
<td>3202 (9.65)</td>
<td>3.07 (2.65 to 3.49)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Face-to-face</th>
<th>2017 (n=31,197), n (%)</th>
<th>2022 (n=33,169), n (%)</th>
<th>Absolute difference (%) from 2017 to 2022 (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>eHealth</td>
<td>1240 (4.34)</td>
<td>8501 (26.3)</td>
<td>22 (21.46 to 22.54)</td>
</tr>
<tr>
<td>Face-to-face</td>
<td>27,321 (95.7)</td>
<td>23,766 (73.7)</td>
<td>__</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Rurality</th>
<th>2017 (n=31,197), n (%)</th>
<th>2022 (n=33,169), n (%)</th>
<th>Absolute difference (%) from 2017 to 2022 (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rural</td>
<td>7312 (25.6)</td>
<td>8499 (26.3)</td>
<td>0.7 (0.04 to 1.43)</td>
</tr>
<tr>
<td>Urban</td>
<td>21,249 (74.4)</td>
<td>23,768 (73.7)</td>
<td>__</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Antidepressants</th>
<th>2017 (n=31,197), n (%)</th>
<th>2022 (n=33,169), n (%)</th>
<th>Absolute difference (%) from 2017 to 2022 (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>25,353 (81.3)</td>
<td>27,777 (83.7)</td>
<td>2.4 (1.87 to 3.06)</td>
</tr>
<tr>
<td>No</td>
<td>5844 (18.7)</td>
<td>5392 (16.3)</td>
<td>__</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Anxiolytics</th>
<th>2017 (n=31,197), n (%)</th>
<th>2022 (n=33,169), n (%)</th>
<th>Absolute difference (%) from 2017 to 2022 (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>15,615 (50.1)</td>
<td>16,519 (49.8)</td>
<td>−0.3 (−1.02 to 052)</td>
</tr>
<tr>
<td>No</td>
<td>15,582 (49.9)</td>
<td>16,650 (50.2)</td>
<td>__</td>
</tr>
</tbody>
</table>

*aStatistically significant difference.

*bGiven the symmetry of the binary variables, the difference and 95% CI have only been expressed for one of the categories.

The results showed that in both 2017 and 2022, the 45-54 years age range had the most diagnoses related to depressive episodes and mood disorders. However, in 2022 there was a significant increase compared to 2017 in diagnoses in the younger age ranges (0-15 years: difference 0.57%, 95% CI 0.33%-0.81%; 16-24 years: difference 2.67%, 95% CI 2.31%-3.04%; 25-34 years: difference 2.46%, 95% CI 1.98%-2.91%). A significant reduction in diagnoses was also observed in the older age ranges.
Regarding gender, it was observed that in both 2017 and 2022, women were the most affected gender (20,950/31,172, 67.2% and 22,078/33,135, 66.6%, respectively). Regarding the MedeA index, a significant increase in diagnoses was observed in rural settings in 2022 with respect to 2017 (difference 0.71%, 95% CI 0.04%-1.43%), while a significant decrease was observed in urban settings, specifically in the urban population with low deprivation (difference –0.7%, 95% CI –1.35% to –0.05%).

Regarding recurrent depressive episodes, there was a pronounced and significant increase in those diagnosed as recurrent compared to those diagnosed with a depressive episode for the first time (difference 3.07%, 95% CI 2.65%-3.49%).

Regarding the use of eHealth for managing these illnesses, a notable shift was observed in the pattern of face-to-face attendance for the management of depressive episodes and mood disorders. A significant increase was observed in the use of eHealth as the approach for these diagnostic codes (difference 22%, 95% CI 21.46%-22.54%). Finally, regarding the most common medication used for depressive episodes and mood disorders, a significant increase in the use of antidepressants was observed in 2022 compared to 2017 (difference 2.4%, 95% CI 1.87%-3.06%), while no differences were observed in the use of anxiolytics.

**Evolution of the Prevalence of Depressive Episodes and Mood Disorders by Year**

The prevalence per year of all selected diagnostic codes was calculated during the period from 2017 to 2022, considering the total population assigned to the Catalan Institute of Health Primary Care Centers throughout Catalonia (Figure 1). Since 2017, an increase in diagnoses was observed. The steepest increase was in 2018, which went from 2.3% (95% CI 2.27%-2.29%) during 2017 to 2.9% (95% CI 2.86%-2.88%) in 2018. From 2019 to 2022, a progressive increase in prevalence was observed, reaching 4.3% (95% CI 4.24%-4.27%) in 2022. In short, the prevalence of depressive episodes and mood disorders increased by 86.6% during the study period. The results can be seen in Table S2 in Multimedia Appendix 1.

**Use of eHealth for Consultations Related to Depression and Mood Disorders by Year**

The percentage of face-to-face and eHealth visits for each year of new diagnoses was calculated to analyze the use of eHealth for consultations related to depression and mood disorders in the public health care system (Figure 2). It was noted that in 2017 eHealth consultations represented 4.34% (1,240/28,561) of all depressive episode and mood disorder visits, with an increase during 2018 and 2019. As of 2020, coinciding with the pandemic lockdown, eHealth consultations accounted for 46.8% (11,122/23,790), representing a more than 6-fold increase over the previous year. In 2021, there was some recovery of face-to-face consultations, but nevertheless, eHealth consultations still accounted for 39.7% (12,119/30,561) of total consultations. Finally, in 2022, eHealth consultations accounted
for 26.3% (8501/32,267), representing just over a quarter of all consultations for diagnoses of depression and mood disorders.

**Figure 2.** Percentage of telematic and face-to-face consultations from 2017 to 2022.

### Comparison of Prevalent Cases in 2022 According to Face-to-Face Attendance

A comparative analysis of the sociodemographic variables of the prevalent 2022 cases was performed according to face-to-face consultations (Table 2). Regarding age, it was observed that the populations that used eHealth consultations the most were those aged 0 to 15 years (difference −1.18%, 95% CI −1.36% to −0.98%), 16 to 24 years (difference −0.64%, 95% CI −0.89% to 0.38%), and those aged 85 years and older (difference −4.37%, 95% CI −4.68% to −4.05). In contrast, in 2022, it was observed that the populations that used face-to-face consultations the most were people aged 35 to 44 years (difference 0.7%, 95% CI 0.33%-1.10%), 45 to 54 years (difference 0.7%, 95% CI 0.27%-1.11%), 55 to 64 years (difference 0.8%, 95% CI 0.43%-1.24%), 65 to 74 years (difference 3.1%, 95% CI 2.84%-3.53%), and 75 to 84 years (difference 0.59%, 95% CI 0.59%-1.31%).

Significant differences were also observed in the MedeA index. People living in urban settings with less deprivation (urban 1) made more use of eHealth (difference −1.9%, 95% CI −2.35% to −1.44%), while those with higher deprivation in urban settings (urban 3 and urban 4) used face-to-face consultations more often (difference 0.9%, 95% CI 0.48%-1.34% and difference 2.4%, 95% CI 2.01%-2.84%, respectively). Regarding recurrence, it was observed that individuals with recurrent depressive episodes sought consultation more frequently via eHealth (difference −2.46%, 95% CI −2.77% to −2.15%). In terms of rurality, individuals living in rural areas demonstrated a higher frequency of eHealth usage (difference −1.4%, 95% CI −1.94% to −0.97%). Finally, individuals receiving treatment with antidepressants or anxiolytics showed a higher tendency to use eHealth services for their health consultations (difference −10.1%, 95% CI −10.59% to −9.50% and difference −4%, 95% CI −4.53% to −3.51%, respectively).
Table 2. Comparative table of prevalent cases in 2022 according to attendance type.

<table>
<thead>
<tr>
<th>Age (years)</th>
<th>Non–face-to-face (n=38,719), n (%)</th>
<th>Face-to-face (n=188,598), n (%)</th>
<th>Absolute difference (%) between attendance types (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-15</td>
<td>1246 (3.22)</td>
<td>3850 (2.04)</td>
<td>−1.18 (−1.36 to −0.98)(^a)</td>
</tr>
<tr>
<td>16-24</td>
<td>2219 (5.73)</td>
<td>9599 (5.09)</td>
<td>−0.64 (−0.89 to −0.38)(^a)</td>
</tr>
<tr>
<td>25-34</td>
<td>3615 (9.34)</td>
<td>17,240 (9.14)</td>
<td>−0.2 (−0.51 to 0.12)(^a)</td>
</tr>
<tr>
<td>35-44</td>
<td>5616 (14.5)</td>
<td>28,711 (15.2)</td>
<td>0.7 (0.33 to 1.10)(^a)</td>
</tr>
<tr>
<td>45-54</td>
<td>6859 (17.7)</td>
<td>34,716 (18.4)</td>
<td>0.7 (0.27 to 1.11)(^a)</td>
</tr>
<tr>
<td>55-64</td>
<td>6388 (16.5)</td>
<td>32,688 (17.3)</td>
<td>0.8 (0.43 to 1.24)(^a)</td>
</tr>
<tr>
<td>65-74</td>
<td>4086 (10.6)</td>
<td>25,913 (13.7)</td>
<td>3.1 (2.84 to 3.53)(^a)</td>
</tr>
<tr>
<td>75-84</td>
<td>4742 (12.2)</td>
<td>24,894 (13.2)</td>
<td>1 (0.59 to 1.31)(^a)</td>
</tr>
<tr>
<td>≥85</td>
<td>3948 (10.2)</td>
<td>10,987 (5.83)</td>
<td>−4.37 (−4.69 to −4.05)(^a)</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Women</td>
<td>25,802 (66.7)</td>
<td>126,652 (67.2)</td>
<td>0.5 (0.00 to 1.03)(^b)</td>
</tr>
<tr>
<td>Men</td>
<td>12,886 (33.3)</td>
<td>61,832 (32.8)</td>
<td></td>
</tr>
<tr>
<td>MedeA index</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rural</td>
<td>10,418 (26.9)</td>
<td>48,003 (25.5)</td>
<td>−1.4 (−1.94 to −0.97)(^a)</td>
</tr>
<tr>
<td>Urban 1</td>
<td>8687 (22.4)</td>
<td>38,726 (20.5)</td>
<td>−1.9 (−2.35 to −1.44)(^a)</td>
</tr>
<tr>
<td>Urban 2</td>
<td>5838 (15.1)</td>
<td>28,469 (15.1)</td>
<td>0 (−0.37 to 0.41)(^a)</td>
</tr>
<tr>
<td>Urban 3</td>
<td>7363 (19)</td>
<td>37,588 (19.9)</td>
<td>0.9 (0.48 to 1.34)(^a)</td>
</tr>
<tr>
<td>Urban 4</td>
<td>6413 (16.6)</td>
<td>35,812 (19)</td>
<td>2.4 (2.01 to 2.84)(^a)</td>
</tr>
<tr>
<td>Recurrent</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>35,209 (90.9)</td>
<td>176,141 (93.39)</td>
<td>—</td>
</tr>
<tr>
<td>Yes</td>
<td>3510 (9.07)</td>
<td>12,457 (6.61)</td>
<td>−2.46 (−2.77 to −2.15)(^a)</td>
</tr>
<tr>
<td>Rurality</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rural</td>
<td>10,418 (26.9)</td>
<td>48,003 (25.5)</td>
<td>−1.4 (−1.94 to −0.97)(^a)</td>
</tr>
<tr>
<td>Urban</td>
<td>28,301 (73.1)</td>
<td>140,595 (74.5)</td>
<td>—</td>
</tr>
<tr>
<td>Antidepressants</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>20,464 (52.9)</td>
<td>80,734 (42.8)</td>
<td>−10.1 (−10.59 to −9.50)(^a)</td>
</tr>
<tr>
<td>No</td>
<td>18,255 (47.1)</td>
<td>107,864 (57.2)</td>
<td>—</td>
</tr>
<tr>
<td>Anxiolytics</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>12,819 (33.1)</td>
<td>54,855 (29.1)</td>
<td>−4 (−4.53 to −3.51)(^a)</td>
</tr>
<tr>
<td>No</td>
<td>25,900 (66.9)</td>
<td>133,743 (70.9)</td>
<td>—</td>
</tr>
</tbody>
</table>

\(^a\)Statistically significant differences.

\(^b\)Given the symmetry of the binary variables, the difference and CI have only been expressed for one of the categories.

Finally, a logistic regression model was used to observe the impact of the studied variables from the 2022 attendance data (Figure 3). Men were 1.06 times more likely (95% CI 1.04–1.09) to use eHealth in consultations than women, and those suffering from recurrent depressive episodes were 1.37 times more likely (95% CI 1.32–1.43) to use eHealth in consultations. It was also observed that the youngest age ranges (0-15 years: odds ratio [OR] 1.98, 95% CI 1.85-2.13; 16-24 years: OR 1.27, 95% CI 1.21-1.34; 25-34 years: OR 1.12, 95% CI 1.08-1.18) were more likely to use eHealth in consultations, as was also the case among those aged 85 and older (OR 1.99, 95% CI 1.90-2.08). In contrast, people living in urban settings were 7% less likely (OR 0.93, 95% CI 0.91-0.95) to have an eHealth consultation. Finally, people who consulted more through eHealth were 1.54
times (95% CI 1.50-1.57) more likely to take an antidepressant and 1.06 times (95% CI 1.03-1.09) more likely to take an anxiolytic. The results can be seen in Table S3 in Multimedia Appendix 1.

**Figure 3.** Logistic regression results. The reference categories were as follows: women, aged 45-54, nonrecurrent, rural, no antidepressants, and no anxiolytics. The black squares indicate the odds ratio estimate, and the horizontal lines indicate the 95% CI.

**Discussion**

**Principal Findings**

The aim of the present study was to analyze the context of depression within the Catalan health care system. Consequently, the prevalence of depression and mood disorders has been successfully analyzed over the years, along with the sociodemographic profile of this population. Additionally, the role played by eHealth and its impact on PC for these conditions has been studied.

Regarding the prevalence of depression and mood disorders in Catalan PC, an increase of 86.6% was observed during the study period. One possible reason explaining this increase in prevalence could be an uptick in the detection of these cases within the public health care system, possibly indicating increased awareness among both professionals and citizens. In 2006 in Catalonia, the Primary Care Support Programme was implemented to coordinate PC with mental health services, providing a budget to hire more psychiatrists and psychologists in order to reinforce PC [32,33]. However, it was not until 2017 that implementation was completed in all PC centers in Catalonia. In this context, despite the existence of critical reports regarding the operation of this support program, its initiation might have positively influenced professionals’ awareness, potentially contributing to an increase in detections in PC since then. The results of the study revealed a more pronounced increase (25.82%) in 2018, coinciding with the end of this program’s operation.

Another possible reason explaining the increase in prevalence is the COVID-19 pandemic. The data analyzed cover the period from before the pandemic to the early postpandemic period, so the results may indeed be influenced by the consequences of lockdown and the isolation of individuals. In this context, the pandemic has generated great concern and an even more marked increase in mental health awareness, especially among the younger population. As the context of social isolation has strongly impacted the mental health of young people, as it is a crucial period for the development of social skills, it is noteworthy that from the prepandemic context in 2019, the overall prevalence increased by 36.52% in 2022 [34,35]. However, in line with the statistics already published, it is worth mentioning that women are the most affected group [6,7]. Over the past decade, the importance of sex- and gender-stratified biomedical research has grown significantly, as demonstrated by the observations of Oertelt-Prigione et al [36] of an increasing number of publications in medicine addressing this issue. It is now crucial to incorporate this perspective into the clinical approach to depression. Understanding how individuals’ behaviors align with social expectations, coping strategies, and help-seeking tendencies can contribute to a more comprehensive approach to treatment. Regarding the medicalization of these conditions, it has been observed that approximately 80% of new
diagnoses of depression and mood disorders in both 2017 and 2022 were prescribed antidepressant drugs. That is, most people who were diagnosed for the first time were associated with antidepressants. In contrast, the prevalence of antidepressant use has been found to be around 40% to 50% in the entire prevalent population. The results suggest that although most people who are diagnosed with depression or mood disorders are prescribed medication, in most cases treatment is withdrawn, as indicated by clinical practice guidelines [37,38].

Regarding the use of eHealth for consultations on depression and mood disorders in the PC context in Catalonia, it has been observed that the use of eHealth is increasingly being integrated into daily clinical practices in Catalonia [18]. In fact, while in 2017 diagnoses made through eHealth consultations accounted for only 4.34% of consultations, in 2022 they accounted for 26.34%. Without a doubt, the pandemic has also been a clear catalyst in the use of eHealth. Although the volume of eHealth consultations in 2020 was not equivalent to that in 2022, the results from the latter year may suggest a shift in the pattern of consultations related to depression and mood disorders. In this sense, eHealth is gaining prominence in routine practice, as has been suggested by other studies conducted in Catalonia [17,39,40]. Nevertheless, the incorporation of eHealth into the health care system must be accompanied by training for the professionals who must use it. In other words, it is not just about providing infrastructure and resources for technology operation. It is also crucial to understand when the use of these tools aligns with the care process and equips the health care professional to make clinical and therapeutic decisions appropriately, as they hold responsibility for the procedure and its outcomes [41]. The results of the present study have also shown that one of the impacts of eHealth on depressive episodes and mood disorders is on the prescription of drugs. People treated with eHealth were prescribed 54% more antidepressant drugs and 6% more anxiolytics. A systematic review by Han et al [42] observed the impact of eHealth on antibiotic prescribing in PC. Although it was not possible to conclude with sufficient confidence that eHealth significantly impacts the prescribing of these drugs, 4 of the 12 studies analyzed reported higher prescribing rates through eHealth [42]. In a study conducted by Wabe et al [43] during the pandemic in Australia's PC, they compared the prescription of medication through face-to-face consultation versus eHealth methods. The findings revealed that prescribing through face-to-face consultations was more prevalent for all drug groups classified in the World Health Organization Anatomical Therapeutic Chemical Classification list [44], except for the group that includes nervous system drugs, such as antidepressants and anxiolytics. In the context of PC in Catalonia, it would be interesting to examine whether the results obtained related to the increased prescription of drugs via eHealth can be applied to other medical conditions. In this regard, the development strategies used in the approach to eHealth will be crucial to ensure consistency in clinical practice.

**Strengths and Limitations**

The most relevant strength of this study is the size of the sample, since it was possible to work with the consultations related to depressive episodes and mood disorders of almost 5.8 million people. As a result, this study has yielded robust and realistic findings regarding the prevalence of depression, the use of eHealth, and the associated implications. Another strength of this study lies in the comprehensive analysis of the role and implications of eHealth in depression-related consultations. It not only highlights the current utilization of eHealth in depression consultations but also underscores the need to develop strategies that promote consistency between in-person and eHealth visits.

The main limitation of this analysis is that the data were collected only from the Catalan Institute of Health within Catalonia’s public health care system. As a result, there is a lack of information from individuals who have sought treatment through private health care providers. Another limitation is that, since this was a cross-sectional study, it was not possible to follow up on each individual user, and therefore, it was not possible to establish a causal relationship between the variables studied, although the relationship between them was observed. However, due to the absence of subdivision in the original database regarding telematic consultations, specifically whether they were synchronous or asynchronous, a detailed breakdown of the proportions of telematic consultations could not be provided. Furthermore, it is essential to acknowledge the limitation that the data analyzed may reflect the impact of the lockdown imposed during de COVID-19 pandemic. The pandemic led to the isolation of individuals, which, in turn, resulted in a notable increase in diagnoses through eHealth and prescriptions of antidepressants and anxiolytics when comparing the period before COVID-19 to the postlockdown situation. In this context, the often-mandatory use of eHealth consultations during the lockdown, along with the isolation of the population, may have contributed to the increase in the detection rates and medication usage reflected in the results. Therefore, it would be valuable to replicate this study in the coming years to reassess the prevalence and sociodemographic characteristics of depression as well as the role of eHealth in addressing the observed diagnostic and pharmacological trends. This follow-up study would enable us to determine whether the findings observed in this study represent short-term consequences of COVID-19 or if they have more lasting implications over time.

**Conclusions**

The prevalence of depression in Catalonia has significantly increased from 2017 to 2022, with the pandemic likely having a profound impact. Although the Catalan health care system has been undergoing digital transformation since 2011, eHealth usage was limited in 2017. During the lockdown, it accounted for nearly half of the care, and by 2022 it represented a quarter of the consultations. In the short post-COVID-19 era, evidence suggests that eHealth plays a role in consultations related to depression and appears to increase the likelihood of taking antidepressants and anxiolytics. However, future studies in a context further removed from the pandemic could provide a more accurate indication of the role of eHealth in diagnosing and treating depression and mood disorders.
**Acknowledgments**

We would like to thank Manuel Moreno for extracting the study variables from the original database. This project was supported by the Department of Health of the Generalitat de Catalunya and received funding in the 2021 grant call from the Strategic Plan for Research and Innovation in Health (PERIS) for the 2022-2024 period. The project falls under the research projects focused on primary care, with the file code SLT021/21/000002.

**Data Availability**

The data sets generated and analyzed during the current study are not publicly available because the manuscript was based on confidential and sensitive health data but are available from the corresponding author upon reasonable request.

**Conflicts of Interest**

None declared.

Multimedia Appendix 1

Selected diagnostic codes, depression prevalence, and logistic regression. [DOCX File, 27 KB - mental_v11i1e52816_appl1.docx ]

**References**

1. Depression. World Health Organization. 2021. URL: [https://www.who.int/health-topics/depression#tab=tab_3](https://www.who.int/health-topics/depression#tab=tab_3) [accessed 2022-11-13]
3. Depressive disorder (depression). World Health Organization. 2021. URL: [https://www.who.int/news-room/fact-sheets/detail/depression](https://www.who.int/news-room/fact-sheets/detail/depression) [accessed 2022-11-03]


**Abbreviations**

- **ICD-10**: International Statistical Classification of Diseases and Related Health Problems
- **OR**: odds ratio
- **PC**: primary care
Incorporating a Stepped Care Approach Into Internet-Based Cognitive Behavioral Therapy for Depression: Randomized Controlled Trial

Jasleen Kaur Jagayat1*, MSc; Anchan Kumar2, MD; Yijia Shao2,3, MPS; Amrita Pannu2, MD; Charmy Patel2, MPH, MACP; Amirhossein Shirazi2,4, MD, PhD; Mohsen Omrami2,4, MD, PhD; Nazanin Alavi1,2*, MD

1Centre for Neuroscience Studies, Queen’s University, Kingston, ON, Canada
2Department of Psychiatry, Queen’s University, Kingston, ON, Canada
3Pastoral Studies, Toronto School of Theology (TST), Emmanuel College, University of Toronto, Toronto, ON, Canada
4Online Psychotherapy Tool, Toronto, ON, Canada

these authors contributed equally

Corresponding Author:
Jasleen Kaur Jagayat, MSc
Centre for Neuroscience Studies
Queen’s University
18 Stuart Street
Kingston, ON, K7L 3N6
Canada
Phone: 1 613 533 2600
Email: jasleenjagayat@gmail.com

Abstract

Background: Depression is a hidden burden, yet it is a leading cause of disability worldwide. Despite the adverse effects of depression, fewer than one-third of patients receive care. Internet-based cognitive behavioral therapy (i-CBT) is an effective treatment for depression, and combining i-CBT with supervised care could make the therapy scalable and effective. A stepped care model is a framework for beginning treatment with an effective and low-intensity intervention while adapting care based on the patient’s needs.

Objective: This study investigated the efficacy of a stepped care i-CBT model for depression based on changes in self-reported depressive symptoms.

Methods: In this single-blinded, randomized controlled trial, participants were allocated to either the i-CBT–only group (28/56, 50%) or the i-CBT with stepped care group (28/56, 50%). Both groups received a 13-week i-CBT program tailored for depression. The i-CBT program was provided through a secure, online mental health clinic called the Online Psychotherapy Tool. Participants read through the sessions and completed the assignments related to each session. Participants in the stepped care group received additional interventions from their care provider based on standard questionnaire scores (ie, Patient Health Questionnaire–9 [PHQ-9], Quick Inventory of Depressive Symptomatology [QIDS], and Quality of Life Enjoyment and Satisfaction Questionnaire–Short Form) and their assignment responses. From lowest to highest intensity, the additional interventions included SMS text messages, phone calls, video calls, or a video call with a psychiatrist.

Results: For this study, 56 participants were recruited to complete an i-CBT program (n=28, 50%; mean age 37.9; SD 13.08 y; 7/28, 27% were men) or an i-CBT with stepped care program (n=28, 50%; mean age 40.6; SD 14.28 y; 11/28, 42% were men). The results of this study indicate that the i-CBT program was effective in significantly reducing depressive symptoms, as measured by the PHQ-9 (F4,80=9.95; P<.001) and QIDS (F2,28=5.73; P=.008); however, there were no significant differences in the reduction of depressive symptoms between the 2 groups (PHQ-9: F4,80=0.43; P=.78; QIDS: F2,28=3.05; P=.06). The stepped care group was not significantly better in reducing depressive symptoms than the i-CBT group (PHQ-9, F2,28=0.43; P=.78; QIDS, F2,28=3.05; P=.06). Although there were no significant differences observed between the number of participants who completed the program between the groups (X2 =2.6; P=.10), participants in the stepped care group, on average, participated in more sessions than those who prematurely terminated participation in the i-CBT group (t53=−2; P=.03; 95% CI −4.83 to −0.002).
Conclusions: Implementing a stepped care approach in i-CBT is an effective treatment for depression, and the stepped care model can assist patients to complete more sessions in their treatment.

Trial Registration: Clinicaltrials.gov NCT04747873; https://clinicaltrials.gov/study/NCT04747873

(KEYWORDS
internet-based cognitive behavioral therapy; i-CBT; major depressive disorder; MDD; stepped care; digital mental health care; mobile phone

Introduction

Background and Rationale
Depression is a leading cause of disability, affecting approximately 3.8% of the population worldwide [1,2]. Major depressive disorder (MDD) is characterized by persistent feelings of sadness, negative mood, or loss of interest in life activities [3]. Detrimental and persistent changes in appetite, sleep, energy, and cognition may accompany these feelings. In addition to its deleterious impacts on mental health, depression is associated with increased morbidity, decreased quality of life, and reduced work productivity [4-6]. Despite the negative consequences of depression, only one-third of individuals receive treatment, and of those, only 3 in 5 people receive sufficient care [7,8].

Cognitive behavioral therapy (CBT) is an effective treatment for MDD [9-11]. It is a form of psychotherapy that focuses on cognitive restructuring strategies and behavioral activation techniques to help individuals with depression overcome their depressive symptoms and modify ineffective thinking patterns. Although effective, CBT on its own presents some challenges such as accessibility issues, lack of follow-ups, and increased costs [12-14]. Over the past 2 decades, issues of accessibility are being increasingly addressed through digital modalities, such as internet-based CBT (i-CBT) [11,15,16]. i-CBT allows for the delivery of CBT through digital media including computers and smartphones. It involves using web-based sessions with interactive components and practices to teach individuals ways to improve their mental health. Many studies have shown the efficacy of i-CBT for depression and anxiety, with results comparable with those of traditional in-person CBT [11,16-20]. The digital format allows for increased accessibility and convenience when delivering and receiving psychotherapy while remaining effective [21,22]; however, the digital format comes with limitations, including low adherence. Adherence rates are variable and can be affected by multiple factors including the digital modalities used and patient-therapist interactions [23-25]. Furthermore, attrition rates in CBT are the highest in patients with depression [26], and adherence can be compromised as disease severity increases [27]. With the benefit of being versatile and adaptable, i-CBT may be adapted to include support to address the issues of adherence.

The stepped care model is an approach to providing individuals with adaptable and effective care that is adjusted based on symptom severity [28-31]. This dynamic approach allows care to be adapted as required throughout an individual’s treatment plan to best tailor care to the individual’s needs. Stepped care models aim to provide the most effective and efficient care by minimizing resources and costs for care. The least intensive intervention, such as self-help materials or psychoeducation, is first provided to patients in a stepped care model. The patient may then advance to the next level of care, which may include more extensive interventions such as increased support or medication if they do not improve or just partially improve. Until the patient gets the desired result or reaches the highest level of care, this process is repeated. This stepped care model allows for efficient allocation of resources, increases access to care, can reduce wait times, can lower costs, and can increase versatility in treatment options [28,31]. This model is effective in treating mental health disorders including MDD and focuses on a patient-centered approach to care [21,31].

In the stepped care model, all patients receive some form of treatment for their symptoms, and based on the symptom severity and progress through treatment, additional care is added to their treatment plan. The addition of therapist-guided support to i-CBT has demonstrated high levels of patient satisfaction and a general decrease in symptom severity [32]. It is also noted that dropout rates decrease as increased therapeutic support is provided to patients. A meta-analysis that gathered dropout rates of CBT based on treatment support levels found that without support, the dropout rate was 74%; with administrative support, the dropout rate was 38%; and with therapeutic support, the patient dropout rate was 28% [16]. This suggests that the type of support provided to patients is important. Therapist-guided support in an i-CBT setting can also include text (ie, messaging and emails) or audio-based (ie, telephone and video) formats, and the mode of therapy may influence treatment outcome [22,33,34]. Incorporating text and call interventions into online therapy can provide an additional support outlet for patients throughout their treatment by providing an additional means of communication while remaining accessible [35]. Phone and video interventions have both been shown to be effective in reducing symptoms and maintaining patients in care compared with in-person care for mood disorders [36]. Phone calls may be more broadly accessible than video calls owing to inherent technological barriers; however, video calls allow for additional nonverbal cues to be assessed in care [36]. They can serve as check-ins, encouragement, and reminders and help monitor patient progress between treatment sessions. Through this additional support, a review of therapeutic concepts and words of encouragement are found to help decrease symptom severity and improve therapy completion [37]. This method can help with building therapist-patient rapport throughout treatment and act as a valuable tool in online therapy.
Objectives
This study was a single-blinded, randomized controlled trial exploring the efficacy of a stepped care model in an i-CBT program compared with an i-CBT–only program for adults with MDD. The i-CBT program for depression was provided through a secure, online mental health clinic called the Online Psychotherapy Tool (OPTT) [38]. The stepped care model considered the varying digital modalities of text-based, phone-based, and video-based interventions to help support individuals in their care.

Methods

Ethical Considerations
The research study was approved by the Health Sciences and Affiliated Teaching Hospitals Research Ethics Board (file number 6031992) at Queen’s University in Kingston, Ontario, Canada. The research study followed the CONSORT-EHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth) guidelines and can be viewed in Multimedia Appendix 1 [39].

No monetary compensation was provided to participants for participation in this study because they were provided with CBT treatment for their symptoms.

Data Privacy
To maintain participant confidentiality, each participant was assigned a randomized participant ID number. This ID was used to identify the participant for data analysis purposes. Only associated care providers had access to the participant’s identity owing to the nature of the CBT intervention. If any technical issues on OPTT arose, this ID was used to resolve the issues to maintain participant anonymity.

Participant information maintained online was password protected, including consent forms, identities, and CBT content. All encrypted files were stored on a safe server run by Queen’s University. Participant identities and consent forms are stored on-site at Queen’s University’s exclusive storage in Kingston, Ontario, Canada, for 5 years following the conclusion of the study because they are considered as medical records. Following the 5 years, the participant records will be destroyed. Participants were informed about the possibility to withdraw from the study at any time if they wished to do so. In upcoming plans for knowledge dissemination and publication of outcomes, participant identification will be secured and maintained anonymously.

The online platform used to deliver the i-CBT program, OPTT [38,40], serves as the repository for all data. OPTT complies with the Personal Information Protection and Electronic Documents Act, Health Insurance Portability and Accountability Act, and Service Organization Control–2. The cloud infrastructure of Amazon Web Service Canada is used to host all servers and databases, and Medstack manages it to ensure compliance with all the local, state, and federal privacy and security laws. For privacy reasons, OPTT does not gather any identifying personal data or IP addresses. Only anonymous metadata were gathered by OPTT to enhance the quality of its services and provide the care provider team access to participant analytics (ie, interaction with the OPTT platform). No OPTT employee had direct access to participant data owing to data encryption procedures (ie, participant ID). All encrypted backups are stored at Queen’s University.

Outcome Evaluation

Patient Health Questionnaire–9
The Patient Health Questionnaire–9 (PHQ-9) is a self-assessment questionnaire designed to examine a participant’s depression severity through a 9-item questionnaire, with an additional question about functional health [41]. Each of the 9 items is scored from 0 (not at all) to 3 (nearly every day) and corresponds to the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition criteria for MDD symptoms [3]. Each item is then scored for a total to indicate depression severity from none to severe, ranging from 0 to 27. High scores indicate high depression.

Quick Inventory of Depressive Symptomatology Questionnaire
The Quick Inventory of Depressive Symptomatology (QIDS) questionnaire is a self-assessment questionnaire designed to screen for depression and measure depression severity based on 16 items [42]. Each item correlates with the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition criteria for MDD symptoms and is scored using a 4-point scale (0–3), for a total score ranging between 0 and 27 [3]. To calculate the final score, the questions are summed using the highest response in the following domains: sleep patterns (questions 1–4), sad mood (questions 5), change in appetite (questions 6–9), concentration and decision-making (questions 10), self-view (questions 11), suicidal thoughts (questions 12), general interest (questions 13), energy level (questions 14), and psychomotor effects (questions 15 and 16). High scores indicate high depression severity.

Quality of Life Enjoyment and Satisfaction Questionnaire–Short Form
The Quality of Life Enjoyment and Satisfaction Questionnaire–Short Form (Q-LES-Q) is a self-assessment questionnaire designed to collect information about the level of enjoyment and satisfaction in various aspects of daily functioning through a series of 16 items [43]. Each item is rated on a scale from 1 (very poor) to 5 (very good). Of the 16 items, only the first 14 items are summed for a total raw score ranging from 14 to 70. High scores indicate great quality of life.

Measured Outcomes
The primary outcomes measured in this study included changes in depressive symptoms, as measured by the PHQ-9 and QIDS. Using the Q-LES-Q to measure changes in the participant’s quality of life was also a primary outcome that was evaluated. Participants completed the PHQ-9 every 3 weeks (weeks 1, 4, 7, 10, and 13) and the QIDS and Q-LES-Q at 3 time points (weeks 1, 7, and 13) throughout the program. All the 3 questionnaires are intended to be completed at the 3-, 6-, 9-, and 12-month follow-up periods during the 1-year follow-up...
phase (Textbox 1). Pretreatment and posttreatment changes in depressive symptoms (ie, PHQ-9 and QIDS) and quality of life (Q-LES-Q) between the control and experimental groups were compared. The number of sessions completed by each participant was a secondary measure that was considered when evaluating compliance with the i-CBT program between the 2 groups.
Textbox 1. Schedule of the assigned clinical questionnaires sent to participants, including the 1-year follow-up period.

<table>
<thead>
<tr>
<th>Session</th>
<th>Questionnaires</th>
</tr>
</thead>
<tbody>
<tr>
<td>Session 1</td>
<td>Patient Health Questionnaire–9 (PHQ-9), Quality of Life Enjoyment and Satisfaction Questionnaire–Short Form (QIDS), Quick Inventory of Depressive Symptomatology (Q-LES-Q), and demographic survey</td>
</tr>
<tr>
<td>Session 2</td>
<td>No questionnaires are provided during these sessions.</td>
</tr>
<tr>
<td>Session 3</td>
<td>No questionnaires are provided during these sessions.</td>
</tr>
<tr>
<td>Session 4</td>
<td>PHQ-9 and initial stepped care assessment (if the participant is in the stepped care group, an intervention decision will be made at the end of every session, starting at session 4. Interventions will be implemented, beginning from session 5)</td>
</tr>
<tr>
<td>Session 5</td>
<td>Interventions begin as needed (if the participant is in the stepped care group, an intervention decision will be made at the end of every session, starting at session 4; Interventions will be implemented, beginning from session 5)</td>
</tr>
<tr>
<td>Session 6</td>
<td>No questionnaires are provided during these sessions.</td>
</tr>
<tr>
<td>Session 7</td>
<td>PHQ-9, QIDS, and Q-LES-Q</td>
</tr>
<tr>
<td>Session 8</td>
<td>No questionnaires are provided during these sessions.</td>
</tr>
<tr>
<td>Session 9</td>
<td>No questionnaires are provided during these sessions.</td>
</tr>
<tr>
<td>Session 10</td>
<td>PHQ-9</td>
</tr>
<tr>
<td>Session 11</td>
<td>No questionnaires are provided during these sessions.</td>
</tr>
<tr>
<td>Session 12</td>
<td>No questionnaires are provided during these sessions.</td>
</tr>
<tr>
<td>Session 13</td>
<td>PHQ-9, QIDS, and Q-LES-Q</td>
</tr>
<tr>
<td>Follow-up 1</td>
<td>PHQ-9, QIDS, and Q-LES-Q</td>
</tr>
<tr>
<td>Follow-up 2</td>
<td>PHQ-9, QIDS, and Q-LES-Q</td>
</tr>
<tr>
<td>Follow-up 3</td>
<td>PHQ-9, QIDS, and Q-LES-Q</td>
</tr>
<tr>
<td>Follow-up 4</td>
<td>PHQ-9, QIDS, and Q-LES-Q</td>
</tr>
</tbody>
</table>
Participants
This study was registered at ClinicalTrial.gov (NCT04747873). Participants were recruited from the Providence Care outpatient psychiatry clinic and Kingston Health Sciences Center sites (Hotel Dieu Hospital and Kingston General Hospital), both of which are located in Kingston, Ontario, Canada. Physicians familiar with the Queen’s Online Psychotherapy Lab (QUOPL) research team were also informed about the study and directed patients who may benefit from CBT toward the study when considered appropriate. Self-referrals were also accepted for this study. Recruitment was managed by the laboratory manager who was the initial point of contact for all participants.

Eligibility Criteria
Individuals interested in the research study were provided with a letter of information and consent form to understand the study design before beginning the eligibility process. Following the informed consent process (written or verbal), a trained research assistant at QUOPL screened the individual based on the eligibility criteria. Individuals were eligible for the study if they met the following criteria: (1) aged ≥18 years, (2) met the criteria of MDD by a trained research assistant according to the Mini International Neuropsychiatric Interview (MINI) [44], (3) showed competence to consent to participate, (4) fluent in English because the i-CBT program was provided in English only, and (5) had consistent and reliable access to the internet. For the MINI assessment, individuals were first assessed by a trained research assistant to support a diagnosis of MDD. This MINI assessment was completed through a secure Microsoft Teams video call. Participants were ineligible for the research study if they presented with active psychosis, acute mania, severe alcohol or substance use disorder, or active suicidal or homicidal ideation. If a participant received or was receiving CBT in the past year at the time of beginning the study, they were excluded from the study to avoid confounding effects on the efficacy of this i-CBT program.

Following the consent and eligibility screening process, participants were randomized in a 1:1 allocation ratio to either the i-CBT group (control group) or the i-CBT with stepped care group (experimental group). Randomization was computer generated in Microsoft Excel using the RANDBETWEEN function. This method returned an integer at random that indicated either 1 (designated as the i-CBT group) or 2 (for the stepped care group). This integer was used to determine the group allocation for the participant.

Using G*Power (version 3.1.9.7 [45]), a priori power analysis was performed to determine the minimum sample size required to test the study hypothesis. The average PHQ-9 score decreased from 16.2 before i-CBT to 11.48 (combined SD 5.45) after 12 sessions in our previous clinical trials and data collection regarding i-CBT for depression [46]. These figures led to an effect size (Hedge g of 0.86). A paired sample t test (1-tailed) would require 14 individuals to detect a significant impact, given the effect size and a power of 0.8. For online CBT programs, compliance and care adherence are common challenges; therefore, we predict 50% dropout rate based on previous clinical trials conducted in our laboratory [47,48]. Consequently, considering the proposed sample size calculation accounted for this dropout rate, we aimed to recruit 28 participants in each group. With the 2 treatment arms, our total sample size was expected to be 56 participants. Figure 1 summarizes the study design in a flow chart.

Figure 1. A flowchart providing an overview of the study design. i-CBT: internet-based cognitive behavioral therapy; MINI: Mini International Neuropsychiatric Interview; RCT: randomized controlled trial.
Participants were assigned a primary care provider to help build rapport through assigning sessions, providing feedback, and providing additional interventions as needed. All care providers were trained in psychotherapy by the lead psychiatrist who is a licensed psychotherapist and expert in internet-based psychotherapy. Each participant's care team consisted of a care provider and a psychiatrist with extensive experience in i-CBT. The care providers were research assistants recruited and trained by the team psychiatrist. Their role was to provide participants with feedback about weekly homework and act as a point of contact for any questions that the participants may have regarding the study design. The care providers included master’s degree students studying in the field of neuroscience and psychology, as well as research assistants who have completed a medical degree. Through training, all care providers were well experienced in internet-based psychotherapy, with a specific focus on CBT techniques. They were all taught the standard care pathway and the aim, and content of each therapeutic session. They also continued receiving specialized training through webinars, CBT workshops led by a psychiatrist trained in CBT, and exercises with feedback during the study. Moreover, they were provided predesigned feedback templates tailored to each session to be used when writing the weekly feedback. The feedback templates helped standardize the feedback by providing a basic structure that ensured all aspects of the homework were acknowledged and that the newly presented CBT concepts were reviewed. Feedback templates varied between sessions, and care providers personalized each template for each participant’s homework. The use of feedback templates that were drafted by the team psychiatrist trained in online psychotherapy allowed some control over treatment consistency across sessions and participants. Before submission to the participant, feedback was always reviewed by another care provider (MSc candidate) trained in psychotherapy and overlooked by a psychiatrist to ensure the quality of the feedback. The team psychiatrist fulfilled the role to overlook the feedback; manage any crisis events if necessary; and provide support in the stepped care model, specifically the highest-intensity intervention, step 4.

**Intervention**

### i-CBT Program

All participants (56/56, 100%) were provided with the same 13-week i-CBT program designed for MDD through OPTT—a secure, cloud-based, digital mental health platform [38]. This i-CBT program, titled “Electronic Cognitive Behavioural Therapy (e-CBT) For Depression,” is a set of predesigned modules created to address depressive symptoms through various CBT techniques to work on cognitive restructuring [46,49-51]. This specific program has undergone validation through previous studies and has been shown to significantly decrease depressive symptoms [47,49]. The program was intended to imitate in-person, standard CBT for MDD in an asynchronous format. In this study, 1 module was assigned to participants on the same day of each week. A module included 20 to 30 slides of CBT content and took approximately 45 to 50 minutes to view (Multimedia Appendix 2). Each module focused on 1 CBT technique (Table 1). Each module followed a similar outline, beginning with an introduction to the topic for the week, an overview of the technique with examples, and concluding with homework related to the weekly technique [50]. The homework consisted of a few questions based on a stressful event that occurred in the participant’s life in the past week and provided participants with the practice of the new CBT technique. Participants were provided with 4 days to asynchronously review the module and submit their homework to their assigned care provider on the OPTT platform. Participant answers were then reviewed by the assigned care provider over 3 days who provided participants with feedback about their answers along with their next session on the assigned day of the week.
Table 1. A brief description of each session for the internet-based cognitive behavioral therapy (i-CBT) program, titled “Electronic Cognitive Behavioural Therapy (e-CBT) For Depression,” on the Online Psychotherapy Tool platform.

<table>
<thead>
<tr>
<th>Module</th>
<th>Title</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>What is depression?</td>
<td>• Introduced i-CBT and discussed common depressive symptoms while setting expectations for the course</td>
</tr>
<tr>
<td>2</td>
<td>The 5-Part Model</td>
<td>• Described the 5-Part Model: connections and interactions between a situation, thoughts, feelings, physical reactions, and behaviors</td>
</tr>
<tr>
<td>3</td>
<td>Sleep hygiene</td>
<td>• Focused on how to improve rest through sleep habits and offered a variety of techniques</td>
</tr>
<tr>
<td>4</td>
<td>Strategies for stressful situations</td>
<td>• Provided a general review of practical strategies that can be applied under difficult circumstances, including breathing techniques and activities</td>
</tr>
<tr>
<td>5</td>
<td>Thoughts, feelings, behavior, physical reactions, and environment</td>
<td>• Explained the 5-Part Model in detail and how modifications to 1 part can influence the remaining 4 parts</td>
</tr>
<tr>
<td>6</td>
<td>The thought record</td>
<td>• Featured the first 3 columns of the thought record: a tool for challenging ineffective thoughts and the “hot thought” • The first 3 columns included the situation, feelings, and automatic thoughts associated with the situation • Common cognitive errors were also discussed</td>
</tr>
<tr>
<td>7</td>
<td>Automatic thoughts</td>
<td>• Explored the function of automatic thoughts and how they affect emotions • This included learning how to recognize automatic thoughts and the “hot thought” • Common cognitive errors were also discussed</td>
</tr>
<tr>
<td>8</td>
<td>Activity scheduling</td>
<td>• Described the activity record: a tool for recording weekly activities and finding connections between activities and associated moods</td>
</tr>
<tr>
<td>9</td>
<td>Evidence</td>
<td>• Back to the thought record, the fourth and fifth columns were explained, which included looking at evidence for and against the most intense thought</td>
</tr>
<tr>
<td>10</td>
<td>Alternative and balanced thinking</td>
<td>• Concluded the 5-Part Model with the final 2 columns outlining alternative or balanced thinking and rerating feelings</td>
</tr>
<tr>
<td>11</td>
<td>Experiments</td>
<td>• Introduced a behavioral activation technique of experiments to promote belief in alternative or balanced thinking</td>
</tr>
<tr>
<td>12</td>
<td>Action plans</td>
<td>• Promoted working on the identified problems using a structured tool called the action plan</td>
</tr>
<tr>
<td>13</td>
<td>Review</td>
<td>• Reviewed the 12 modules in the program and summarized the key tools and techniques</td>
</tr>
</tbody>
</table>

The feedback was based on predesigned feedback templates created by the lead psychiatrist and i-CBT expert at QUOPL. The feedback for each session followed a similar structure and was delivered as a letter addressed to the participant. The feedback template mirrored the following structure: (1) addressed the participant and thanked them for their work over the week while acknowledging any adverse events that occurred; (2) commented about their mood and sleep quality in the previous week; (3) summarized the content of the module, highlighting the main concepts; (4) summarized the participant’s answers; (5) empathized with the participant based on their shared experiences and encouraged them to use the learned CBT techniques; and (6) thanked them again for their time and provided the participant with a brief introduction about the next module’s content (Multimedia Appendix 3—sample feedback template for session 1). Overall, the feedback itself focused on the participant’s mood and sleep patterns over the week, progress on their weekly goals, and their understanding of CBT concepts. The care providers personalized the feedback template based on the participant’s responses.

The OPTT platform provided participants and care providers with a modality to communicate asynchronously regarding the availability of the next module and any questions or concerns regarding the OPTT operations or the study design. These questions and concerns were limited to the study design and technical issues, rather than personal or therapy-related concerns. Care providers viewed and replied to these OPTT messages at least once a week. For concerns regarding the OPTT platform and technical issues, participants were redirected to OPTT’s technical support team.

**i-CBT With Stepped Care**

Although all participants (56/56, 100%) had access to the i-CBT program, a designated care provider, and weekly feedback, only participants in the stepped care group received additional interventions as per the proposed stepped care model. The stepped care interventions were provided by their care provider or the team psychiatrist associated with their care as required. The structure of the stepped care model used in this study followed the 4 identified steps of a preventative stepped care model based on 10 randomized controlled trials [52]. These
four steps included the following: (1) watchful waiting, (2) self-help psychotherapy, (3) face-to-face psychotherapy, and (4) referral to specialists. The intervention that participants received was dictated by the participant’s care team: the care provider and the team psychiatrist. This model initiated all participants at the lowest intervention of i-CBT beginning at session 1 and provided additional interventions beginning at session 5 after monitoring their progress through the first 4 sessions and obtaining 2 PHQ-9 scores as a quantitative assessment measure, to help with the decision (Textbox 1). Subsequently, an intervention could be added or changed in each session by the care team, following session 5. Beginning stepped care at session 5 allowed for a watchful waiting period as the initial step to understand the patient’s symptoms better, as seen in preventative stepped care models for anxiety and depressive disorders [52]. The first 3 interventions, steps 1 to 3, are similar in content and care provider, but the delivery modality varies. These first 3 steps included varying intensities to care and effort required by the care provider. Step 4 differed from the first 3 steps in that the team psychiatrist was involved and the content of care in this intervention was different. Details of each step in the intervention are provided in the following sections.

Step 0: i-CBT Only

This intervention was considered to have the lowest intensity because the participant was provided with an i-CBT program for depression that is effective in treating MDD [49]. This was also known as the starting point for all participants. All participants began at step 0, which included completing the i-CBT program. In this step, no intervention was added to the participants’ care, which allows for patient monitoring during a watchful waiting phase as they receive a low-intensity treatment [52]. Furthermore, many studies have control groups that encompass waitlist groups; however, more studies should focus on implementing i-CBT treatments as the control group for a wholesome review of the effects of stepped care models placed in these i-CBT programs. Therefore, step 0 mimics the procedure for the control group to allow for comparable effects between the 2 groups.

Step 1: i-CBT With Messaging

This intervention was the first added intervention following step 0 and included the addition of asynchronous check-in messages to the participant from their care provider on the OPTTT platform. When the care provider sent the next session, they also sent a personalized message based on the message templates for that weekly session (Multimedia Appendix 3—sample message template). The message focused on addressing the participant and checking in with them by asking them how their week was so far. This intervention was a brief exchange between the participant and their care provider and a way to add active human support to the participant’s care through direct messaging [53,54]. Upon receiving the participant’s response, care providers acknowledged their experiences and reminded them to complete their next session.

Step 2: i-CBT With a Telephone Call

This intervention included a brief telephone call from the care provider for live, verbal support [55-57]. The call is limited to one 15 to 20–minute call during the week [58]. After sending the weekly session, care providers called the participant before the next session. The telephone call focused on asking participants how their week was and acknowledging the participant’s experience (Multimedia Appendix 3—sample call template). Care providers also asked for updates regarding module completion for the week and reminded participants to submit their weekly homework. This template was similar to that of step-1 messages, except that it included direct verbal encouragement.

Step 3: i-CBT With a Video Call

This intervention included a brief, secure, Microsoft Teams video call from the care provider for live support and visual contact between the participant and care provider, in addition to verbal cues. The video call is limited to one 20 to 30–minute call during the week [22,59]. Care providers set up a video call with the participant before the next session. The video call focused on the participant’s experience over the past week and how they are doing (Multimedia Appendix 3—sample message template). Care providers also encouraged participants to submit their weekly homework if they had not submitted yet. This template was similar to that of step-2 telephone calls and included direct, live encouragement in hopes to mimic a face-to-face setting.

Step 4: i-CBT With Psychiatrist Call

This intervention was the highest-intensity intervention provided in the stepped care model. This intervention included a web-based, one-on-one psychiatrist appointment with the team psychiatrist using Microsoft Teams. The psychiatrist discussed the participant’s current challenges and potential options for the participant. This may include the possibility of adding medication to the participant’s care based on the severity of the case. The selection of medication was based on the Canadian Network for Mood and Anxiety Treatments guidelines, and these guidelines were referenced to decide the first or second line of treatment for each participant [60].

When deciding the appropriate intervention, a participant’s care was not always increased sequentially; instead, a participant who exhibited severe symptoms by the end of session 4 was provided a high step as opposed to step 1. The decision about which intervention the participant would receive was dependent on a few factors considered by the care providers. These included (1) changes in participant’s PHQ-9 scores, (2) engagement with treatment, (3) progress in weekly goals, and (4) homework submission.

Changes in PHQ-9 Scores

PHQ-9 scores for each participant were collected every 3 weeks (ie, weeks 1, 4, 7, 10, and 13). The initial step-up decision was based on the first 2 PHQ-9 scores collected at weeks 1 and 4. If the participant’s PHQ-9 score at week 4 increased by >2 points compared with week 1, the participant was stepped up in their care to either step 1 (message) or step 2 (phone call). During the subsequent weeks, changes in the subsequent PHQ-9
scores were compared with the previous week’s score (ie, increase of >2 points) to determine whether an additional intervention was required [61].

Treatment Engagement
A systematic review including 35 studies showed that great treatment engagement significantly improved postintervention mental health outcomes [62]. Thus, participant engagement with treatment through the OPTT platform was another factor that was considered when determining which intervention would be most suitable for the participant. If the participant exhibited limited interaction with the platform (eg, read receipts for messages indicating that messages were not seen when notifying participants about session availability) or expressed difficulties in navigating the platform, the participant was provided with the step-2 (phone call) intervention. This limited engagement indicated that step 1 (messages) would be ineffective because the participant was not viewing the intervention. Therefore, the step-2 (phone call) intervention was identified as a more effective modality than the step-1 (messages) intervention. In addition, if participants did not respond to step-1 (messages) intervention, step-2 (phone call) interventions were provided as a follow-up.

Homework Submission
Participants’ answers to weekly homework were also considered as a factor when deciding their care intensity [63]. If participants indicated ongoing depressive symptoms that have not changed within the past 2 sessions, they were stepped up in care intensity. Similarly, if depressive symptoms did not improve within the past 2 weeks and if they did not find the current intervention of steps 1, 2, or 3 helpful, step 4 (psychiatrist support) was provided to the participant. Finally, if the care provider sensed any form of suicidal or homicidal ideation or any severe depressive behavior at any point, step 4 (psychiatrist support) was referred to the participant.

Goal Progress
Beginning in session 1, participants were encouraged to set a goal that they would like to achieve by the end of the program. Each week, participants were prompted to complete a small step toward the goal to help them achieve it. The weekly progress toward this goal rather than absolute achievement was used as another indication of whether a step-up in intervention was required [64]. Studies have shown some support for goal planning in positive treatment outcomes for mental health care through assisting the therapeutic relationship by building rapport and allowing for open communication [65,66]. If participants struggled with the progression in their weekly small step for >2 consecutive sessions, they were stepped up in their intervention.

Data Analysis
A combination of descriptive statistics, independent t tests, and ANOVA were used to determine any difference between the primary outcomes across the 2 groups. All analyses were performed at a 1-tailed significance level of α=.05. To determine any significant difference in the clinical questionnaire scores between the 2 groups, ANOVA was used. A 2×5 repeated-measures ANOVA was conducted for PHQ-9 scores, and a 2×3 repeated-measures ANOVA was performed for Q-LES-Q and QIDS scores. The time points at which meaningful differences appear were examined using the Bonferroni post hoc method. If Mauchly test of sphericity was significant, a Huynh-Feldt correction was applied. A chi-square test was conducted to compare treatment compliance across the 2 groups based on the number of participants completing all 13 sessions of the program. An independent samples t test was also performed to compare the number of sessions completed by participants in the 2 groups. In addition, an intention-to-treat (ITT) analysis was completed to assess the clinical outcomes of treatment on participants who withdrew prematurely. No intermediate analyses were performed, and all statistical analyses were conducted after the trial. IBM SPSS Statistics (version 28) was used to conduct all analyses [67].

Results
Participants
Recruitment was initiated in May 2021 after receiving approval from the Queen’s University Health Sciences and Affiliated Teaching Hospitals Research Ethics Board. Recruitment was conducted between May 2021 and June 2022, and 69 individuals were found to be eligible for this study. Table 2 displays the participant demographics. Of the 69 eligible individuals, 34 (49%) were randomized to the control group and only 28 (41%) initiated treatment. Of the 28 participants in the control group, 9 (32%) completed the full round of therapy (ie, 13 sessions of i-CBT). Of the 35 individuals randomized to the stepped care group, 28 (80%) initiated treatment. In the stepped care group, only 54% (15/28) of the participants completed all 13 sessions of the i-CBT program. Unfortunately, owing to some data collection errors, some questionnaire scores were not collected during the program, leading to some gaps in pretreatment, midtreatment, or posttreatment scores. These scores were treated as missing and were not imputed during data analysis. Figure 2 shows the participant flow.

Participants in the stepped care group were allocated to the stepped care intervention as follows: step 0 (8/28, 29%), step 1 (message; 2/28, 7%), step 2 (phone call; 15/28, 54%), step 3 (video call; 1/28, 4%), step 4 (psychiatrist consultation; 2/28, 7%). Overall, of the 28 participants, a total of 20 (71%) participants were stepped up in their care. This includes participants who dropped out prematurely and those who completed all 13 sessions. Of the 15 participants who completed all 13 sessions, 12 (80%) participants were stepped up in their care at some point during the program. Table 3 provides information regarding the number of occurrences of each step, and Table 4 provides a summary of each participant’s progression in the stepped care program throughout the treatment, beginning at session 5. Currently, the 1-year follow-up period is ongoing, in which PHQ-9, QIDS, and Q-LES-Q scores are collected at the 3-, 6-, 9-, and 12-month follow-up periods after the treatment. The follow-up period is expected to be completed in November 2023.
Table 2. Demographics and characteristics of the participants who began treatment, categorized based on treatment group: internet-based cognitive behavioral therapy (i-CBT) and i-CBT with stepped care.

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>i-CBT (n=28)</th>
<th>Stepped care (n=28)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (y), mean (SD)</td>
<td>37.88 (13.08)</td>
<td>40.57 (14.28)</td>
</tr>
<tr>
<td>Baseline PHQ-9 score, mean (SD)</td>
<td>16.63 (4.40)</td>
<td>17.75 (5.33)</td>
</tr>
<tr>
<td><strong>Sex, n (%)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>17 (61)</td>
<td>17 (61)</td>
</tr>
<tr>
<td>Male</td>
<td>7 (25)</td>
<td>11 (39)</td>
</tr>
<tr>
<td>Other</td>
<td>2 (7)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Missing</td>
<td>2 (7)</td>
<td>0 (0)</td>
</tr>
<tr>
<td><strong>Ethnicity, n (%)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asian</td>
<td>1 (4)</td>
<td>1 (4)</td>
</tr>
<tr>
<td>Hispanic</td>
<td>1 (4)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>White</td>
<td>7 (25)</td>
<td>7 (25)</td>
</tr>
<tr>
<td>Other</td>
<td>17 (61)</td>
<td>20 (71)</td>
</tr>
<tr>
<td>Missing</td>
<td>2 (7)</td>
<td>0 (0)</td>
</tr>
<tr>
<td><strong>First language, n (%)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>English</td>
<td>24 (86)</td>
<td>26 (93)</td>
</tr>
<tr>
<td>Hebrew</td>
<td>0 (0)</td>
<td>1 (4)</td>
</tr>
<tr>
<td>Hindi</td>
<td>0 (0)</td>
<td>1 (4)</td>
</tr>
<tr>
<td>Cantonese</td>
<td>1 (4)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Spanish</td>
<td>1 (4)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Missing</td>
<td>2 (7)</td>
<td>0 (0)</td>
</tr>
<tr>
<td><strong>Immigration status, n (%)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Born in Canada</td>
<td>12 (43)</td>
<td>13 (46)</td>
</tr>
<tr>
<td>Immigrated to Canada</td>
<td>3 (11)</td>
<td>2 (7)</td>
</tr>
<tr>
<td>Missing</td>
<td>13 (46)</td>
<td>13 (46)</td>
</tr>
<tr>
<td><strong>Employment, n (%)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Full time</td>
<td>13 (46)</td>
<td>15 (54)</td>
</tr>
<tr>
<td>Part time</td>
<td>3 (11)</td>
<td>4 (14)</td>
</tr>
<tr>
<td>Unemployed</td>
<td>6 (21)</td>
<td>8 (29)</td>
</tr>
<tr>
<td>Student</td>
<td>4 (14)</td>
<td>1 (4)</td>
</tr>
<tr>
<td>Missing</td>
<td>2 (7)</td>
<td>0 (0)</td>
</tr>
<tr>
<td><strong>Marital status, n (%)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Married</td>
<td>7 (25)</td>
<td>10 (36)</td>
</tr>
<tr>
<td>Never married</td>
<td>14 (50)</td>
<td>13 (36)</td>
</tr>
<tr>
<td>Divorced</td>
<td>1 (4)</td>
<td>2 (46)</td>
</tr>
<tr>
<td>Widowed</td>
<td>2 (7)</td>
<td>1 (7)</td>
</tr>
<tr>
<td>Other</td>
<td>2 (7)</td>
<td>2 (4)</td>
</tr>
<tr>
<td>Missing</td>
<td>2 (7)</td>
<td>0 (0)</td>
</tr>
<tr>
<td><strong>Children, n (%)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>5 (18)</td>
<td>13 (46)</td>
</tr>
<tr>
<td>No</td>
<td>21 (75)</td>
<td>15 (54)</td>
</tr>
<tr>
<td>Missing</td>
<td>2 (7)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Characteristics</td>
<td>i-CBT (n=28)</td>
<td>Stepped care (n=28)</td>
</tr>
<tr>
<td>-----------------</td>
<td>-------------</td>
<td>--------------------</td>
</tr>
<tr>
<td>Income (CAD $; CAD $1=US 0.74), n (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;20,000</td>
<td>7 (25)</td>
<td>6 (21)</td>
</tr>
<tr>
<td>20,000-34,999</td>
<td>5 (18)</td>
<td>3 (11)</td>
</tr>
<tr>
<td>35,000-49,999</td>
<td>5 (18)</td>
<td>8 (29)</td>
</tr>
<tr>
<td>50,000-74,999</td>
<td>4 (14)</td>
<td>6 (21)</td>
</tr>
<tr>
<td>75,000-99,999</td>
<td>0 (0)</td>
<td>2 (7)</td>
</tr>
<tr>
<td>&gt;100,000</td>
<td>3 (11)</td>
<td>3 (11)</td>
</tr>
<tr>
<td>Missing</td>
<td>4 (14)</td>
<td>0 (0)</td>
</tr>
</tbody>
</table>

aPHQ-9: Patient Health Questionnaire–9.
Figure 2. CONSORT (Consolidated Standards of Reporting Trials) flow diagram of participant flow through the study. i-CBT: internet-based cognitive behavioral therapy; MINI: Mini International Neuropsychiatric Interview; PHQ-9: Patient Health Questionnaire–9; QIDS: Quick Inventory of Depressive Symptomatology; Q-LES-Q: Quality of Life Enjoyment and Satisfaction Questionnaire–Short Form.

Table 3. Sample size of the highest intervention provided for participants in the stepped care group, sorted based on completion and dropout from the internet-based cognitive behavioral therapy (i-CBT) program.

<table>
<thead>
<tr>
<th>Stepped care interventions</th>
<th>Step 0: i-CBT only, n (%)</th>
<th>Step 1: message, n (%)</th>
<th>Step 2: phone call, n (%)</th>
<th>Step 3: video call, n (%)</th>
<th>Step 4: psychiatrist, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dropout (n=13)</td>
<td>5 (38)</td>
<td>1 (8)</td>
<td>6 (46)</td>
<td>0 (0)</td>
<td>1 (8)</td>
</tr>
<tr>
<td>Completed (n=15)</td>
<td>3 (20)</td>
<td>1 (7)</td>
<td>9 (60)</td>
<td>1 (7)</td>
<td>1 (7)</td>
</tr>
<tr>
<td>Total (n=28)</td>
<td>8 (29)</td>
<td>2 (7)</td>
<td>15 (54)</td>
<td>1 (4)</td>
<td>2 (7)</td>
</tr>
</tbody>
</table>
**Table 4.** The number of participants provided with each stepped care intervention for each session, from session 5 to session 13 (n=28).

<table>
<thead>
<tr>
<th>Step</th>
<th>Session 5, n (%)</th>
<th>Session 6, n (%)</th>
<th>Session 7, n (%)</th>
<th>Session 8, n (%)</th>
<th>Session 9, n (%)</th>
<th>Session 10, n (%)</th>
<th>Session 11, n (%)</th>
<th>Session 12, n (%)</th>
<th>Session 13, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>N/A</td>
<td>3 (11)</td>
<td>5 (18)</td>
<td>6 (21)</td>
<td>6 (21)</td>
<td>8 (29)</td>
<td>9 (32)</td>
<td>11 (39)</td>
<td>13 (46)</td>
<td>13 (46)</td>
</tr>
<tr>
<td>0</td>
<td>16 (57)</td>
<td>17 (61)</td>
<td>16 (57)</td>
<td>15 (54)</td>
<td>12 (43)</td>
<td>15 (54)</td>
<td>10 (36)</td>
<td>12 (43)</td>
<td>8 (29)</td>
</tr>
<tr>
<td>1</td>
<td>4 (14)</td>
<td>1 (4)</td>
<td>2 (7)</td>
<td>1 (4)</td>
<td>4 (14)</td>
<td>0 (0)</td>
<td>3 (11)</td>
<td>1 (4)</td>
<td>2 (7)</td>
</tr>
<tr>
<td>2</td>
<td>5 (18)</td>
<td>5 (18)</td>
<td>4 (14)</td>
<td>6 (21)</td>
<td>4 (14)</td>
<td>2 (7)</td>
<td>2 (7)</td>
<td>1 (4)</td>
<td>3 (11)</td>
</tr>
<tr>
<td>3</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>2 (7)</td>
<td>1 (4)</td>
<td>1 (4)</td>
<td>1 (4)</td>
</tr>
<tr>
<td>4</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>1 (4)</td>
<td>0 (0)</td>
<td>0 (0)</td>
</tr>
</tbody>
</table>

*N/A*: not applicable.

**Measured Outcomes**

**PHQ-9 Score**

At baseline, for the participants who initiated the i-CBT program, the mean PHQ-9 score was 16.63 (SD 4.40; 27/28, 96%) for the control group and 17.75 (SD 5.33; 28/28, 100%) for the stepped care group, showing no statistically significant difference in pretreatment scores between the 2 groups ($t_{53}=-0.85; P=.40; 95\% CI -3.77 to -1.53$). A 2×5 repeated-measures ANOVA determined that the mean PHQ-9 scores differed significantly between time points ($F_{4,80}=9.95; P<.001$; **Figure 3**), but there was no significant difference at different time points between the 2 groups ($F_{4,80}=0.43; P=.78$; **Figure 4**). Post hoc analysis with Bonferroni adjustment revealed that PHQ-9 scores significantly decreased from pretreatment period (week 1) to the second time point (week 4; 4.554, 95% CI 1.264-7.843; $P=.003$), from pretreatment period (week 1) to the third time point (week 7; 4.973, 95% CI 1.904-8.042; $P<.001$), from pretreatment period (week 1) to the fourth time point (week 10; 4.384, 95% CI 0.862-7.906; $P=.008$), and from pretreatment period (week 1) to posttreatment period (week 13; 5.357, 95% CI 2.169-8.546; $P<.001$). All between–time point scores were statistically significant (**Figure 3**). A detailed breakdown of the results are summarized in **Multimedia Appendix 4**.
Figure 4. Estimated marginal means of Patient Health Questionnaire–9 (PHQ-9) scores at 5 intervals of treatment (sessions 0, 4, 7, 10, and 13), in both internet-based cognitive behavioral therapy (i-CBT; blue) and i-CBT with stepped care (green) treatment conditions. Error bars depict –2 or +2 SE.

QIDS Score

At baseline, for the participants who initiated the i-CBT program, the mean QIDS score was 14.62 (SD 4.64; 21/28, 75%) for the control group and 16.83 (SD 4.43; 23/28, 82%) for the stepped care group, showing no statistical significance in pretreatment scores between the 2 groups ($t_{42}$=−1.61; $P$=.11; 95% CI −4.97 to 0.55; independent samples $t$ test). A 2×3 repeated-measures ANOVA determined that the mean QIDS scores differed significantly between time points ($F_{2,28}$=5.73; $P$=.008; Figure 5), but there was no significant difference at different time points between the 2 groups ($F_{2,28}$=3.05; $P$=.06; Figure 6). Post hoc analysis with Bonferroni adjustment revealed that QIDS scores significantly decreased from pretreatment period (week 1) to posttreatment period (week 13; 4.12, 95% CI 0.76-7.47; $P$=.02). There was no significant difference between pretreatment period (week 1) and midtreatment period (week 7; 2.08, 95% CI −1.10 to 5.26; $P$=.29) or between midtreatment period (week 7) and posttreatment period (week 13; 2.03, 95% CI −1.34 to 5.41; $P$=.37; Figure 5). A detailed breakdown of the results are summarized in Multimedia Appendix 4.

Figure 5. Estimated marginal means of Quick Inventory of Depressive Symptomatology (QIDS) scores at 3 treatment time intervals corresponding to sessions 1, 7, and 13, including both groups—internet-based cognitive behavioral therapy (i-CBT) only and i-CBT with stepped care. Error bars depict –2 or +2 SE.
Figure 6. Estimated marginal means of Quick Inventory of Depressive Symptomatology (QIDS) scores at 3 intervals of treatment (sessions 0, 7, and 13), in both internet-based cognitive behavioral therapy (i-CBT; blue) and i-CBT with stepped care (green) treatment conditions. Error bars depict −2 or +2 SE.

**Q-LES-Q Score**

At baseline, for the participants who initiated the i-CBT program, the mean Q-LES-Q score was 36.88 (SD 7.80; 26/28, 93%) for the control group and 36.25 (SD 7.81; 28/28, 100%) for the stepped care group, showing no statistical significance in pretreatment scores between the 2 groups ($t_{52}=0.30; P=.77$; 95% CI −3.63 to 4.90; independent samples $t$ test). A 2×3 repeated-measures ANOVA determined that the mean Q-LES-Q scores differed significantly between time points ($F_{2,38}=4.18; P=.02$; Figure 7), but there was no significant difference at different time points between the 2 groups ($F_{2,38}=0.19; P=.83$; Figure 8). Post hoc analysis with Bonferroni adjustment revealed that Q-LES-Q scores significantly increased from pretreatment period (week 1) to midtreatment period (week 7; −5.50, 95% CI −10.91 to −0.94; $P=.045$). There were no significant differences between pretreatment period (week 1) and posttreatment period (week 13; −4.77, 95% CI −11.03 to 1.49; $P=.18$) and between midtreatment period (week 7) and posttreatment period (week 13; 0.73, 95% CI −5.18 to 3.72; $P=.99$; Figure 7). A detailed breakdown of the results are summarized in Multimedia Appendix 4.
Figure 7. Estimated marginal means of Quality of Life Enjoyment and Satisfaction Questionnaire–Short Form (Q-LES-Q) scores at 3 treatment time intervals corresponding to sessions 1, 7, and 13, including both groups—internet-based cognitive behavioral therapy (i-CBT) only and i-CBT with stepped care. Error bars depict –2 or +2 SE.

Figure 8. Estimated marginal means of Quality of Life Enjoyment and Satisfaction Questionnaire–Short Form (Q-LES-Q) scores at 3 intervals of treatment (sessions 0, 7, and 13), in both internet-based cognitive behavioral therapy (i-CBT; blue) and i-CBT with stepped care (green) treatment conditions. Error bars depict –2 or +2 SE.

Treatment Compliance
The proportion of participants completing the full round of therapy (ie, 13 sessions) was 32% (9/28) in the i-CBT–only group and 54% (15/28) in the stepped care group. There was no significant difference between the 2 groups regarding completion of the i-CBT program for depression ($\chi^2 = 2.6; P=.10$; chi-square test). An independent samples t test was conducted to determine whether there was a difference in the number of completed sessions between the i-CBT group and the stepped care group for the individuals who had begun the program. The results indicate significant difference between the average number of sessions completed by participants in the i-CBT–only group (mean 7, SD 4.64; 28/28, 100%) and stepped care group (mean 9.41, SD 4.44; 28/28, 100%; $t_{55} = -2; P=.03; 95\% \text{ CI} \approx -4.83 \text{ to } -0.002$; independent samples t test). Multimedia Appendix 5 shows the distribution of the sessions completed in each group.
ITT Analysis

A linear mixed model analysis was used to conduct the intent-to-treat analysis for each participant across the program. The group (i-CBT or stepped care) and evaluation time points were indicated as fixed factors. This analysis included participants who did not complete the full program and dropped out of the study before completion. As seen previously, there was significant change in PHQ-9 scores across the time points \((F_{4,48.79}=10.98; P<.001)\), but there was no significant difference between the 2 groups \((F_{1,107.07}=2.79; P=.10)\). In addition, for the QIDS scores, there was significant change in scores across the time points \((F_{2,41.08}=8.53; P<.001)\), but no significant difference between the 2 groups \((F_{1,54.44}=2.72; P=.10)\). Q-LES-Q scores imitated a similar pattern, with significant difference in scores across the time points \((F_{2,44.90}=9.36; P<.001)\), but no significant difference between the 2 groups \((F_{1,60.73}=3.28; P=.08)\).

Discussion

Principal Findings

This study evaluated the effectiveness of an i-CBT program with and without stepped care for depression among adults. The results of this study indicate that the proposed stepped care model was not significantly better in reducing depressive symptoms, as measured by PHQ-9 and QIDS, than the i-CBT program alone. Some previous studies have also found no significant differences between stepped care treatments and care as usual in improving depressive symptoms [52,68]. Ho et al [52], who reviewed stepped care for both depressive and anxiety disorders, found that stepped care treatment was significantly better in improving anxiety symptoms but did not find any significant difference in reducing depressive symptoms. This suggests some differences in the populations and the sensitivity of the stepped care structure in addressing different mental health issues. When viewing the results of both groups together, this study indicated that the i-CBT program itself was effective in significantly reducing depressive symptoms from pretreatment period to posttreatment period, as measured by the PHQ-9 and QIDS; however, there were no significant differences in the reduction of depressive symptoms between the 2 groups. The stepped care group did not show to be significantly better in reducing depressive symptoms than the i-CBT group. This finding is consistent with other studies of i-CBT programs, which show improvements in depressive symptoms for mild to moderate depression [69,70]. This also reinforces the efficacy of the i-CBT program for depression on OPTT used in this study, based on similar results from previous clinical trials using this program [47,49].

Furthermore, the results suggest that the i-CBT program indicated significant difference in the quality of life, as measured by Q-LES-Q; however, post hoc analysis revealed that overall, there were no significant differences before and after treatment in improving the quality of life. Significant difference was observed between pretreatment period and midtreatment period, which may suggest that the i-CBT program had the greatest impact on improving quality of life in the early stages of the program. No significant differences were observed between the i-CBT and stepped care groups. A systematic review and meta-analysis reviewed 3 studies of the effects of i-CBT on the quality of life and found inconclusive evidence of i-CBT compared with in-person outcomes [71]. We found no significant difference between the effects of i-CBT with stepped care and i-CBT without stepped care on the quality of life.

Although the completion rates between the 2 groups were not significantly different, participants in the stepped care group significantly completed, on average, 2 more sessions of therapy than those in the i-CBT group. The increase in the number of completed sessions may have been caused by the added support offered in the stepped care group. This is consistent with previous studies that report improvement in treatment adherence in mental health treatment with additional interventions in care [21,28,63]. Specifically, phone and video calls assist with keeping patients in care compared with in-person treatment [36]. Compared with in-person treatment, receiving care for depression over the phone [36,72] or via video [22] is associated with high completion rates and few dropouts. Further studies are required to decipher whether the stepped care group’s high completion rate of sessions is associated with better treatment outcomes than the i-CBT group.

Furthermore, Multimedia Appendix 5 shows that although participant dropout occurs uniformly across the first 10 sessions in the stepped care group, most participants (19/28, 68%) in the i-CBT group dropped out in the first half of the program (ie, first 7 sessions): 25% (7/28) of participants in the stepped care group dropped out of the program before the midpoint of treatment (ie, session 7) and 68% (19/28) of participants in the i-CBT group. Previous studies have noted that most patients drop out of treatment programs after 2 to 4 sessions [16,31,73]. During this time, participants were provided with i-CBT care only in both groups (additional interventions for the stepped care group were introduced at session 5 after monitoring their PHQ-9 scores and interaction with the i-CBT program); however, introducing a possibility of intervention soon in the stepped care group may have helped participants who were indicated as an early dropout (ie, before session 5) complete more sessions. Concurrently, it is noteworthy that the i-CBT program’s 13 sessions were not completed by several participants (32/56, 57%) in both groups. This is consistent with previous studies that reveal high dropout rates in depression i-CBT programs. Compliance is generally a challenge for i-CBT programs, with dropout rates averaging approximately 32% and ranging between 0% and 75% [26,37], whereas traditional CBT dropout rates are approximately 25% on average, ranging between 0% and 68% [74]. However, another study found that additional human support results in a large effect size regarding the efficacy of i-CBT (g=0.673) than providing no additional support (g=0.239) [34]. Thus, it was expected that the stepped care group would exhibit significant improvements following treatment, but this was not observed in this study. This may be owing to the nature of the study design, as the i-CBT group received some support during the treatment through the homework feedback and did not encompass a truly no-support intervention. Upon reviewing participant progression in the stepped care model (Table 4), it was seen that on average, most
participants (13/28, 46%) spent their time in step 0 of the stepped care model, which mimics the control group (i-CBT only). This may be a factor in the observed nonsignificant differences between the 2 groups because not many participants were stepped up in their care across treatment. Moreover, the small sample size may have influenced the outcomes. A previous study has shown that patients who terminate CBT prematurely show high symptom severity compared with patients who complete therapy, but the 2 groups did not vary in the rate of symptom change [75]. This suggests that the increase in symptom severity may arise from completing few sessions of treatment in the dropout group compared with the completer group and explains the differing results observed between the 2 groups as dropout was high. Given the effectiveness of i-CBT treatment and the high dropout rates, it is important to assist patients in completing more sessions by providing sufficient resources and care. Our proposed stepped care model assisted participants in completing a great number of sessions. Stepped care has been associated with high treatment satisfaction, which may have assisted in participants completing more sessions in this group [68,76]. Future studies should examine the variables that may be responsible for treatment attrition in i-CBT programs and devise methods to raise the rates of treatment engagement and completion.

Limitations

It is important to consider further limitations of this study. This includes the relatively small sample size; observed sex imbalance, with 61% (34/56) women in the study (approximately 15% more women in the i-CBT group compared with the stepped care group); and lack of long-term follow-up, which is currently ongoing. It is important to note that the sample size for analysis of the i-CBT group was smaller than that of the stepped care group. In addition, the participants in this study may not be representative of the general population, as they were predominantly women, English speaking (owing to the limitations of the i-CBT program), employed full time, and recruited from a specific clinical setting of self-referrals and clinics limited to Kingston, Ontario (Table 2). Missing data were also a challenge that affected the total sample size. A large portion of the data was unavailable owing to collection errors and dropouts (Figure 2). We attempted to address this issue by using ITT analysis and observed similar results with no significant differences between groups but observed significant differences across various time points for PHQ-9, QIDS, and Q-LES-Q scores.

This study was not specifically designed to investigate the effects of the different treatment interventions in the stepped care group and how they influenced the results. In hindsight, using 4 different factors to decide about the stepped intervention results in increased variability in the stepped care decision and limits our ability to decipher the effectiveness of the approach. Most stepped care models use clinical questionnaire scores, such as the PHQ-9 score, to decide when to step up or step down care [68,76-78]; however, our study did not include strict cutoff guidelines and instead adopted a subjective monitoring approach of PHQ-9 scores as one of the evaluating factors. Changes in PHQ-9 scores (scale ranging from 0-27) and homework submission (submitted or not submitted) can be quantized predictors for stepped care; however, monitoring engagement with the OPTT platform and goal progression are variables that are multifactorial and are subject to interpretation. This design prevented us from making conclusions about a best-fit intervention model. Furthermore, beginning the stepped care interventions in session 5 allowed for a watchful waiting period that allowed reflection about patient status to determine the correct intensity of care if they did not improve with a low-resource intervention (ie, i-CBT). However, this may be a limitation of the proposed stepped care model, as the waiting period may be harmful because it may delay optimal treatment [68]. With a small sample size, it is difficult to analyze the effects of these deciding factors. Future studies need to be conducted to explore the effects of the individual stepped interventions provided in this study and any potential relationships among the modalities of care.

Therapeutic alliance in psychotherapy is another significant factor in predicting treatment outcomes [79-81]. The i-CBT–only group had limited interaction with their assigned care provider, and the ability to build rapport with participants was limited. In the stepped care group, participants were potentially able to gain a deep connection with their care provider, beginning in session 5, based on the intervention provided to them. Therapeutic alliance follows two important phases: (1) initial alliance development, usually occurring within the first 5 sessions, and (2) challenging the patient more actively, which can cause strain to the therapeutic relationship [79]. With stepped care interventions being provided in session 5, the first phase seems to be neglected and may affect the potential relationship, thus the structure of the program hinders participants from achieving the desired treatment outcome and reducing their depressive symptoms. Some studies demonstrate that a positive outcome is more predictive by the quality of the alliance rather than the type of the intervention [79,82-84]. In this case, the quality of the alliance was affected by the limited interaction with the care provider and may reflect one of the limitations of online therapy. Some studies found that building therapeutic alliances face to face is significantly more effective than online psychological treatments [85-89]. In contrast, other studies have found the opposite and report no correlation between the therapeutic alliance and treatment outcomes in digital interventions [90-93]. Further studies are required to better understand the characteristics of therapeutic alliance in digital contexts [25,94]. It would also be interesting to assess participant and care provider opinions about their therapeutic alliance during the study to better understand this limitation.

Conclusions

This study provides further evidence that i-CBT is effective for treating depressive symptoms. However, the study showed no significant evidence that the proposed stepped care model was more effective than i-CBT alone. The stepped care model allowed participants to complete 2 more sessions on average than the i-CBT–only group, indicating that stepped care is an effective method for guiding patients to treatment completion. Future studies should examine the long-term effects of such interventions and the efficacy of specific stepped care interventions in large and more diverse groups. To improve the design and implementation of such a model, studies might also...
investigate the processes through which stepped care quality of life interventions reduce depression symptoms and enhance the

Acknowledgments
The authors would like to thank all the care providers involved in this program for ensuring that participants were supported throughout the study. The authors are also grateful to the participants of this study who provided their time and trust in the process to help the authors complete and further investigate this online cognitive behavioral therapy approach.

Data Availability
The data sets generated during and analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest
NA and MO have cofounded Online Psychotherapy Tool (OPTT), the platform used to provide care to participants. They have ownership stakes in OPTT Inc.

Multimedia Appendix 1
CONSORT-eHEALTH checklist (V 1.6.1).
[PDF File (Adobe PDF File), 7464 KB - mental_v11i1e51704_app1.pdf ]

Multimedia Appendix 2
Sample of the e–cognitive behavioral therapy program for depression used in the study, showcasing session 5.
[PDF File (Adobe PDF File), 3875 KB - mental_v11i1e51704_app2.pdf ]

Multimedia Appendix 3
Feedback templates for e–cognitive behavioral therapy used throughout the study by care providers including feedback for session homework and live interactions between the care provider and participants.
[DOCX File , 20 KB - mental_v11i1e51704_app3.docx ]

Multimedia Appendix 4
Statistical analysis results.
[DOCX File , 31 KB - mental_v11i1e51704_app4.docx ]

Multimedia Appendix 5
The frequency of participants’ last completed sessions organized by 2 groups—internet-based cognitive behavioral therapy (i-CBT) only and i-CBT with stepped care.
[PNG File , 62 KB - mental_v11i1e51704_app5.png ]

References


Abbreviations

CBT: cognitive behavioral therapy
CONSORT-EHEALTH: Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth
i-CBT: internet-based cognitive behavioral therapy
ITT: intention-to-treat
MDD: major depressive disorder
MINI: Mini International Neuropsychiatric Interview
OPTT: Online Psychotherapy Tool
PHQ-9: Patient Health Questionnaire–9
Q-LES-Q: Quality of Life Enjoyment and Satisfaction Questionnaire–Short Form
QIDS: Quick Inventory of Depressive Symptomatology
QUOPL: Queen’s Online Psychotherapy Lab
HealthySMS Text Messaging System Adjunct to Adolescent Group Cognitive Behavioral Therapy in the Context of COVID-19 (Let’s Text!): Pilot Feasibility and Acceptability Study

Lauren M Haack, PhD; Courtney C Armstrong, BSc; Kate Travis, MD; Adrian Aguilera, PhD; Sabrina M Darrow, PhD

1Department of Psychiatry and Behavioral Sciences, University of California San Francisco, San Francisco, CA, United States
2School of Social Welfare, University of California Berkeley, Berkeley, CA, United States

Corresponding Author:
Lauren M Haack, PhD
Department of Psychiatry and Behavioral Sciences
University of California San Francisco
675 18th Street
San Francisco, CA, 94107
United States
Phone: 1 415 502 8060
Email: lauren.haack@ucsf.edu

Abstract

Background: The widespread occurrence and devastating impact of adolescent depression warrant health service research focused on feasible and acceptable digital health tools to supplement evidence-based intervention (EBI) efforts, particularly in the context of shelter-in-place guidelines disrupting youth socialization and service use in the wake of the COVID-19 pandemic. Given the promise of SMS text message interventions to enhance EBI engagement, our team developed the HealthySMS system as an adjunct to one of the most empirically supported interventions for adolescent depression: cognitive behavioral therapy (CBT) group services. The system sends daily SMS text messages requesting responses assessing mood, thoughts, and activities; weekly attendance reminder messages; daily tips about adherence (eg, a prompt for activity completion); and personalized responses based on participants’ texts.

Objective: This study aims to evaluate the feasibility and acceptability of HealthySMS in a real-world setting and explore potential mechanisms of change in EBI engagement, before evaluating the system’s impact on adolescents’ group CBT engagement and, ultimately, depression outcomes.

Methods: Over the course of 2020, we invited all 20 adolescents receiving CBT group services for depression at an outpatient psychiatry clinic to enroll in our HealthySMS study; ultimately, 17 (85%) adolescents agreed to participate. We tracked participant initiation and engagement with the HealthySMS system as well as the content of SMS text message responses to HealthySMS. We also invited each participant to engage in a semistructured interview to gather additional qualitative inputs on the system.

Results: All (n=17, 100%) research participants invited agreed to receive HealthySMS messages, and 94% (16/17) of the participants maintained use during the first month without opting out. We uncovered meaningful qualitative themes regarding the feasibility and acceptability of HealthySMS, as well as its potential impact on EBI engagement.

Conclusions: Taken together, the results of this pilot study suggest that HealthySMS adjunct to adolescent CBT group depression services is feasible and acceptable, as evidenced by high rates of HealthySMS initiation and low rates of dropout, as well as meaningful themes uncovered from participants’ qualitative feedback. In addition, the findings provide evidence regarding iterative improvements to the HealthySMS system and research protocol, as well as potential mechanisms of change for enhanced EBI engagement and, ultimately, adolescent depression outcomes, which can be used in future effectiveness research.

(JMIR Ment Health 2024;11:e49317) doi:10.2196/49317
Introduction

Background

Depression is a major public health concern for adolescents [1-3]; it is a risk factor for mental health problems that occur in adults, medical illnesses, disability, substance abuse, and suicide [2-6]. Evidence-based efforts to improve adolescent depression and related risks became more imperative during the COVID-19 pandemic owing to concerns about adolescents’ mental health deterioration in the context of shelter-in-place (SIP) guidelines causing substantial educational and social disruptions [7-10]. Further, most in-person services were halted during the implementation of SIP guidelines, rapidly necessitating a transition to telehealth and digital health (dHealth) supports [7,11-14]. In response, the need for pilot feasibility and acceptability testing of dHealth tools targeting adolescent depression in real-world settings became glaringly apparent. Such efforts would provide an essential first step toward testing the effectiveness of dHealth tools for adolescent depression. The impact of this work would extend beyond the initial COVID-19 SIP context, given that research suggests that many systems will and should continue using telehealth and dHealth supports in the foreseeable future [13,15,16], which may be particularly beneficial in reducing disparities for historically marginalized populations [15].

Evidence-Based Intervention for Adolescents With Depression

There are several effective evidence-based interventions (EBIs) for adolescent depression [17-19]. The most well-supported psychosocial approach is cognitive behavioral therapy (CBT) [18,19]), which targets unhelpful thinking patterns and the avoidance of goal-directed and social activities, which are characteristic of depression [20-22]. CBT can be effectively delivered via a group format [23], demonstrating efficacy similar to that of individual CBT in decreasing depression in a more cost-effective modality [24]. CBT can also be delivered electronically (electronic CBT [eCBT]), enhancing accessibility by eliminating the requirement for patients to travel for in-person services [25-29]. The potential benefits of electronically delivered EBIs became increasingly evident in the context of the implementation SIP guidelines during the COVID-19 pandemic, which prohibited many individuals from receiving in-person mental health treatment. Importantly, our society may experience future SIP guidelines during COVID-19 surges, safety lockdowns, or natural disasters in response to climate change. Fortunately, recent research supports the effectiveness and efficacy of eCBT in decreasing depression [25-29].

Despite the promising efficacy research supporting CBT as an EBI for adolescent depression, the effect sizes are heterogeneous and low, demonstrating room for improvement [23,30]. Optimal treatment of adolescent depression requires patient engagement, including the initiation of treatment after referral, the attendance of sessions, the completion of homework between sessions, and the continued engagement in treatment (i.e., getting the whole “dose” of treatment [31]). Thus, poor initiation, poor adherence, and treatment dropout are barriers to effective services in real-world settings and are known to mediate treatment outcome [31]. A meta-analysis of EBIs concluded that the average treatment dropout rate is approximately 29% and that patients with depression are at a higher risk for dropout [32]. Treatment dropout appears particularly pronounced among youths and in web-based mental health interventions; therefore, efforts to enhance engagement with eCBT among youth populations are especially called for [33].

SMS Text Message Interventions as Automated dHealth Supports Aiming to Enhance EBI Engagement

There is a strong scientific premise and public health call to develop and evaluate dHealth supports adjunct to EBIs that are easy to implement, efficient, and sustainable. Using SMS text messages is a promising approach aiming to enhance adolescent EBI engagement by making use of a tool that is readily available and widely used by adolescents. The use of mobile phones is ubiquitous among adolescents, and SMS text messages are used at high rates. As of 2015, more than 85% of adolescents across races and ethnicities had access to a cell phone, and 90% of them used cell phones to send SMS text messages [34]. In fact, two-thirds of adolescents reported that they are more likely to use their cell phones to text rather than talk to friends, and a typical adolescent in the United States sends ≥30 SMS text messages each day [34]. Importantly, in mobile health (mHealth) research, youths report high satisfaction and readily engage with technology [35-37]. In addition, digital technology allows for the automation of SMS text messages; thus, an SMS text message intervention leveraging automation should not add to providers’ burden, making it a more sustainable and efficient services intervention.

SMS text messaging interventions have several advantages over other forms of dHealth interventions, such as app-based and website-based interventions, and are particularly well suited to improve engagement. SMS text messages allow for a more equitable delivery of care, given that they can reach anyone with a phone and do not rely on smartphone ownership or internet access, which are affected by socioeconomic, racial, and ethnic disparities [38]. In addition, although dHealth interventions have demonstrated promise, web-based and mobile app–based interventions are subject to more difficulties with engagement, such as problems with adherence and dropout [26,39].

Some dHealth SMS text message interventions were designed as stand-alone supports (eg, those in the works of Aguilera et al [40] and Bendsten et al [41]; MacDougall et al [42] conducted a scoping review of SMS text message–delivered adolescent mental health interventions); however, there may be unique benefits to SMS text message interventions designed to...
supplement EBIs and enhance treatment engagement. There are many effective examples of adjunctive SMS text message interventions for disease management and health behavior change (eg, smoking cessation and diabetes management [43]), including SMS text message interventions for adolescents [44]. In the emerging literature on SMS text message interventions, adolescents generally react favorably and show good compliance [35,42,45,46]. Pilot research on integrating SMS text messages into individual CBT has also demonstrated encouraging results [47].

Given the promise of SMS text message interventions to enhance EBI engagement, our team developed the HealthySMS system as an adjunct to group CBT depression services. HealthySMS sends customized SMS text messages to participants, inquiring about mood as well as reminding them to attend CBT and practice strategies learned during group treatment. For safety, adolescent SMS text message responses are monitored for keywords or phrases that could indicate whether someone is expressing suicide risk; providers are sent immediate alerts if the system is triggered by an adolescent SMS text message response with these keywords or phrases. Safety keyword triggers are important features of any automated dHealth system implemented in real-world settings to keep participants safe and providers informed; in addition, this feature may be particularly helpful during the implementation of SIP guidelines, given the probable decrease in adolescent safety monitoring from other settings, such as schools. The addition of HealthySMS to CBT for adults with depression in a public sector treatment setting was associated with an increased number of sessions attended and a longer duration of treatment [48]. The mood ratings sent by adults through HealthySMS also predicted attendance [49]. Thus, adding HealthySMS to EBIs for adolescent depression may be a promising change to existing services, targeting increased engagement and, ultimately, improving outcomes.

This Study
As a first step in the process of implementing and evaluating the HealthySMS system adjunct to the most evidence-based treatment for adolescent depression (ie, CBT), we conducted a feasibility pilot study in an outpatient clinic embedded within an academic medical center’s department of psychiatry and behavioral sciences. Our primary objectives were to investigate the feasibility and acceptability of HealthySMS in a real-world setting and explore potential mechanisms of change in EBI engagement, before evaluating the system’s impact on adolescents’ group CBT engagement and, ultimately, depression outcomes. We also aimed to monitor the HealthySMS safety keyword alert triggers and provider responses to inform system and research protocol adjustments before future HealthySMS research. This multiphase design featuring a preliminary feasibility pilot to inform decisions about future effectiveness testing is aligned with the Medical Research Council framework [50,51], which has been prolifically used in mental health and dHealth intervention research. We predicted the following hypotheses:

1. Hypothesis 1: most adolescents invited to use HealthySMS (ie, ≥75%) would initiate and maintain use without opting out during their group CBT experience.
2. Hypothesis 2: adolescents enrolled in HealthySMS would display high rates of engagement with the SMS text message system (ie, ≥50% response rate).
3. Hypothesis 3: our team would uncover meaningful qualitative themes from participants’ SMS text message responses and semistructured interviews about the feasibility and acceptability of HealthySMS to inform iterative system and research protocol improvements supporting future HealthySMS effectiveness research.
4. Hypothesis 4: our team would uncover meaningful qualitative themes from participants’ SMS text message responses and semistructured interviews about the potential impact of HealthySMS on EBI engagement to inform decisions on which mechanisms of change to evaluate in future HealthySMS effectiveness research.

Methods
Participants
Participants were adolescents aged 13 to 18 years. They were recruited from the University of California San Francisco (UCSF) Department of Psychiatry and Behavioral Sciences Child and Adolescent Services clinic between December 2019 and September 2020. Adolescents were eligible to participate if they were enrolled in the clinic’s group CBT for depression running from January 2020 to September 2020.

Of the 20 eligible adolescents who started the group CBT for depression, 17 (85%) agreed to participate in the study. Participant characteristics are listed in Table 1. Briefly, the average age of the participants was 15.4 (SD 1.5) years, and most participants (15/17, 88%) were diagnosed with major depressive disorder. Most participants (12/17, 71%) had engaged in prior mental health care. Moreover, 18% (3/17) of the participants had a history of at least 1 suicide attempt, with most participants (11/17, 65%) reporting suicidal ideation in the past year. A little more than half (9/17, 53%) of the participants engaged in concurrent individual therapy, family therapy, or a combination.
Table 1. Participants’ demographic and clinical information (N=17).

<table>
<thead>
<tr>
<th>Demographic information</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (years), mean (SD)</td>
<td>15.4 (1.5)</td>
</tr>
<tr>
<td><strong>Gender, n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>Woman</td>
<td>8 (47)</td>
</tr>
<tr>
<td>Man</td>
<td>8 (47)</td>
</tr>
<tr>
<td>Transwoman</td>
<td>1 (6)</td>
</tr>
<tr>
<td><strong>Ethnicity, n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>Latinx</td>
<td>7 (41)</td>
</tr>
<tr>
<td>Non-Latinx</td>
<td>10 (59)</td>
</tr>
<tr>
<td><strong>Race, n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>Asian</td>
<td>1 (6)</td>
</tr>
<tr>
<td>Black</td>
<td>1 (6)</td>
</tr>
<tr>
<td>White</td>
<td>10 (59)</td>
</tr>
<tr>
<td>Biracial or multiracial</td>
<td>4 (24)</td>
</tr>
<tr>
<td>Something else (Hispanic or Mexican)</td>
<td>1 (6)</td>
</tr>
<tr>
<td><strong>Sexual orientation, n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>Heterosexual</td>
<td>8 (47)</td>
</tr>
<tr>
<td>Lesbian</td>
<td>1 (6)</td>
</tr>
<tr>
<td>Bisexual</td>
<td>3 (18)</td>
</tr>
<tr>
<td>Gay</td>
<td>1 (6)</td>
</tr>
<tr>
<td>Pansexual</td>
<td>1 (6)</td>
</tr>
<tr>
<td>Not sure or do not care</td>
<td>3 (18)</td>
</tr>
<tr>
<td><strong>Clinical history, n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>Prior therapy experience&lt;sup&gt;a&lt;/sup&gt;</td>
<td></td>
</tr>
<tr>
<td>Participated in prior outpatient therapy</td>
<td>9 (53)</td>
</tr>
<tr>
<td>Participated in IOP&lt;sup&gt;b&lt;/sup&gt; or PHP&lt;sup&gt;c&lt;/sup&gt;</td>
<td>2 (12)</td>
</tr>
<tr>
<td>Psychiatric hospitalization</td>
<td>1 (6)</td>
</tr>
<tr>
<td>Residential treatment facility</td>
<td>2 (12)</td>
</tr>
<tr>
<td>None</td>
<td>5 (29)</td>
</tr>
<tr>
<td>Prior suicide attempt</td>
<td>3 (18)</td>
</tr>
<tr>
<td><strong>SI&lt;sup&gt;d&lt;/sup&gt; at intake</strong></td>
<td></td>
</tr>
<tr>
<td>Denied</td>
<td>6 (35)</td>
</tr>
<tr>
<td>SI in the past week</td>
<td>3 (18)</td>
</tr>
<tr>
<td>SI in the past month</td>
<td>2 (12)</td>
</tr>
<tr>
<td>SI in the past year</td>
<td>6 (35)</td>
</tr>
<tr>
<td><strong>NSSI&lt;sup&gt;e&lt;/sup&gt; at intake</strong></td>
<td></td>
</tr>
<tr>
<td>Denied</td>
<td>10 (59)</td>
</tr>
<tr>
<td>NSSI in the past week</td>
<td>1 (6)</td>
</tr>
<tr>
<td>NSSI in the past month</td>
<td>2 (12)</td>
</tr>
<tr>
<td>NSSI in the past year</td>
<td>4 (24)</td>
</tr>
<tr>
<td><strong>School problems&lt;sup&gt;d&lt;/sup&gt;</strong></td>
<td></td>
</tr>
</tbody>
</table>

<sup>a</sup> Prior therapy experience includes outpatient therapy, inpatient therapy, residential treatment, and suicide attempts.

<sup>b</sup> IOP = Intensive Outpatient Program

<sup>c</sup> PHP = Partial Hospitalization Program

<sup>d</sup> SI = Suicide ideation

<sup>e</sup> NSSI = Non-suicidal self-injury
Values

None reported 9 (53)
IEP\(^{f}\) or accommodations 6 (35)
School refusal 3 (18)

**Major depressive disorder**
- Met criteria in the past 1 (6)
- Reports some symptoms 1 (6)
- Currently meets criteria 15 (88)

**Symptoms from comorbid disorders**
- Generalized anxiety disorder 5 (29)
- Panic disorder 2 (12)
- Social anxiety disorder 1 (6)
- Eating disorder 1 (6)
- ADHD\(^{g}\) 1 (6)
- None 7 (41)

**Concurrent therapy\(^{a}\)**
- Individual CBT\(^{b}\) 6 (35)
- Family therapy 4 (24)
- None 8 (47)
- Concurrent medication management 15 (88)

\(^{a}\)Response options are not mutually exclusive.
\(^{b}\)IOP: intensive outpatient treatment.
\(^{c}\)PHP: partial hospitalization program.
\(^{d}\)SI: suicidal ideation.
\(^{e}\)NSSI: nonsuicidal self-injury.
\(^{f}\)IEP: individualized education plan.
\(^{g}\)ADHD: attention-deficit/hyperactivity disorder.
\(^{h}\)CBT: cognitive behavioral therapy.

Of note, our research protocol did not make any changes to how the real-world clinic provided services. Adolescents were not incentivized to start treatment, attend sessions, or complete their homework; they only were incentivized for completing additional research tasks (see below). Adolescents in the clinic could be engaged in individual therapy, family therapy, or both while attending group treatment.

**Ethical Considerations**

All study procedures were approved by the UCSF Institutional Review Board (reference number 255820). When participants were minors (ie, age <18 years), their parent or legal guardian completed informed consent and they completed assent procedures; 18-year-old participants completed informed consent procedures. As part of these procedures, participants (and their parents or legal guardians, if applicable) were informed that their information would be kept private and housed on a secure UCSF server only accessible to the study team; they were informed that participants would be compensated with a US $30 gift card for attending the interview).

**CBT Group Intervention**

The Cognitive Behavioral Therapy for Depression Group for Adolescents (CBT-D) consists of three 4-week modules on thoughts, activities, and people (ie, 12 group sessions in total). It is based on the Building Recovery by Improving Goals, Habits, and Thoughts (BRIGHT) group CBT manual for depression for adults developed by Miranda et al [52], which was subsequently adapted for adolescents with a diagnosis of major depression or persistent depressive disorder. The thought module involves cognitive interventions with a focus on awareness of helpful and harmful thoughts, the activity module focuses on pleasant activities and behavioral activation, and the people module encourages group members to improve relationships and evaluate the impact of positive and negative social relationships on mood. The sessions are structured with an initial homework check-in followed by a didactic discussion on the covered topic, an interactive discussion led by 2 group providers, and the setting of homework goals related to skill use. In this study, the first 7 group sessions were held in person; the remainder were held via the Health Insurance Portability and Accountability Act–compliant Zoom (Zoom Video
Communications, Inc) platform after SIP orders were implemented in March 2020.

The HealthySMS System

We used the Health Insurance Portability and Accountability Act–compliant, web-based texting platform called HealthySMS, developed by our team member (AA), to send and receive SMS text messages, administer weekly surveys, and track attendance. HealthySMS sent four types of automated SMS text messages to participants: (1) daily mood prompts asking participants to rate their mood and reflect on their mood, thoughts, and behavior ("[First name], what is your mood right now on a scale of 1 to 9 (9 being best)? Please respond with a number and a message about what you are doing or thinking"); (2) personalized and reinforcing responses to 20% of participants’ mood ratings (ie, encouraging the participants to engage in behavioral activation in response to a low mood rating and reinforcing the participants in response to a high mood rating); (3) daily reminders about the concepts and skills learned during the corresponding CBT module for that month; and (4) weekly reminders to attend and come prepared to CBT sent the day before the group session. The participants could opt out of receiving SMS text messages at any point by texting “stop.”

Our clinical research team oriented adolescent participants to the HealthySMS system. We also oriented providers to the SMS text messages and HealthySMS web-based provider dashboard, which visualizes client responses to the SMS text messages, including graphs of mood ratings. HealthySMS monitors participants’ SMS text message responses and alerts providers to words and phrases that may correspond to suicidal behaviors (eg, “die,” “kill,” and “cut”), and providers were trained on how to respond to these alerts in alignment with existing clinic policies and procedures.

Measures

Participants’ Demographics and Clinical History

The participants were asked to complete a demographic survey before beginning group CBT. CBT-D providers recorded psychiatric diagnoses and treatment history information for each participant, including other therapy or medication management services, based on their evaluation and review of the medical record.

HealthySMS Engagement

We tracked the number of participants who were invited to receive HealthySMS messages, as well as those who agreed to initiate receipt of HealthySMS messages. We also tracked the number of participants who texted “stop” to opt out of SMS text messages before ending participation in the group, as well as how long each participant received the SMS text messages before opting out. We measured engagement with the HealthySMS messages by tracking the number of mood ratings that participants texted in response to mood rating request messages.

HealthySMS Safety Keyword Triggers

We tracked the number and content of participant SMS text message responses that were flagged for the risk of suicide (Multimedia Appendix 1 provides the trigger words). We also kept observation notes about the provider responses.

Qualitative Feedback

We collected qualitative feedback in several ways. We tracked responses to our monthly SMS text message prompts to participants asking, “What is the most positive part of receiving these text messages?” and “What do you not like about receiving the text messages?” We also invited all adolescent participants to share feedback in a semistructured interview after their CBT-D completion with an incentive of US $30 Amazon gift card. The interviews were moderated by a member of our clinical research team who followed a semistructured guide containing the study objectives to explain, questions to pose, and prompts to use when needed. Specifically, we explained that researchers hoped to obtain information about mHealth interventions such as HealthySMS and group CBT services for depression. Next, we asked for general feedback about the HealthySMS system and then specifically asked about different aspects of HealthySMS, such as the mood prompts, responses to mood ratings, and group reminders. We also showed participants a list of the HealthySMS texts and asked for feedback on their impact, content, and phrasing. Textbox 1 lists the interview questions and prompts. Each interview lasted between 19 and 45 (mean 32, SD 8.8) minutes, with the length depending on the amount of details provided by the respondents.
Textbox 1. Qualitative semistructured interview questions and prompts.

**How did you like the LET’S TEXT! message program overall?**
- Was there anything that made it difficult for you to receive the messages? For example, how did you like: the timing of the messages, the amount of messages, the phrasing of messages?
- Was there anything that made it difficult for you to respond to the messages? For example, how did you like: the timing of the messages, the amount of messages, the phrasing of messages?
- Is there anything about it you would suggest we change?

**How did you like the LET’S TEXT! Mood Prompts and Responses?** (after general feedback was given, participants were shown the list of printed texts for specific feedback on this question)
- What did you think worked well?
- What was difficult or did not go well?
- Is there anything you would change?
- Did these messages change your mood, thoughts, or behavior?

**How did you like the LET’S TEXT! Skill Practice Reminders?** (after general feedback was given, participants were shown the list of printed texts for specific feedback on this question)
- What did you think worked well?
- What was difficult or did not go well?
- Is there anything you would change?
- Did these messages change your mood, thoughts, or behavior?
- For example, did you need more/less help with any of the skills; was the purpose of the reminders clear; -were the messages too few/many; were the reminders relevant to your goals; were they phrased appropriately?

**How did you like the LET’S TEXT! Group Reminders?**
- Did these messages change your mood, thoughts, or behavior?

**Some teens find...**
- ...it easier to participate if they feel: comfortable, respected, and understood by the group members and group leader attached to the messages. How was your relationship with the group and group leader and did that affect your experience with the text messages?
- ...that the messages help them become more active. Do you think your activity completion was impacted by the text messages?
- ...the messages help them go to group and participate more often and/or effectively. How was your group engagement and did that change with the text messages?
- ...the messages are reinforcing and help them practice the group skills more often and/or effectively. How was your practice of the group skills and did that change with the text messages?

**Is there ANYTHING ELSE you would like to share that we haven’t asked you?**

**Data Analytic Plan**
We analyzed quantitative data on participant characteristics, HealthySMS engagement, and safety keyword triggers using descriptive statistics in SPSS (IBM Corp). We calculated the SMS text message response rates by dividing the number of responses by the number of SMS text messages received for mood ratings.

We analyzed the qualitative data in a multistep process using thematic analysis principles [53]. First, we developed a hierarchical coding system based on recurrent concepts that we uncovered when conducting the qualitative interviews and in consideration of the related theoretical literature. Next, members of our research team reviewed the qualitative message content and interview transcriptions to collaboratively refine recurrent themes while iteratively updating the coding system. Our team selected exemplary quotes for each theme.

**Results**

**Hypothesis 1**
Our first hypothesis was that most of the adolescents (ie, >75%) invited to HealthySMS would initiate and maintain use. All participants who agreed to the research study (n=17, 100%) opted to enroll in the HealthySMS system and initiate the receipt of messages upon starting CBT-D; 94% (16/17) of the participants maintained use during the CBT-D group experience. Only 1 (6%) participant opted out of the SMS text messages by texting “stop” 30 days after initiation.
Hypothesis 2
Our second hypothesis was that adolescents enrolled in HealthySMS would display high rates of engagement with the SMS text message system (ie, ≥75% response rate). As shown in Table 2, the HealthySMS response rate varied among participants. The average response rate to daily mood ratings was 61%, with a range of 0.00 to 1.77 responses per message. Of the 17 participants, only 1 (6%) participant did not respond to any of the SMS text messages prompting mood ratings, and 2 (12%) participants responded multiple times to several prompts (indicated by a response proportion >1). Most participants (10/17, 59%) responded to >50% of the daily mood rating prompts.

Table 2. HealthySMS engagement.

<table>
<thead>
<tr>
<th>Participant</th>
<th>Proportion of responses to mood ratings</th>
<th>Weeks until opting out</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.04</td>
<td>N/Aa</td>
</tr>
<tr>
<td>2</td>
<td>0.16</td>
<td>N/A</td>
</tr>
<tr>
<td>3</td>
<td>0.30</td>
<td>N/A</td>
</tr>
<tr>
<td>4</td>
<td>0.64</td>
<td>N/A</td>
</tr>
<tr>
<td>5</td>
<td>0.01</td>
<td>N/A</td>
</tr>
<tr>
<td>6</td>
<td>1.54</td>
<td>N/A</td>
</tr>
<tr>
<td>7</td>
<td>0.00</td>
<td>N/A</td>
</tr>
<tr>
<td>8</td>
<td>0.87</td>
<td>N/A</td>
</tr>
<tr>
<td>9</td>
<td>0.05</td>
<td>N/A</td>
</tr>
<tr>
<td>10</td>
<td>0.82</td>
<td>N/A</td>
</tr>
<tr>
<td>11</td>
<td>1.77</td>
<td>N/A</td>
</tr>
<tr>
<td>12</td>
<td>0.12</td>
<td>N/A</td>
</tr>
<tr>
<td>13</td>
<td>0.87</td>
<td>N/A</td>
</tr>
<tr>
<td>14</td>
<td>0.77</td>
<td>N/A</td>
</tr>
<tr>
<td>15</td>
<td>0.79</td>
<td>N/A</td>
</tr>
<tr>
<td>16</td>
<td>0.75</td>
<td>N/A</td>
</tr>
<tr>
<td>17</td>
<td>0.84</td>
<td>4</td>
</tr>
</tbody>
</table>

aN/A: not applicable.

Hypothesis 3
Our third hypothesis was that we would uncover meaningful qualitative themes from participants’ SMS text message responses and semistructured interviews about the feasibility and acceptability of HealthySMS. We posited that meaningful themes would be beneficial in informing iterative system and research protocol improvements supporting future HealthySMS effectiveness research. When examining the context of participants’ SMS text message responses to our monthly message prompts and semistructured interviews asking for feedback, we identified themes and exemplary quotes regarding the feasibility and acceptability of HealthySMS (Tables 3-5).
<table>
<thead>
<tr>
<th>Theme</th>
<th>Example quote supporting feasibility</th>
<th>Example quote about limited feasibility</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text modality</td>
<td>“I am not really great with the emails. I think just going through texts is almost always better.”</td>
<td>“For a while, I just didn’t even open the texts.”</td>
</tr>
<tr>
<td>Consistency of delivery</td>
<td>“I always got the text and sometimes if I didn’t reply, like with my mood, I would get like a reminder, maybe 10 minutes later, respond your mood. So I would say technically or technologically, it was all good. I didn’t run into any issues.”</td>
<td>“Sometimes my mood ratings just wouldn’t come in. Like some days, like towards the ending of it, they just came in periodically like not every day.”</td>
</tr>
<tr>
<td>Amount of effort to use</td>
<td>“[It was fine to] like, give a number, how was your day? But if it was like kind of describe your day, I don’t think anyone would want to do that because it would take too long.”</td>
<td>“If you didn’t like type in your mood right away, it would like send you a reminder a lot. So I got those a lot if I wasn’t doing it.”</td>
</tr>
<tr>
<td>Timing of texts</td>
<td>“I definitely didn’t get anything like super early or late. I would say they did a pretty good job of, like, changing up time, so it wasn’t like the same time every like 8AM and 10PM, like it was pretty good at switching up times so you could get like different times of the day.”</td>
<td>“I guess sometimes they came at weird times, like really early in the morning or really late at late night...It kind of felt less helpful if they came later in the day because either I had already figured it out or like got past it or it just didn’t help anymore.”</td>
</tr>
<tr>
<td>Amount of texts</td>
<td>“I don’t remember getting, like, bombarded with them. So I would say the amount is probably pretty good and reasonable.”</td>
<td>“It felt sometimes like it was getting too many. But that was mainly because that was like at a time when I didn’t need them and so it just felt like a waste of message if that makes sense.”</td>
</tr>
<tr>
<td>Text length</td>
<td>“They’re all quick...It’s short and simple and sweet...”</td>
<td>“I liked it, but it was also like kind of a lot sometimes.”</td>
</tr>
<tr>
<td>Theme</td>
<td>Example quote supporting acceptability</td>
<td>Example quote about limited acceptability</td>
</tr>
<tr>
<td>-----------------------------</td>
<td>--------------------------------------------------------------------------------------------------------</td>
<td>---------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Group reminders</td>
<td>“I liked it, because sometimes when things got a little hectic, I would forget about group so it’s a useful reminder.”</td>
<td>“They weren’t the most helpful, especially because I’m pretty sure I only ever got one text and it was just ’there is group tomorrow. Don’t forget your binder,’ which I realize is probably works a lot better when you actually have to go somewhere to do it. But like I already said, I have nothing else going on [in quarantine], so I remember yep, that’s tomorrow. I felt like there definitely could be some variation in things like maybe it could bring up, like, remember the skills you learned or don’t be scared to share or something like that, that could be a little more emotional based.”</td>
</tr>
<tr>
<td>Personalized response texts</td>
<td>“I did like how if your mood was like a bad mood, it would give you like, it wouldn’t it just be like that sucks feel better, but it would give you like advice and strategies of how you can get better. I think that was nice. And I liked when if you were in a good mood, it would kind of still give you like a different kind of advice to keep you in that space and like be like, like you could use that experience to feel better later when you remember.”</td>
<td>“...having a little bit of background about like who it is I’m responding to, like, is it a computer or is it a person like this data is being used for what, kind of thing might be helpful?”</td>
</tr>
<tr>
<td>Mood ratings</td>
<td>“The rating of the mood...it kind of helped me figure out how I was feeling, like in a number form.”</td>
<td>“I would say it was kind of easy when I was like doing super well or super bad..., but little bit harder, I guess, like to be: ‘OK, well, I don’t really know. I’m average.’”</td>
</tr>
<tr>
<td>Skill reminder texts overall</td>
<td>“I thought they really were helpful...a really nice boost, and it was and it would remind me of the other things I had learned that I could also use to feel better.”</td>
<td>“Um if I was able to do something like kind of just the reminder was helpful. But with COVID a lot of the times, it wasn’t applicable.”</td>
</tr>
<tr>
<td>Activity skill reminders</td>
<td>“I always think it’s good to like set goals for yourself so you have something to work towards. So it’s almost like a little bit of motivation.”</td>
<td>“...like: ’do a new activity’...a lot of activities have like kind of gotten harder to do with everything [in the context of the COVID-19 pandemic].”</td>
</tr>
<tr>
<td>Social skill reminders</td>
<td>“I liked the people [texts], because I struggle with my relationships with people.”</td>
<td>“Sometimes I couldn’t hang out with friends. So then sometimes that even made me, like, a little frustrated.”</td>
</tr>
<tr>
<td>Cognition skill reminders</td>
<td>“I think sometimes like I get so stuck in like the past or just like in the moment that it’s good to just like think about your future and like the good things that are to come.”</td>
<td>“...it was just odd getting text telling me to, like, change the way I’m thinking because I know like the way my brain works, it’s not going to just happen...So it was just annoying because I would like to make [the thoughts] go away, but they’re not going to.”</td>
</tr>
<tr>
<td>Statements vs questions</td>
<td>“I felt that they were good, I like them more when they were more statement based...I felt that just a clear like statement or like advice boost helped more for me personally.”</td>
<td>“I got confused a bit with the questions they’d ask because I wasn’t sure if I was supposed to, like, respond to them and it would respond back or it was just kind of a moment to reflect.”</td>
</tr>
</tbody>
</table>
Table 5. Qualitative themes uncovered and example quotes regarding HealthySMS’s potential impact on evidence-based intervention (EBI) engagement.

<table>
<thead>
<tr>
<th>Theme</th>
<th>Example quote supporting impact on EBI engagement</th>
<th>Example quote about limited impact on EBI engagement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participation in group sessions</td>
<td>“It was just like a reminder to like start thinking about group. So like before group, I’d just like start thinking about it so I’d like have more to say.”</td>
<td>“I wouldn’t really say that the text messages changed my experience in the actual group. It felt more like kind of a recap throughout the week, and less than, sort of a second part of group, if that makes sense. Yeah like we talked about the text messages a bit in group, but the text messages were always about the thing that happened last week, and so we’d want to move on. It could be helpful if the text messages like cover, like some newer things to introduce you, but um yeah.”</td>
</tr>
<tr>
<td>Group homework completion</td>
<td>“That was kind of like a good check in where I was like, OK, did I have homework? What was it? Could that help me? So I would keep that in there.”</td>
<td>“I would definitely do techniques more if I got a text message reminding me to do it [the specific homework rather than the general skill reminder] in the week.”</td>
</tr>
<tr>
<td>Connection with others in the group</td>
<td>“It felt like we were closer because we were all getting the same messages and we were all in the same boat together.”</td>
<td>“My experience with my group and group leader was pretty good, but I wouldn’t really say my experience in group and my experience with the text messages were linked.”</td>
</tr>
<tr>
<td>Validation and support</td>
<td>“And it felt like someone was like caring about me.”</td>
<td>“But it also could backfire on if you don’t do it [the suggestion in the text], feeling guilty.”</td>
</tr>
<tr>
<td>Impact on mood</td>
<td>“It kind of gave me some positivity boost, and it was nice to have someone or something to talk to me while I was feeling that way.”</td>
<td>“I wasn’t doing anything because I couldn’t go out and it just made me think about it more...So it just made me a little bit more sad.”</td>
</tr>
<tr>
<td>Behavior change and skill use</td>
<td>“I got something like the. call or spend time with people who make you feel happy...And I ended up calling one of my closest friends and she did bring up my mood a little bit...”</td>
<td>“I didn’t usually act on it, probably just because it was like, OK, well, right now I’m doing something. So then I wouldn’t really remember to do it later, but I can see how, what the idea was.”</td>
</tr>
</tbody>
</table>

Hypothesis 4

Our final hypothesis was that we would uncover meaningful qualitative themes from participants’ SMS text message responses and semistructured interviews about the potential impact of HealthySMS on EBI engagement. We posited that meaningful themes would be beneficial in informing decisions on which mechanisms of change to evaluate in future HealthySMS effectiveness research. When examining the context of participants’ SMS text message responses to our monthly message prompts and semistructured interviews asking for feedback, we identified themes and exemplary quotes regarding the potential impact of HealthySMS on EBI engagement (Table 5).

Safety During the HealthySMS Intervention

A secondary aim of our pilot study was to monitor the HealthySMS safety keyword alert triggers and provider responses to inform system and research protocol adjustments before future HealthySMS research. During the study period, 76 (7.58%) of the 1002 total SMS text messages sent by participants alerted providers to potential suicide risk throughout our flagged keyword system. When examining the content of participants’ SMS text message responses that were flagged for risk of suicide, only 2 (3%) of these 76 messages were determined to contain true risk-related content (eg, texts about wanting to hurt oneself or die by suicide). In one case, the notified provider determined that the text may be an indication that the participant was about to self-harm, and they followed the clinic’s safety protocol; no indication that harm occurred was received, and the participant continued attending the group sessions and engaging in the study. In the second case, the notified provider determined that the text did not represent an increase in risk and addressed the client’s worry about the future in their following session.

Discussion

Principal Findings

The results of this feasibility pilot study demonstrate that the use of HealthySMS adjunct to adolescent group CBT depression services (CBT-D) appears feasible and acceptable, as evidenced by high rates of HealthySMS initiation and low rates of dropout, as well as meaningful themes uncovered from participants’ qualitative feedback. Importantly, the findings also provide evidence regarding iterative improvements to the HealthySMS system and research protocol, as well as potential mechanisms of change for enhanced EBI engagement and, ultimately, adolescent depression outcomes, which can be used in future effectiveness research. It is compelling that the results of this study were obtained in the context of ongoing clinical services at a real-world outpatient clinic experiencing a transition to telehealth services amidst the onset of the COVID-19 pandemic and that the research protocol did not alter the clinical service procedures in any way, such as by incentivizing adolescents to attend the group sessions. Furthermore, it should be noted that we were able to implement this intervention in a safe manner during at a time when an increasing number of youths were at a risk for suicide; the 2 instances of HealthySMS alerts indicating risk were managed via clinical procedures, and no adverse outcomes occurred to any participant during the study.
HealthySMS Feasibility and Acceptability

As predicted, adolescents enrolled in and maintained the use of HealthySMS at high rates; in fact, no adolescent who agreed to the research study declined initiation of HealthySMS, and only 1 (6%) of the 17 enrolled adolescents opted out after a month of use. HealthySMS response rate was slightly lower than predicted but, of note, varied among participants. Some adolescents had very low response rates (ie, n=1, 6% never responded to any messages), and some adolescents had high response rates (ie, n=1, 6% responded multiple times to most prompts). Overall, most adolescents responded more than half of the time to daily mood rating prompts (ie, response rates averaging >60%). Qualitative feedback suggested that quick and short messages, as well as midday versus early or late message timing, may be the most feasible to respond to. In addition, some adolescents shared that they were motivated to respond by the “interactive” nature of the mood rating texts, which triggered responses 20% of the time. We received feedback that the number of HealthySMS messages and reminders felt appropriate for some participants; however, we received other feedback indicating a preference for less frequent messages and reminders. A future direction to explore is whether personalizing the message and reminder timing and frequency (eg, by requesting participants to share their preferences in the initial survey and monthly feedback prompts to adapt the system to allow for differences by preference) may increase response rates. Interestingly, some adolescents also sent SMS text messages back to the skill reminder messages (ie, response rates averaging 12%), although there were no explicit requests for responses. Adding explicit response requests to skill reminder messages may be another way to explore the potential benefits of HealthySMS.

Regarding the HealthySMS safety triggers, the system appeared to appropriately monitor risk. Throughout the 10 months of the study, 76 (7.58%) of the 1002 total messages sent by participants during the study period triggered provider alerts via flagged keywords (eg, cut, kill, and die; listed in Multimedia Appendix 1). We added words to the system for the current trial after consulting with data from the adolescent crisis text line. Importantly, only 2 of the alerts in this study indicated a potential suicide risk and required clinical follow-up. Providers complied with the established clinical policies and procedures, and no known adverse outcomes occurred among participants throughout the study. In fact, the vast majority of alerts were false positives (eg, “I just got my haircut.” “I went for a bike ride over the bridge.” and “my throat hurts”). One of the most sensitive keywords appeared to be “end,” given that it is a relatively frequent used word and is included in many words (eg, “almost the end of the day” and “FaceTime with friend”). This information can be used to iteratively improve the HealthySMS system to balance the need to detect safety concerns with high sensitivity while trying to avoid false positives and, therefore, reducing providers’ burden in reviewing the alerted SMS text messages, such as by programming keyword alerts to trigger only when the word “end” by itself is sent and not when it is included in other words, such as “friend.” It will be important for future HealthySMS efforts to consider how future HealthySMS research designed to evaluate the appropriateness of the safety keyword triggers and subsequent provider responses to prevent self-harming behaviors would be beneficial.

Potential Impact on EBI Engagement to Be Evaluated in Future Effectiveness Research

When asked about the utility of HealthySMS, we received input providing initial evidence that the system may indeed enhance EBI engagement. Our identified themes may be used to inform decisions on which mechanisms of change to evaluate in future HealthySMS effectiveness research designed to measure the impact of the system on EBI engagement and, ultimately, adolescent depression outcomes. To begin, most adolescents gave feedback that group reminder texts may have increased their likelihood to attend and meaningfully participate in the group sessions, although one of the participants felt that the group reminder texts were intrusive and did not impact engagement in the group sessions. Variation in responses again highlighted that a personalized SMS text messaging system would be ideal for accommodating individual needs and desires. Qualitative responses also suggest that the HealthySMS skill messages provided beneficial reminders and motivation for some adolescents to engage in CBT strategies, including behavioral activation and helpful thinking. Some participants did suggest that we sync skill messages with the content being delivered each week (rather than with each module, as is the current setup) to optimize relevance. Although weekly agendas are generally known from the beginning, given the CBT-D format in this study, group providers may adjust agendas to meet the needs of the current group members. Thus, personalizing the timing and content of messages and syncing messages more precisely to weekly content would likely be perceived as helpful by participants but would need to be weighed against the feasibility of such measures, which could increase providers’ burden.

We were also interested in the impact of HealthySMS on the interpersonal connection of the adolescents in the current trial, given the reports of adults in our team’s prior trial of HealthySMS that participants felt cared for and supported by the messages, as well as closer to the CBT group (sources removed for masked review). We received feedback suggesting that HealthySMS may help some adolescents feel more connected to the group leader and other members, as well as validated and supported even with the knowledge that it is a “bot” responding. One of the providers shared that a participant felt that HealthySMS was like a “friend.” Feedback from participant interviews indicated that this sense of connection was especially true when they received the responses to their mood rating texts, which were aligned with how they rated their mood (ie, different responses were sent for high, medium, or low mood ratings).

Limitations and Future Directions

Our study has several limitations that should be acknowledged and addressed in future work. First, our pilot study investigating the feasibility and acceptability of HealthySMS in a real-world treatment setting was not designed to evaluate the system’s role in improving the ultimate target of adolescent depression. To accomplish this, future effectiveness research building on the
lessons learned and iteratively updated deliverables (ie, the HealthySMS system and research protocol) from this study is warranted. Specifically, although we were able to explore the potential impact of HealthySMS on EBI engagement in participants’ qualitative feedback, we did not have the data or sample size required to quantitatively evaluate this construct. In addition, our decision to conduct this pilot study in a real-world setting without making changes to existing clinical procedures created several potentially confounding variables among participants, such as the number of other EBI services engaged in adjunct to CBT-D. Subsequent studies fully powered to detect mechanisms of change as well as control for potential covariates to EBI engagement and outcomes using a control group are called for. Finally, this pilot study took place in the initial months of SIP enforcement owing to the COVID-19 pandemic; thus, it is not known how well our findings will generalize beyond this context.

**Conclusions**

Our pilot study suggests that HealthySMS adjunct to the most evidence-based-treatment for adolescent depression (ie, CBT) is feasible and acceptable, warranting future effectiveness research evaluating the system’s impact on adolescent EBI engagement and subsequent depression outcomes. As SMS text messaging is cheap and uses technology already in the hands of most adolescents, it is well suited to be added to existing clinical services. Feedback from the participants in our studies suggested that mHealth may be particularly helpful during times of SIP enforcement, given the limited ability for adolescents to engage in activities, social interactions, in-person mental health treatment, and safety monitoring by adults in their lives (eg, teachers and providers). However, it is possible that SMS text message systems are beneficial adjuncts to EBIs in all contexts, given their potential to increase the likelihood and effectiveness of service participation, enhance feelings of connectedness and validation, and encourage skill use in-between sessions. Continued service research on the implementation and effectiveness of mHealth tools has the potential to improve mental health services for a population experiencing drastic increases in depression and suicide risk: adolescents.

**Conflicts of Interest**

AA is the owner of the HealthySMS program license and has licensed it to other researchers for use in their studies. He was not paid for the license for this study.

Multimedia Appendix 1

Risk alert words.

[DOCX File, 15 KB - mental_v11i1e49317_app1.docx ]

**References**


Abbreviations

BRIGHT: Building Recovery by Improving Goals, Habits, and Thoughts
CBT: cognitive behavioral therapy
CBT-D: Cognitive Behavioral Therapy for Depression Group for Adolescents
dHealth: digital health
EBI: evidence-based intervention
eCBT: electronic cognitive behavioral therapy
mHealth: mobile health
SIP: shelter-in-place
UCSF: University of California San Francisco
Translating Suicide Safety Planning Components Into the Design of mHealth App Features: Systematic Review

Kim Gryglewicz1*, MSW, PhD; Victoria L Orr2, BS; Marissa J McNeil2, MA; Lindsay A Taliaferro3, MPH, PhD, CHES; Serenea Hines2, MSW; Taylor L Duffy2, BS; Pamela J Wisniewski4*, PhD

1School of Social Work, University of Central Florida, Orlando, FL, United States
2Center for Behavioral Health Research & Training, University of Central Florida, Orlando, FL, United States
3Department of Population Health Sciences, University of Central Florida, Orlando, FL, United States
4Department of Computer Science, Vanderbilt University, Nashville, TN, United States
*these authors contributed equally

Corresponding Author:
Kim Gryglewicz, MSW, PhD
School of Social Work
University of Central Florida
12805 Pegasus Drive HS I
Orlando, FL, 32816
United States
Phone: 1 14078232954
Email: kgryglew@ucf.edu

Abstract

Background: Suicide safety planning is an evidence-based approach used to help individuals identify strategies to keep themselves safe during a mental health crisis. This study systematically reviewed the literature focused on mobile health (mHealth) suicide safety planning apps.

Objective: This study aims to evaluate the extent to which apps integrated components of the safety planning intervention (SPI), and if so, how these safety planning components were integrated into the design-based features of the apps.

Methods: Following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines, we systematically analyzed 14 peer-reviewed studies specific to mHealth apps for suicide safety planning. We conducted an analysis of the literature to evaluate how the apps incorporated SPI components and examined similarities and differences among the apps by conducting a comparative analysis of app features. An independent review of SPI components and app features was conducted by downloading the available apps.

Results: Most of the mHealth apps (5/7, 71%) integrated SPI components and provided customizable features that expanded upon traditional paper-based safety planning processes. App design features were categorized into 5 themes, including interactive features, individualized user experiences, interface design, guidance and training, and privacy and sharing. All apps included access to community supports and revisable safety plans. Fewer mHealth apps (3/7, 43%) included interactive features, such as associating coping strategies with specific stressors. Most studies (10/14, 71%) examined the usability, feasibility, and acceptability of the safety planning mHealth apps. Usability findings were generally positive, as users often found these apps easy to use and visually appealing. In terms of feasibility, users preferred using mHealth apps during times of crisis, but the continuous use of the apps outside of crisis situations received less support. Few studies (4/14, 29%) examined the effectiveness of mHealth apps for suicide-related outcomes. Positive shifts in attitudes and desire to live, improved coping strategies, enhanced emotional stability, and a decrease in suicidal thoughts or self-harm behaviors were examined in these studies.

Conclusions: Our study highlights the need for researchers, clinicians, and app designers to continue to work together to align evidence-based research on mHealth suicide safety planning apps with lessons learned for how to best deliver these technologies to end users. Our review brings to light mHealth suicide safety planning strategies needing further development and testing, such as lethal means guidance, collaborative safety planning, and the opportunity to embed more interactive features that leverage the advanced capabilities of technology to improve client outcomes as well as foster sustained user engagement beyond a crisis. Although preliminary evidence shows that these apps may help to mitigate suicide risk, clinical trials with larger sample sizes and more robust research designs are needed to validate their efficacy before the widespread adoption and use.
suicide prevention; suicide safety planning; mobile health; mHealth apps; eHealth; digital health; systematic review; Preferred Reporting Items for Systematic Reviews and Meta-Analyses; PRISMA

Introduction

Background

Suicide is one of the leading causes of death in the United States, accounting for >45,000 deaths annually [1]. Over the last decade, suicide rates have doubled for youth aged 10 to 24 years [2] and have steadily increased for racial and ethnic minority youth [1,3,4]. Suicide ideation and attempt rates have also risen [5,6], especially among youth and minoritized populations [5,7-11]. Numerous studies have shown that untreated mental illness, limited or lack of available care, and low perceived need for mental health treatment are common, yet preventable, suicide risk antecedents [12-19]. Moreover, stigma, difficulties recognizing suicide warning signs, preferences for self-reliance and autonomy, fear of burdening others, and negative treatment experiences can negatively affect help-seeking intentions and engagement in mental health services [20-24].

Researchers have identified various suicide prevention strategies to reduce the public health problem of suicide [25,26]. Safety planning is an integral component of suicide care [27] and has been empirically validated for reducing suicidality [28,29]. The process of safety planning involves collaboration between a clinical and client, as well as with the at-risk individual and their support network. This means that the support network could also be part of the safety planning process [30]. Safety planning involves jointly identifying, problem-solving, and communicating strategies to keep an individual safe if a crisis arises. Core strategies focus on uncovering warning signs or triggers that precede an emotional event, identifying and reinforcing the use of healthful self-management strategies to cope with distress, encouraging the use of positive socialization strategies for distraction and support, creating a network of external support and professional contacts to solicit assistance and support, and reducing access to lethal means [31]. The individualized nature of creating a safety plan (ie, a written document detailing the plan to keep an individual safe during a crisis) allows the person at risk of suicide the ability to incorporate culturally relevant and meaningful strategies, thereby making these plans useful and relevant for diverse populations [30,32].

Suicide safety planning is a brief intervention that has been used in both acute and clinical settings [31,33,34] and as a self-help tool [35]. Overall, researchers have found this intervention to be feasible, acceptable, and useful to facilitate support and reduce suicide risk [32,33,35-37]. Researchers have found safety plans and related interventions, such as crisis response planning [38], to be effective in reducing the risk of hospitalization, increasing engagement in mental health treatment, and promoting the use of healthful coping strategies when used alongside other therapeutic approaches [33,34,36,39,40]. Although safety planning has shown initial success in reducing suicidal urges and offering a sense of hope to individuals in crisis [41], some clinicians and researchers have criticized this process [42,43]. For example, safety planning encourages clinicians to revisit and update safety plans with their clients over time [44], which can prove challenging if service use barriers prevent clients from reaccessing care or if clients misplace or throw away their paper-based safety plan.

Considering these challenges, mobile health (mHealth) technologies could offer a timely and effective solution to address some of the criticisms directed at traditional safety planning methods. mHealth, particularly the use of apps, represents a common tool used by consumers with access to mobile phones [45,46]. In addition, mHealth has garnered attention as a practical and convenient method for implementing mental health interventions [47], with increase in the quantity and functionality of applications and tools resulting in increased use [48]. In general, mHealth apps have been used to effectively help individuals identify and manage symptoms of various mental health problems and conditions such as depression, anxiety, substance abuse, posttraumatic stress, and eating disorders [49,50]. Thus, incorporating mHealth apps into mental health treatment and adjunctive interventions may prove beneficial.

Furthermore, incorporating mHealth apps into established evidence-based interventions may also serve as a culturally inclusive way of disseminating treatment to younger, more technologically savvy generations who also happen to demonstrate higher rates of suicidal thoughts and behaviors than adults [6]. mHealth apps may also help address service use barriers and risk factors (eg, stigma) that hinder individuals from seeking help and participating in treatment for suicidality. Combining suicide safety planning practices with mHealth apps may combat accessibility concerns as well, including a commonly reported flaw of the traditional intervention—the reliance on a paper format [35]. Given the widespread proliferation of mHealth apps for suicide prevention, there is a need to examine the components and features that have been incorporated into the design of suicide safety planning apps.

Objectives

The purpose of this systematic literature review was to first assess the extent to which suicide safety planning mHealth apps integrated the 6 steps or components of a widely used safety planning intervention (SPI) developed by Stanley and Brown [31] (research question [RQ] 1). Next, we independently reviewed available mHealth suicide safety planning apps via download from iOS and Android app stores to assess the integration of SPI components and to categorize different app design features used to personalize the end users’ experience (RQ2). We also examined the evidence on the effectiveness of these apps in terms of usability, acceptability, app engagement, and suicide-related outcomes (RQ3). This review aims to synthesize the extant research to inform suicide prevention
efforts, clinical practice, and future development of suicide safety planning mHealth apps.

Methods

Overview

In accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 statement guidelines (Multimedia Appendix 1 [51]), a comprehensive systematic review of existing literature on suicide safety planning via mHealth apps was conducted. The process is described in the following sections.

Systematic Literature Review

Eligibility Criteria

The inclusion criteria for the reviewed research studies were as follows: (1) a primary focus on suicide safety planning involving the use of a mHealth app, (2) publication in a peer-reviewed article written in English, and (3) availability of the full text of the article. Studies were excluded if (1) the word suicide, safety plan, or app was not included in the title; (2) they included other forms of mHealth technologies as the primary focus (eg, web-based applications); (3) the apps were designed with safety planning as a secondary focus (ie, not exclusively for suicide safety planning, not intended as a crisis intervention, or use of safety planning as a secondary tool to other treatment modalities); and (4) they were part of other systematic reviews or meta-analyses. We included studies across the entire system development life cycle (eg, formative evaluations and 1 group pre-posttest designs) owing to limited research on the topic and the relatively recent emergence of such research.

Information Sources

The following 5 bibliographic databases were used to systematically review the literature: PsycINFO, PubMed, ACM Digital Libraries, Academic Search Premier, and ERIC. We limited our results to articles published between January 2000 and May 2023. All databases were last searched on July 2, 2023.

Search Strategy

We used the following keywords to search for the topic of interest in each scientific database: “Safety Plan*” AND (“Applications” OR “Apps”); (“Suicide” OR “Safety Plan*”) AND (“Applications” OR “Apps”); “Suicide Interven*” AND (“Applications” OR “Apps”); “Suicide Prevent*” AND (“Applications” OR “Apps”); “Suicide Contract” AND (“Applications” OR “Apps”); “mHealth” AND “Suicide”; “Crisis Response” AND “Plan*.” Asterisks were added to search for words that began with the preceding letters (eg, prevent*: prevent, prevention, and preventing). An example of the search strategy outlined above is provided in Multimedia Appendix 2.

Selection Process

Citations obtained from electronic databases were imported into Zotero (version 6.0.16). Two reviewers (KG and VLO) independently screened the articles to remove duplicates and assessed inclusion and exclusion criteria by title and abstract. For articles about which the reviewers were uncertain after the title and abstract review, 4 reviewers independently analyzed the full-text articles to determine whether they met the inclusion criteria. The reviewers discussed discrepancies until they reached a consensus. The references of all articles that met the inclusion criteria were reviewed and cross-referenced for additional relevant articles. We included all eligible studies (N=14) in this systematic review (Figure 1 [51]).
Figure 1. Flowchart of the studies in line with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines. mHealth: mobile health.

Data Collection Process

Data from eligible studies were analyzed using the Cochrane Collaboration’s data extraction template for included studies (version 1.8) [52]. We added study-specific items to the template to answer RQ1 and RQ2. Specifically, to answer RQ1, we reviewed articles describing each mHealth app and coded, using a dichotomous (yes or no) coding scheme, for the following SPI components: (1) personal warning signs, (2) coping strategies, (3) ways to distract oneself through social activities, (4) identification of and ways to access trusted individuals (eg, family and friends) for support, (5) identification of and ways to access community supports (eg, mental health professionals, nonmental health adult supports, crisis, or emergency services), and (6) information about keeping the environment safe (eg, restricting access to lethal means). To answer RQ2, we downloaded available mHealth apps via the Apple App Store or Google Play Store or contacted app developers to conduct an independent review of SPI components and app features described in the articles. Next, we created codes to describe app features, organized and categorized codes based on similarities, and generated 5 themes to capture the core aspects of features. To answer RQ3, we extracted both qualitative and quantitative findings reported on primary and secondary outcomes. We categorized the study outcomes into 3 main research themes. Two reviewers coded 2 research articles to assess interrater reliability based on the coding template and made refinements as necessary (eg, added operational definitions to describe SPI components and provided examples of app features). Once finalized, the reviewers used the template to extract the data from the remaining studies. Data items included (1) general article information (eg, author, publication year, and country); (2) study methods (eg, aims and research design); (3) study characteristics (eg, sample size, sample demographics, and setting); (4) SPI intervention characteristics (RQ1); (5) mHealth app design features (RQ2) and primary and secondary outcomes (RQ3); and (6) study implications and future directions (Multimedia Appendix 3 [42,43,53-64]). A similar process was used to independently code the SPI components and app features of the mHealth apps available for download.

Risk of Bias Assessment

The risk of bias for each study was assessed by 2 independent reviewers (KG and VLO) using Joanna Briggs Institute (JBI) appraisal tools for quasi-experimental [65] and qualitative research study designs [66]. For studies that included mixed methods designs, we used both tools as recommended by the
JBI. Each appraisal tool used a rating scale with yes, no, unclear, and nonapplicable responses. The overall appraisal rating was based on the following categories: include, exclude, and seek further information. Disagreements between the reviewers were discussed until they reached a consensus (Multimedia Appendix 4 [65,66]).

Synthesis of Results
Owing to the heterogeneity of the study designs, participants, and outcomes collected, we could not perform a meta-analysis of the identified studies in this review. Therefore, we present a narrative synthesis of the study findings.

Results

Study Selection
The initial search of electronic databases and hand-searched references resulted in a total of 46,397 peer-reviewed articles. After duplicate records were removed, 21,151 studies remained. Titles were screened for relevancy (eg, relating to suicide, suicide safety planning, and mHealth apps), and 20,970 articles were excluded. A total of 181 abstracts were reviewed. Following full-text reviews of 54 articles, 40 articles were removed (15 studies did not include an mHealth app and 25 were not intended as a suicide safety planning app). A total of 14 articles met the inclusion criteria (refer to Figure 1 for breakdown).

Study Characteristics

Overview
The detailed study characteristics of the selected articles (N=14) are presented in Multimedia Appendix 3. Most studies (12/14, 86%) were conducted outside the United States [42,53-63]. The year range of the selected articles was between 2015 and 2023.

Study Design
As shown in Tables 1 and 2, a total of 7 mHealth suicide safety planning apps were studied across the 14 articles in our data set (Multimedia Appendix 3). We classified the articles based on the research design (ie, formative feedback, usability assessment, single cohort pre-posttest, and random control trial protocol). Formative designs assessed SPI components and features to guide app development [43,56,61,64], whereas usability designs assessed interface design issues and functionality (eg, task difficulty and time to complete tasks) [55,60,61,64]. Other studies evaluated the acceptability or feasibility of a fully developed mHealth app [54,58-60,62,63]. Across these studies, participants rated the frequency and duration of app use; ease of navigation; and level of satisfaction, comfort, confidence, or engagement in using the app.

Table 1. Formative, usability, and acceptability assessments of mobile health suicide safety planning apps (n=8 articles).

<table>
<thead>
<tr>
<th>Safety planning app</th>
<th>Year span</th>
<th>Formative feedback only</th>
<th>Usability assessment</th>
<th>Random control trial</th>
<th>Use period</th>
<th>Key findings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unnamed [43]</td>
<td>2015</td>
<td>Clinicians (n=9), at-risk youth, and guardians (n=20)</td>
<td>N/A(^a)</td>
<td>None</td>
<td>N/A</td>
<td>Qualitative feedback to inform app design; no app developed</td>
</tr>
<tr>
<td>SERO [56]</td>
<td>2022</td>
<td>Clinicians, at-risk individuals (n=11), and relatives (n not reported)</td>
<td>N/A</td>
<td>None</td>
<td>N/A</td>
<td>Summary of 6 suicide prevention strategies aligned with app design</td>
</tr>
<tr>
<td>ED-SAFE [64]</td>
<td>2023</td>
<td>Clinicians and subject matter experts (n=7), and at-risk adults (n=6)</td>
<td>Emergency department patients after discharge (N=14)</td>
<td>N/A</td>
<td>Unclear</td>
<td>High usability scores and low uptake (use); no significant clinical outcomes reported</td>
</tr>
<tr>
<td>MYPLAN [53,54,62]</td>
<td>2016, 2017, and 2020</td>
<td>At-risk youth, adults, relatives, and clinicians (n=26)</td>
<td>N/A</td>
<td>Protocol only (N=546)</td>
<td>Familiarity with app</td>
<td>Qualitative feedback only; no clinical outcomes assessed</td>
</tr>
<tr>
<td>SafePlan [57,61]</td>
<td>2020 and 2023</td>
<td>Clinicians and experts (N=15)</td>
<td>Students (N=18)</td>
<td>Protocol only (N=80)</td>
<td>Single use</td>
<td>High usability scores; no clinical outcomes assessed</td>
</tr>
</tbody>
</table>

\(^a\)N/A: not applicable.
Table 2. Pre-posttest assessments of mobile health suicide safety planning apps (n=6 articles).

<table>
<thead>
<tr>
<th>Safety planning app name</th>
<th>Year span</th>
<th>Formative feedback only</th>
<th>Usability assessment</th>
<th>Single cohort pre-posttest</th>
<th>Use period</th>
<th>Key findings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brake of My Mind [55]</td>
<td>2020</td>
<td>Expert heuristic evaluations (N=5)</td>
<td>Clinician interviews (N=6)</td>
<td>At-risk youth (N=3)</td>
<td>1 week</td>
<td>High usability scores; sample size too low to assess clinical efficacy</td>
</tr>
<tr>
<td>BackUp [42,60,63]</td>
<td>2017, 2018, and 2022</td>
<td>Designed informed by the expert panel (n=8) and at-risk adults (n=21)</td>
<td>Usability assessment combined with pre-posttests</td>
<td>Protocol only, single cohort design (N=80); at-risk adults (n=21) and at-risk adults (n=12)</td>
<td>1 week and 3 months</td>
<td>High usability scores, no significant decrease in suicidal thoughts; low clinical and patient uptake (use)</td>
</tr>
<tr>
<td>BeyondNow [58,59]</td>
<td>2019 and 2020</td>
<td>N/A^a</td>
<td>N/A</td>
<td>At-risk youth and adults (n=22) as well as at-risk youth (n=17)</td>
<td>2 months and 6 weeks</td>
<td>High usability scores. Significant reduction in severity and intensity of suicidal ideation; significant increase in coping; and no significant change in suicide resilience for mixed samples. For youth only sample, no significant decrease in suicidal thoughts and significant increase in suicidal resilience. No conclusions regarding clinical efficacy.</td>
</tr>
</tbody>
</table>

^aN/A: not applicable.

Sample Characteristics

Across studies, the study sample varied in age, type of participant (eg, youth or adults at risk of suicide and clinicians collaborating with suicidal clients), and setting (eg, suicide prevention clinic and pediatric inpatient facility). Among studies that recruited participants to inform or evaluate mHealth suicide safety planning apps [43,54-56,58-64], the sample size ranged from 11 to 36 participants. However, after reporting dropout rates, sample sizes dropped to as low as 2 participants and as high as 22 participants.

Integration of SPI Components Within mHealth Apps

Most articles (5/7, 71%) describing the mHealth apps incorporated SPI components into the design of their apps [54,58,61,63,64] (Table 3). Creating a safe environment from lethal means was the missing component in 29% (2/7) of the apps [55,56].
Table 3. Safety planning intervention (SPI) components and app features.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SPC&lt;sup&gt;b&lt;/sup&gt;</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SPI 1: warning signs</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>SPI 2: coping strategies</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>SPI 3: distractions social activities</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>SPI 4: trusted supports (family and friends)</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>SPI 5: community supports (MHP&lt;sup&gt;d&lt;/sup&gt;)</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>SPI 6: safe environment (lethal means)</td>
<td>X</td>
<td>_&lt;sup&gt;c&lt;/sup&gt;</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td><strong>Interactive features&lt;sup&gt;f&lt;/sup&gt;</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Links stressors to cope strategies; SPI 1 and 2</td>
<td>X</td>
<td>—</td>
<td>—</td>
<td>X</td>
<td>—</td>
<td>—</td>
<td>X</td>
</tr>
<tr>
<td>Inclusion of media (distraction); SPI 3</td>
<td>X</td>
<td>X</td>
<td>✔&lt;sup&gt;g&lt;/sup&gt;</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>—</td>
</tr>
<tr>
<td>Access to trusted supports; SPI 4</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Access to community supports; SPI 5</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>GPS tracking; SPI 5 and 6</td>
<td>X</td>
<td>X</td>
<td>—</td>
<td>✔</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td><strong>Individualized user experience</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Revisable safety plan</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Personality and mood exercises or tracking</td>
<td>—</td>
<td>X</td>
<td>—</td>
<td>X</td>
<td>—</td>
<td>—</td>
<td>X</td>
</tr>
<tr>
<td>Visual customization</td>
<td>✔</td>
<td>X</td>
<td>✔</td>
<td>X</td>
<td>X</td>
<td>—</td>
<td>X</td>
</tr>
<tr>
<td>Enabling notifications</td>
<td>X</td>
<td>—</td>
<td>—</td>
<td>X</td>
<td>✔</td>
<td>—</td>
<td>✔</td>
</tr>
<tr>
<td><strong>Interface design</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Easy to navigate</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>✔</td>
</tr>
<tr>
<td><strong>Guidance and training</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>In-app tutorial</td>
<td>X</td>
<td>X</td>
<td>✔</td>
<td>—</td>
<td>X</td>
<td>X</td>
<td>✔</td>
</tr>
<tr>
<td><strong>Privacy and sharing</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Secure username and password</td>
<td>✔</td>
<td>X</td>
<td>—</td>
<td>X</td>
<td>✔</td>
<td>X</td>
<td>✔</td>
</tr>
<tr>
<td>Shareable data and safety plan</td>
<td>X</td>
<td>—</td>
<td>X</td>
<td>X</td>
<td>—</td>
<td>X</td>
<td>✔</td>
</tr>
</tbody>
</table>

<sup>a</sup>BoMM: Brake of My Mind.
<sup>b</sup>SPC: safety planning component.
<sup>c</sup>SPC or app feature included in the app.
<sup>d</sup>MHP: mental health professional.
<sup>e</sup>SPC or app feature missing in the app.
<sup>f</sup>Denotes innovative app features aligned with SPI components.
<sup>g</sup>Feature included in the app that was not mentioned in the article.

We used the JBI quasi-experimental appraisal tool [65] to assess the risk of bias across 5 studies [55,58-60,63]. These studies did not include a control or comparison group, increasing the threat to internal validity. Pre- and posttest measures were used to assess the immediate effects of the mHealth apps. However, the lack of repeated outcome measures over time, selection bias
Each mHealth app also included a built-in feature for users to content to inspire hope and distract users from negative thinking. The use of media content in these apps was intended to serve as a coping mechanism for users experiencing distress. Three of the suicide safety planning mHealth apps \[54,56,61\] allowed users to upload or view media content (e.g., pictures, quotes, music, and videos). Users had access to their phone’s camera with the ability to upload media content. The ability to access community supports such as mental health professionals (SPI 5) was emphasized. Three apps \[54,55,61\] included GPS capabilities, which enabled users to search for nearby counseling agencies or emergency services, and, after selecting a search result, users received directions for quick access (SPI 5 and 6).

### Individualized User Experience

All apps (7/7, 100%) allowed users to continually add to or revise their safety plans. Examples included the addition of new warning signs, reasons for living, and identifying coping strategies. None of the apps maintained a historical record of the previous safety plans or provided a visual mechanism to track daily, weekly, or monthly patterns based on stressors encountered or coping strategies used. Other personalization aspects included the ability to enable or disable therapeutic modalities \[61\], the inclusion of web-based resources to take an aptitude and personality test \[55\], exercises to express moods \[55,56,61\], and mood tracking \[55,56,61\]. In addition, all apps had built-in features to make esthetic customizations, such as personalizing the home screen, changing the color palate, and adding background pictures \[54-56,58,61,63\]. In 57% (4/7) of the apps, notifications were enabled to remind users about using their safety plan or skills to practice \[54,55,61,63\].

### Interface Design

Several studies used iterative feedback from content and app design experts to create easy-to-navigate interfaces \[58,61,63\]. To enhance the navigation experience, a simple layout, clear or user-friendly language, and accessibility features were important. The use of icons, buttons, and text was optimized for easy navigation. The inclusion of web-based resources to take an aptitude and personality test \[55\], exercises to express moods \[55,56,61\], and mood tracking \[55,56,61\] was emphasized. In addition, all apps had built-in features to make esthetic customizations, such as personalizing the home screen, changing the color palate, and adding background pictures \[54-56,58,61,63\]. In 57% (4/7) of the apps, notifications were enabled to remind users about using their safety plan or skills to practice \[54,55,61,63\].

### Guidance and Training

In-app tutorials or instructional videos were included in 86% (6/7) of the suicide safety planning mHealth apps \[54-56,58,63,64\]. Some of these tutorials focused on how to use the app, whereas others explained the safety planning process. For example, the BeyondNow app \[58\] included a video tutorial on how to create a safety plan and how to use the app features. The app also included a feature to allow users to save and contact trusted individuals within their social support networks (SPI 4). Typically, users entered contact information into the mHealth app directly or linked to their contact directories. A unique feature of the MYPLAN app \[54\] allowed users to create prewritten messages that they could send to their social supports during times of distress. Although this feature was created to inform loved ones of the app user’s emotional state during a crisis, participants (i.e., app users) noted concerns about messages being misunderstood, whereas relatives felt that messages could minimize emotional states or provide inaccurate information about the app user’s safety. All apps included the ability to access community supports such as mental health professionals (SPI 5). Three apps \[54,55,61\] included GPS capabilities, which enabled users to search for nearby counseling agencies or emergency services, and, after selecting a search result, users received directions for quick access (SPI 5 and 6).

### Conclusion

In summary, the qualitative appraisal results are included in Multimedia Appendix 4. The comparative analysis of SPI components and app features was synthesized in Table 3. The qualitative appraisal tool \[66\] was used to assess the risk of bias in 4 studies \[43,54,56,62\]. Across 2 studies \[43,54\], the cultural or theoretical orientation of the researchers and their influence on the research process was unclear. These issues were noted in the other 2 studies \[56,62\] as well. In these studies \[56,62\], it was also difficult to identify the philosophical perspective and congruity between the research methods, data analysis, and interpretation. The studies included more of a description of the design of the apps and included general perceptions from stakeholders.

The remaining studies \[61,64\] were assessed using both the quasi-experimental and qualitative appraisal tools owing to their mixed methods designs. In both studies, it was unclear whether the researchers’ cultural or theoretical orientation, their influence on the research, and the adequate representation of the participants and their voices were addressed. Other key issues included the lack of a control or comparison group, nonrandom and small sample sizes, and the use of posttest measures to assess usability at only 1 time point. JBI appraisal results are included in Multimedia Appendix 4.
outlining the process of safety planning and links to other helpful information. The most extensive tutorials were seen in the companion app to ED-SAFE [64], where tutorials could be received from a female provider, a male community member, or an avatar. The mHealth suite of apps also included self-care education materials about suicidality, safety plans, and life plans. In addition, the BackUp app [63] provided supportive contacts with web-based information on ways to identify warning signs and strategies to talk with suicidal individuals. The Brake of My Mind app [55] included an introduction from the developer with additional web-based resources to increase app usability.

Privacy and Sharing

Researchers also highlighted app privacy and sharing capabilities as important features to consider when designing mHealth suicide safety planning apps. Given the personal nature of the information saved, most mHealth apps required a username and password to log in [54-56,61,63,64]. For example, ED-SAFE [64] used the username and password feature to verify user identity and connect information collected in the emergency department setting to the mHealth app. Other apps disabled GPS for location tracking or did not use external servers to store users’ information for privacy and security concerns [61,63]. Several apps (5/7, 71%) included features allowing users to share self-monitoring data or share safety plans with clinicians or trusted individuals [54,56,58,61,64]. For instance, ED-SAFE [64] allowed users to share safety plans as well as appointment information, self-care education, helplines, referrals, and distractions through password-protected privileges given to authorized family members.

mHealth App Evidence of Effectiveness

The qualitative and quantitative findings were categorized into 3 main research themes: app usability and acceptability, app use and engagement, and suicide-related outcomes.

App Usability and Acceptability Findings

Across 71% (10/14) of the studies [54-56,58-64] that assessed the initial usability or acceptability of mHealth suicide safety planning apps, stakeholders’ experiences testing the mHealth apps were generally positive. Four studies [55,60,61,64] included standard rating scales (ie, System Usability Scale [68]) to assess the perceived usability of their apps, and scores exceeded the minimum usability standards (ie, >70). The remaining studies used qualitative feedback from focus groups, case reports, and open-ended questionnaires. For example, in the study by Buus et al [54], participants found the MYPLAN safety planning app useful in recognizing patterns of impending crises and for reinforcing personalized strategies to cope with distress. In describing the benefits of the BeyondNow safety planning app, participants in the study by Melvin et al [58] reported developing a sense of hope and connection from using the app. Researchers have attributed these findings to the accessibility of the app and its customizable features. According to the authors, stakeholders regarded apps as highly intuitive, easy to use, and visually appealing interface in terms of the design [59,61,62,64].

App Use and App Engagement

Five studies examined app use over time [58-60,63,64]. Overall app engagement and use were minimal. Across 3 studies, >70% of the participants used the apps at least once during the testing period, which ranged from 1 to 10 weeks [58,59,63]. In the study by Melvin et al [58], 77% (17/22) of the participants reported using the mHealth app “occasionally” or “a lot,” including to make changes to safety plans. Most participants also reported using the mHealth app during a suicidal crisis (15/22, 68%) or when experiencing suicidal ideation (18/22, 82%). Increased frequency of app use during a crisis or among participants with high levels of suicide ideation was reported in studies by Pauwels et al [63] and Muscara et al [59]. Larkin et al [64] reported that 2 (40%) out of 5 participants reported downloading the ED-SAFE patient mHealth app after discharge. Low uptake rates were mostly attributed to the participants’ forgetfulness to download the app. Although most participants acknowledged the benefits of using mHealth suicide safety planning apps during times of crisis [58,63], participant feedback from the study by Muscara et al [59] suggested that participants did not believe or were unsure whether the use of the BeyondNow safety planning app could help them manage their symptoms or keep individuals safe during a crisis. Only 35% (6/17) of the participants favored using the app in the future. Conversely, participants in the study by Nuij et al [60] noted that easy access to the Backup mHealth app provided a sense of reassurance and helped to deter suicidal thoughts.

Suicide-Related Outcomes

Suicide-related outcomes were examined across 29% (4/14) of the small-scale pilot studies (with sample sizes ranging between 3 and 22) [55,58,59,63]. The study by Jeong et al [55] assessed the Theory of Planned Behavior constructs, including attitudes, subjective norms, perceived behavioral control, and intentions toward engaging in suicide attempts, using a pre-posttest design with a small (N=3) sample of adolescent survivors of suicide attempts. The results showed statistically significant changes in attitudes, perceived behavioral control, and intentions, suggesting that the suicide safety planning app helped to positively shift attitudes toward life and reduce beliefs and intentions to engage in self-harm behavior.

Suicide coping or resilience was evaluated in 2 studies using post-test designs [58,59]. Both studies used the same safety planning app (ie, BeyondNow) to examine the changes in protective factors. Melvin et al [58] found a statistically significant increase in suicide-related coping among youth and adult participants (n=22). This finding suggests an increase in knowledge and confidence to use internal coping strategies and external resources to manage suicide ideation. However, the researchers did not observe statistically significant changes in suicide resilience (ie, the perceived ability to manage suicidal thoughts and feelings). In contrast, Muscara et al [59] found a significant increase in 1 subscale of suicide resilience, emotional stability (ie, the ability to regulate emotions), among youth participants (N=17) in their study.

Suicidal ideation or self-harm behavior were measured in 3 studies [58,59,63]. In an open-label, single-group design, Melvin et al [58] found statistically significant reductions in both the
severity and intensity of suicide ideation following exposure to an 8-week trial that evaluated the clinical effectiveness of using the Beyond2Now suicide safety planning app as an adjunct to treatment as usual (ie, existing mental health services). In an evaluation of the same mHealth app, but with the addition of a personalized toolbox app (ie, BlueIce), instead of treatment as usual, Muscara et al [59] also found a reduction in suicide ideation and self-harm behaviors (ie, attempts to harm oneself with and without suicidal intent). However, these findings were not conclusive or statistically significant owing to the small sample size and lack of a control group. Pauwels et al [63] found a similar, nonsignificant decrease in suicide ideation scores in a study examining pre-posttest changes following exposure to the BackUp suicide safety planning app. Although these studies provide some evidence of clinical utility, these researchers noted study limitations and the need for further evaluation using randomized controlled trials (RCTs).

Discussion

Principal Findings

The primary aim of this study was to conduct a comprehensive analysis of the integration and inclusion of the SPI components developed by Stanley and Brown [31] in the design of mHealth suicide safety planning apps. The secondary aim was to synthesize and assess the research methods of studies that reported on the effectiveness of these apps. Implications of these findings and practical recommendations for future directions in mHealth suicide safety planning research are described in the following sections.

Integrating Components of Suicide Safety Planning Into mHealth Apps

Overall, most apps included the core components of the SPI developed by Stanley and Brown [31], such as the identification of suicide warning signs, coping strategies, and supportive persons. Therefore, the results from this review provide evidence of some level of successful integration of SPI components into mHealth suicide safety planning apps (RQ1). Lethal means safety was 1 component that was not incorporated in 2 of the apps reviewed. Reducing access to lethal means is a critical part of suicide safety planning [31] and warrants inclusion in mHealth apps as it brings attention to methods that could be used to attempt or die by suicide if not removed from a user’s environment.

An important aspect of suicide safety planning is access to one’s safety plan. In this review, having access to safety plans at any time [54,55,58,60,61] and being able to continually revise the plan were considered benefits over traditional paper-based safety planning. In some apps, users could create associations between different suicide safety planning components (SPCs; eg, triggers and coping strategies) to better contextualize their experiences and create actional plans for mitigating crises [54,56,61]. We recommend that additional linkages between the SPCs be included to further personalize users’ experiences.

Despite the integration of SPI components within mHealth suicide safety planning app designs, we also identified important gaps in the literature that warrant the attention of app designers, researchers, and mental health professionals who may use this type of technology within their clinical practice. For instance, researchers have consistently emphasized the importance of completing the initial safety plan alongside a knowledgeable clinician [42,54,58,61] to ensure that at-risk users and loved ones understand the components and purpose of a safety plan. However, many of the analyzed apps allowed users to complete the safety plan without the recommended clinical support, and in some cases, they lacked disclaimers. Therefore, additional guidance from a professional when using mHealth suicide safety planning apps would further serve to assist users and ensure that the safety planning process is carried out as intended.

This review also found that most of the apps did not go beyond the traditional SPCs of paper-based protocols to integrate more interactive features that could potentially improve adherence or engagement. For instance, daily or weekly check-ins have been shown to improve adherence in other mHealth contexts, such as for smoking cessation [69] and the management of schizophrenia [70]. Visualization graphs of patterns or trends in suicide warning signs, triggers, and coping behaviors logged over time may serve to increase engagement and improve outcomes, as visualizing behavior change over time has been recommended in other mHealth contexts [71], such as alcohol reduction [72]. Furthermore, other meaningful ways to actively and continuously engage one’s support contacts (eg, clinicians, parents, and family members) and to reinforce the use of helpful coping strategies would be an advantageous direction for future exploration in mHealth app design. Beyond general support contacts, prior research has found that parental support is a significant protective factor against youth suicide [73,74]. For youth, in particular, it may be advantageous to include parents, family members, or other trusted adults in the mHealth suicide safety planning process to increase uptake, enhance help-seeking and coping behaviors, and reinforce ways to keep one’s environment safe. However, future research would need to carefully design and evaluate such interventions to ensure they are effective before making these interventions widely available through the dissemination of mHealth apps for suicide safety planning.

Another variation across the apps was that some apps provided default values for suicide SPCs (eg, suggested coping strategies), whereas others did not. Therefore, an area of future research could be to study whether providing default values is beneficial or detrimental to the safety planning process. Finally, rather than training focused on the technical aspects of using the mHealth app, there is a need to include psychoeducation for suicide safety planning [75], especially related to coping strategies and lethal means restriction, which should be modeled as a collaborative process between at-risk users and their support systems [76].

Usability and Design Considerations for mHealth Suicide Safety Planning Apps

Overall, our review highlights three important recommendations to consider when designing safety planning mHealth apps (RQ2): the need to (1) encourage end user collaboration in the design and implementation of the intervention, (2) incorporate personalization or customization capabilities, and (3) develop...
appropriate privacy safeguards to prevent liability and address other safety concerns that may arise when integrating mental health care and technology. A key strength of most studies in our review was the interdisciplinary collaboration between app developers, computer scientists, and clinical researchers that facilitated the design, development, and evaluation of the various mHealth suicide safety planning apps. In addition, multiple stakeholders were included in the design process, including individuals at risk of suicide, clinicians, usability experts, parents, and extended family members. Only in 1 instance, end users engaged who were not considered part of the target population of at-risk users (eg, students). We strongly recommend that future research continue to include researchers from across multiple disciplines (eg, psychology, public health, social work, medicine, computer science, and human-computer interaction), intended end users, and mental health professionals across each stage of the research process. For instance, researchers from different disciplines may be able to raise important threats to validity during the research design process that could lead to more robust study designs.

A key weakness highlighted within several studies was limited uptake or sustained use of the mHealth suicide safety planning apps over time. Such findings shed suspicion on the feasibility of this type of intervention being effective outside of research, regardless of the high usability and acceptability ratings. Some studies attributed lack of use to the reduction of suicidal behaviors over time, but others suggested that the suicide safety planning process, as designed to be carried out within the apps, was only suited for in-crisis situations and not appropriate for sustained use over time. Although this may be the case, it is also possible that the lack of interactive or engaging features within the apps made them less appealing to users. Being able to customize and personalize app features may help to enhance the user’s experience and increase app engagement. Many of the apps included social distractions (ie, music and pictures) or other features, such as diary cards, which might help increase overall app engagement during noncrisis periods. However, as suicidality is episodic, future research should be conducted to understand how different modalities or features (eg, mood tracking, journaling, mindfulness, and art) could be combined with suicide safety planning in a complementary way for long-term use and engagement. Future work should also consider leveraging advanced technologies and assessments, such as artificial intelligence and ecological momentary assessments [77,78], that could be used to anticipate heightened suicide risk and prompt users to engage in the mHealth app suicide safety planning process when they need it most.

Threats to Validity and Inconclusive Clinical Outcomes Associated With the Use of mHealth Suicide Safety Planning Apps

This review provides some preliminary evidence suggesting that suicide safety planning via mHealth apps could be an easy-to-use mechanism to provide individualized care to those who may otherwise go unserved due to common treatment barriers (RQ3), such as poor accessibility to service providers, lack of knowledge about suicide, and stigmatizing beliefs about help seeking [20-24]. At the same time, several threats to validity were uncovered by our assessment of risk bias, which can inform directions for future research. First, the robustness of the qualitative studies could be improved by stating the positionality of the researchers as well as a clear justification for the design of the mHealth apps. In some cases, articles were published by interdisciplinary teams, whereas in other cases, authors appeared to be from a single discipline (eg, computer science). Details about the composition and expertise of the research teams are important, as well-implemented mHealth apps require interdisciplinary skill sets that span clinical, design-based, and technical expertise. Furthermore, the quantitative studies analyzed in our review were constrained by small sample sizes and no published RCTs. Among the pre-posttest studies conducted thus far, the clinical outcomes were inconclusive.

As such, RCTs with control groups, random assignment, and repeated measure outcomes assessed over time are needed in the future to evaluate the efficacy of using suicide safety planning mHealth apps compared with traditional paper-based safety plans [54,57], specifically related to reducing suicidal urges and behaviors and increasing use of coping strategies, as well as increased engagement in crisis and mental health services after the crisis. When doing so, researchers should recruit larger samples to ensure that the results are conclusive and can be generalized to the populations of interest. Furthermore, additional use metrics collected by the apps to track behavioral data associated with using different app features, such as user engagement with the 6 components of the SPI developed by Stanley and Brown [31], should be considered to better understand the potential mediating factors and behaviors that may influence clinical outcomes. Although the usability of the apps would be an important consideration to control for in future studies, it is necessary to move beyond such measures to determine the efficacy of mHealth apps in reducing suicide-related outcomes. In summary, the inclusion of more advanced study design methodologies and recommendations from lessons learned in future mHealth apps could serve to mitigate suicide risk and promote overall safety.

Limitations and Future Research

This systematic review included 14 peer-reviewed articles that designed, developed, and evaluated mHealth apps for suicide safety planning. There are several limitations of this study that should be addressed in future research. First, although our search process was comprehensive, it is possible that our keywords missed relevant articles and mHealth apps that should have been included in the review. Second, as many of the apps described in the articles were not publicly available for download, we requested access from the corresponding authors to conduct our review. In 2 cases, we were unable to gain access to the apps; therefore, our analysis was based on the description of those apps based on the published paper. As such, it may be possible that some features were not described in the original papers; thus, they were not included in our review. Future research should also consider conducting a systematic feature analysis of mHealth suicide safety planning apps that are publicly available for download but not studied within the peer-reviewed literature. Finally, a limited number of published RCTs at the time of the review restricted our ability to report on app use and...
suicide-related outcomes. As such, the main call-to-action from this review is the need to move beyond usability studies of newly developed mHealth suicide safety planning apps to robust clinical research designs to examine their efficacy in reducing suicidality among at-risk user populations.

Conclusions
Overall, most articles included in this review did little to evaluate the efficacy of mHealth suicide safety planning apps beyond usability assessments, signaling that these apps and corresponding research are still in their infancy in terms of validating clinical outcomes. Although most of the mHealth safety planning apps included in our review are not yet downloadable and broadly available for public use, the prevalence and popularity of mHealth suicide prevention and mental health support apps on the open market that have been deployed without rigorous peer-reviewed research is a concern. As such, there is a critical need for future research to ensure that mHealth apps for suicide safety planning integrate the lessons learned from empirical user-based and clinical research, are upheld to high ethical mental health care standards, and show clinical efficacy for reducing suicidality before the apps are released to end users. This is especially true given the delicate and important goal of preventing suicide among at-risk populations. It is promising to see that future randomized clinical trials have been registered to build upon this important preliminary work on mHealth suicide safety planning apps.

Acknowledgments
The authors would like to thank the following authors who provided access to their apps to conduct this feature analysis: Niels Buus and JL Stovgaard Larsen (MYPLAN); Glenn Melvin (BeyondNow); James Duggan (SafePlan); and Lea Meier, Caroline Gurtner, and François von Kaenel (SERO).

Conflicts of Interest
None declared.

Multimedia Appendix 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[PDF File (Adobe PDF File), 182 KB - mental_v11i1e52763_app1.pdf ]

Multimedia Appendix 2
Example search strategy.
[DOCX File , 18 KB - mental_v11i1e52763_app2.docx ]

Multimedia Appendix 3
Detailed summary of the selected articles and key findings (N=14).
[DOCX File , 40 KB - mental_v11i1e52763_app3.docx ]

Multimedia Appendix 4
Critical appraisal results.
[DOCX File , 20 KB - mental_v11i1e52763_app4.docx ]

References


Abbreviations

JBI: Joanna Briggs Institute
mHealth: mobile health
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RCT: randomized controlled trial
RQ: research question
SPC: safety planning component
SPI: safety planning intervention

©Kim Gryglewicz, Victoria L Orr, Marissa J McNeil, Lindsay A Taliaferro, Serenea Hines, Taylor L Duffy, Pamela J Wisniewski. Originally published in JMIR Mental Health (https://mental.jmir.org), 28.03.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Mental Health, is properly cited. The complete bibliographic information, a link to the original publication on https://mental.jmir.org/, as well as this copyright and license information must be included.
The Urgent Need for an Evidence-Based Digital Mental Health Practice Model of Care for Youth

Brad Ridout¹, MAPS, PhD; Rowena Forsyth¹, PhD; Krestina L Amon¹, PhD; Pablo Navarro², MHLthSc, MAppPsych, M ClinPsych, PhD; Andrew J Campbell³, MAPS, PhD

Abstract

Australian providers of mental health services and support for young people include private and public allied health providers, government initiatives (eg, headspace), nongovernment organizations (eg, Kids Helpline), general practitioners (GPs), and the hospital system. Over 20 years of research has established that many young people prefer to seek mental health support online; however, clear client pathways within and between online and offline mental health services are currently lacking. The authors propose a Digital Mental Health Practice model of care for youth to assist with digital mental health service mapping. The proposed model offers accessible pathways for a client to engage with digital mental health services, provides clear navigation to access support for individual needs, and facilitates a seamless connection with offline mental health services using a transferable electronic health records system. This future-looking model also includes emerging technologies, such as artificial intelligence and the metaverse, which must be accounted for as potential tools to be leveraged for digital therapies and support systems. The urgent need for a user-centered Digital Mental Health Practice model of care for youth in Australia is discussed, highlighting the shortcomings of traditional and existing online triage models evident during the COVID-19 pandemic, and the complex challenges that must be overcome, such as the integration of diverse mental health care providers and establishment of a robust electronic health records system. Potential benefits of such a model include reduced pressure on emergency rooms, improved identification of immediate needs, enhanced referral practices, and the establishment of a cost-efficient national digital mental health care model with global applicability. The authors conclude by stressing the consequences of inaction, warning that delays may lead to more complex challenges as new technologies emerge and exacerbate the long-term negative consequences of poor mental health management on the economic and biopsychosocial well-being of young Australians.

(JMIR Ment Health 2024;11:e48441) doi:10.2196/48441

KEYWORDS

mental health; internet; digital health; telecounselling; social networking; telehealth; telemedicine; counseling; counselling; digital health; service; services; healthcare delivery; youth; model

Introduction

The COVID-19 pandemic highlighted what was already well-established about Australia’s mental health system: that demand and ease of access were not being successfully met [1-5], especially for individuals younger than 25 years, who are the most at-risk age group [6,7]. Over the past 20 years, the Australian youth mental health system has slowly transitioned from being fragmented and piecemeal, with poor integration between inpatient and outpatient services, to include online and offline offerings nationally [8]. However, these offerings are not always clearly connected for clients to navigate [1,7] and are often difficult to access in regional and remote areas (especially for young indigenous and marginalized populations of Australia), where online and offline services are seldom integrated or inclusive [9].

Young people are increasingly using the web as the first step in seeking support and information regarding their mental health [10,11]; yet, the digital services currently on offer often poorly integrate with existing offline mental health and hospital systems of care [1,12,13]. There is also a lack of sharing of clients’ mental health records between private and nongovernment organization (NGO) providers who offer online support (eg, ReachOut, BeyondBlue, Kids Helpline, and SANE), telephone centric services (eg, Lifeline and Kids Helpline), and traditional face-to-face clinical support (eg, headspace). This is due to several ethical and practical challenges associated with sharing electronic health records (EHRs) in mental health care [14], including issues of interoperability, missing data, privacy and confidentiality concerns, and legal compliance.

Currently, EHRs of young people are typically not part of digital mental health service offerings, with data often only collected upon entrance to determine the level of care required (ie, the
triage stage). This means that when young people later return to online services at differing stages of need, or via a different modality, they often must “start again” with programs or offerings that may no longer suit their needs, rather than being directed to more suitable supports that reflect their prior engagement and the current stage of their mental health journey. Furthermore, the current lack of transferable EHRs between service providers implies that young people often have to retell their story multiple times. This can lead to incomplete patient histories, duplicate assessments, fragmented care, and safety risks associated with delayed or inappropriate treatment [15].

We propose a Digital Mental Health Practice model of care for youth (DMHP-Y) to aid digital mental health service mapping. It provides clear and accessible pathways for a young person to enter a digital mental health system, and then accurately navigate it to access support for their individual needs, beyond initial triage. This model proposes transferable EHRs regarding mental health service use to inform clinicians of historical online and offline service access, including any digital mental health assessments and treatments previously engaged in by a client. This will provide clinicians with essential background information on a patient’s mental health journey to date, assist in decision-making about future assessments and treatments, and facilitate continuity of care.

A Brief Overview of Current Digital Mental Health Practices in Australia

Established in 2006, headspace is an NGO funded by the Australian Government as part of its national approach to service the mental health needs of young people (younger than 25 years) [16]. While headspace has been widely lauded for providing no-gap Medicare access to psychologists under a general practitioner (GP)–referral mental health care plan, the service has been under strain, given its high client demand nationwide [17,18] and the limitations to ongoing mental health management beyond the “Better Access” to mental health Medicare system [19-21]. “headspace,” the digital provision of mental health support, was established to provide online resources for young people to connect with mental health information and support through forums, digital tools, and one-to-one counseling via web chat and telephone. While not an emergency service, headspace duplicated some of the resources offered by other NGOs’ online services (eg, ReachOut and Kids Helpline), which have also struggled in connecting online resource offerings to offline clinician services (often because clients may choose to contact these services anonymously). Furthermore, the lack of integration between headspace and face-to-face headspace services demonstrates the challenges involved in integrating online and offline services, even within the same service environment. Herein lies a major systemic issue: the lack of secure, transferable, digital mental health recordkeeping within and between services [22-24].

Notwithstanding this challenge, new online platforms have emerged in recent years that leverage health ITs (HITs) to integrate digital mental health tools with existing face-to-face mental health services. The “Innowell” platform [25] is a customizable digital dashboard for mental health services, designed for assessing, monitoring, and managing the mental health of their clients. It collects and reports personal information to clients and their health professionals to promote collaborative and measurement-based care, and offers a range of online clinical content and assessment tools that are made available to clients as determined by their face-to-face mental health service. Another Australian example is Orygen’s “Moderated Online Social Therapy” (MOST) platform [26,27], a free online digital mental health service that young people can connect to following a referral from participating offline services. The MOST model of care is designed to support young people with, and in between, their face-to-face sessions with a clinician, or while they are waiting for care. It offers a range of on-demand digital services, including one-on-one support, self-directed online therapy, and moderated peer-to-peer online social networking.

These 2 platforms reflect emerging models of digital mental health care internationally, which also use HIT to augment face-to-face sessions and facilitate measurement-based care. A notable example from the United States is the “Digital Clinic” [28,29], which uses smartphone apps to augment and extend care at the Beth Israel Deaconess Medical Center in Boston, Massachusetts. This evolving hybrid model of care emphasizes therapeutic alliance, measurement-based care, and shared decision-making between clients and clinicians. A key feature of the model is mindLAMP (Learn, Assess, Manage, Prevent), an open-source app that provides customizable education, assessment, and management tools, and facilitates data sharing with patients and clinician support. mindLAMP can also be used to capture metadata regarding clients’ use of the app, as well as health and physical activity patterns via a smartwatch (eg, step count and heart rate) to enable “digital phenotyping.” Digital phenotyping is a new and evolving multidisciplinary field of science that uses data from smart devices to create a holistic digital picture of behavior and health, and has the potential to augment clinical care at the individual level (eg, by identifying clinically significant behavior changes) [30].

By integrating HIT with existing mental health services, the models described above represent the beginning of a new era of digital mental health service delivery. Something these models have in common is the requirement for clients to first connect with a face-to-face mental health service to access online services. However, young people are often reluctant to seek professional face-to-face help and face barriers to treatment such as cost, stigma, confidentiality concerns, poor mental health literacy, and inaccessibility to or lack of knowledge of resources [31,32]. These factors contribute to an increasing trend for young people to use the web as the first step in seeking support and information regarding their mental health [10,11], usually by typing a text-based query into an internet search engine [12]. Discussions of Rickwood et al’s [33] help seeking model in the context of the online environment emphasize the importance of internet-based resources and interventions in addressing the mental health concerns of young people [34] and ensuring that young people are guided to appropriate services [35]. Reputable digital mental health services that do not require prior face-to-face referrals, such as ReachOut and Kids Helpline, therefore play an important role in early intervention and as...
gateway services for young people in Australia who are seeking mental health information and support [34,36,37]. They also provide opportunities for young people to anonymously engage with peers online about mental health concerns, using moderated social media platforms, which itself has been shown to increase the likelihood of seeking formal mental health treatment [38,39].

A DMHP-Y is therefore required to map client-led online help seeking and engagement pathways to and within digital mental health services, and help clients navigate access to support for their individual needs, including seamless connections with offline mental health services.

**Considerations in Evidence-Basing a DMHP-Y**

Digital mental health approaches are typically evidence-based as stand-alone interventions (eg, apps, video counseling, websites, and forums) rather than as holistic models of care. For a DMHP-Y to be effective and well-used, more research is needed to build an evidence base for the following, in order to inform design decisions that reflect youths’ help seeking behaviors and preferences, and the navigation of their mental health journey.

**Understand and Monitor Digital Engagement With Mental Health Care Over Time**

This is necessary to map how and at what stages of their mental health journey young people engage with the digital mental health system over time, as use of digital mental health information and tools is often not linear. It also needs to be understood why some demographics might be more likely to arrive at a digital mental health service via an internet search, rather than via clinician referral or social networks. Further, in relation to online search words, search engine optimization (SEO) strategies must account for the different search terms individuals use to find information on mental health, given that the vocabulary for mental health help seeking has been shown to differ based on developmental age [40].

**Understanding Preferred Digital Tools of Young People for Supporting Their Mental Health**

The online support services and modalities offered as part of a DMHP-Y must reflect and keep up with changing preferences and use patterns among young people. For example, there is growing evidence suggesting that young people want to use dynamic social media environments to seek information about their mental health and treatment options, and give and receive peer support from others facing similar mental health challenges [41-43]. While public social media platforms have significant privacy and security risks, recent studies have shown that custom-built social media–based services such as Kids Helpline’s “My Circle” (formerly “KHL Circles”) can be a safe and effective way for young people to engage in counselor-guided peer support for their mental health [25,44-46]. Next on the horizon is the increasing incorporation of virtual, augmented, and mixed reality technologies into social media, giving rise to the metaverse, which presents both opportunities and challenges for the future of digital mental health treatment and support [47,48]. A future-looking DMHP-Y model of care must therefore be dynamic enough to adapt quickly as technology and platform preferences continue to evolve, yet ensure that new service offerings are tested and evidence-based before implementation, and user privacy and safety remain the first priority.

**Ensuring an Optimum and Inclusive User Experience**

It is well established that participatory design research methodologies involving all stakeholders and end users in the design and development of digital mental health systems and interventions are needed to ensure an optimum user experience for clients, clinicians, and other support staff [49,50]. However, further research into understanding individual human factors across different ages, genders, education, locations, abilities, neurotypes, and cultural needs is crucial to providing an inclusive and dynamic experience, leading to better service matching, more engagement, and less dropout from services. Furthermore, all DMHP-Y web content should be created using “universal design” principles to ensure usability by people with the widest possible range of abilities in the widest possible range of situations [51]. For example, Kids Helpline has already adopted a mobile-first approach and has implemented a range of inclusivity features to allow for people with hearing or sight impairment to access their services, such as “voice to text” across all modalities, ensuring that all website text is accessible via screen readers, and that all video content includes subtitles and audio cues. Kids Helpline has also introduced design initiatives to support children and young people with limited access to devices, data, and network coverage.

**Ensuring Continuity of Care**

As discussed, a DMHP-Y model of care must provide for ongoing connections and transferability of information between online modalities and with offline services via EHRs. The need for and benefits of a transferrable EHR should be made clear to clients upon first engagement with a digital mental health system. For example, so that they can easily navigate their mental health journey across modalities (ie, within and between online and offline services) and avoid having to repeat initial onboarding processes in the future. Participatory action research is needed to ensure that such a transferrable EHR system is developed on the basis of user-centered design principles so that it meets the needs of both clinicians and clients, and effectively facilitates continuity of care (especially for nonlinear mental health treatment journeys). Once the evidence base for this is established, a standardized EHR system should ideally be implemented nationally across all community and private mental health services to ensure comprehensive and seamless record keeping between online and offline services.

My Health Records (Australia’s national digital health record system managed by the Australian Digital Health Agency) provides a case study in how difficult it can be to roll out an EHR system nationally. Challenges include low uptake, incomplete information, and concerns around privacy and security [52]. While it is possible for My Health Records to be used for mental health care, it is currently underused for this purpose, with concerns raised about the inclusion of sensitive and potentially stigmatizing mental health information alongside physical health information [24]. Furthermore, psychologists
and other allied health professionals are excluded from engaging meaningfully with My Health Records, as conformant software is not currently available for allied health practices. My Health Records also does not allow for the inclusion of records of engagement with NGOs’ digital mental health services. Therefore, our perspective is that a DMHP-Y model of care requires a national EHR system specifically for mental health, which facilitates the transferring of records between digital mental health services and other health providers. While the considerations are complex and beyond the scope of this paper, any EHR system that seeks to support the mental health journeys of young people needs to include records of engagement with digital mental health services (most of which in Australia, at present, are NGOs and are outside of the health care system).

**Proposing a Future-Looking DMHP-Y**

**Overview**

As outlined, current service provision for youth mental health in Australia offers both offline (eg, headspace and private practice allied health) and online support services (eg, eheadspace, Kids Helpline, and ReachOut). At the time of writing, the only emerging evidence-based digital platforms in Australia connecting clinicians and clients across online and offline service modalities are Innowell [25] and MOST [26,27]. While these models of care provide some basis for the model proposed below in Figure 1, they can only be accessed by clients after obtaining a referral from a participating offline mental health service.

In contrast, the DMHP-Y model of care proposed by the authors is focused on client-led online help seeking and engagement pathways. It aims to provide a map of the client-led engagement pathways that would be possible from the point at which a user first engages with a digital mental health service and is provided with a transferable EHR, regardless of whether they have previously engaged with offline mental health services or how far along their overall mental health journey they might be. While the model proposes that initial access to digital services (with the exception of phone counseling) should be via a common landing point for onboarding, it accounts for there being numerous engagement and re-engagement points (such as following engagement with offline services). The model also accounts for technologies that will change the way we interact online over the coming years, such as the incorporation of natural language processing and other artificial intelligence (AI) modalities, and avatar and social network use in the metaverse. The model presented is, therefore, a future-looking starting point for a model of care that would improve connections between services and transferability of client information, as technological advances and user preferences continue to evolve.

The proposed DMHP-Y model aims for what is herein dubbed the 3 Is of digital mental health care: inclusivity (equal access to all), individuality (choice based on preferences), and integrity (reliable, ethical, and evidence-based care). The 3 Is are addressed, where applicable, across each step of a client’s movement through the system of care.
Step 1: First Engagement

This step represents the individual journeys and modalities that lead young people to first engage with a digital mental health service. As stated previously, young people often use the web as their first step in seeking support and information regarding their mental health [10,11], usually via a text-based query in an internet search engine, which is the most common help seeking approach among young people aged <25 years [12]. Other potential initial engagement channels include suggested links on websites or social media, recommendations from text-based chat services and social media, as well as clinician referral and word of mouth.

Research on young people’s mental health–related search word practices suggests that digital mental health services need to ensure that they can be easily found not only using diagnostic search terms (eg, depression, anxiety, etc), but also using simple affective phrases (eg, sad, scared, or lonely) [40]. SEO strategies also need to account for differences in demographics, ability, neurotype, language, and cultural backgrounds to ensure inclusivity. Moreover, ongoing analysis of common search phrases for mental health issues to inform SEO updates is key to a DMHP-Y model remaining a primary online source of help (integrity) over less evidence-based or predatory mental health offerings.
Step 2: Landing Page

No matter what leads a potential client to first engage with a DMHP-Y system, they should be directed to a common landing page that quickly explains the service modality options available to them (individuality), as well as a clearly identifiable link to “immediate help” (ie, a link to a national mental health crisis line [53]) to reduce cognitive load on what could be a distressed individual (integrity). Often, landing pages are cluttered with dense information or vague navigation. Studies clearly show that time spent engaging with a digital platform is directly related to intuitive design and consideration of user digital literacy [54,55]. The design of the landing page is therefore crucial and should not be dictated solely by the service provider but rather be designed in consultation with young people from a diverse range of backgrounds and with differing levels of mental health literacy (inclusivity). This is to ensure that the right balance is achieved between the need to provide essential information to the client and the need of the client to progress to an appropriate service quickly and easily. This essential information should include explaining the need for and benefit of creating a transferable EHR to navigate their mental health journey, as the next step in the model may require this for progress into the service.

Step 3: Onboarding and Virtual Triage

The next step is to create a transferable EHR for the client (if they do not already have one) so they can onboard into the DMHP-Y system and complete a mental health screening assessment to inform “virtual triage.” It has been proposed that virtual triage could also leverage AI, specifically natural language processing, which shows promise for adaptation for mental health services [55-58] and is currently being used successfully in corporate operations and retail services. It is not the authors’ suggestion that such a sensitive system and service model be reliant on AI alone, rather, a hybrid model of triage engagement be offered (ie, human and AI), depending on the level of severity of a client upon initial screening. An example of this would be when a client has completed the initial onboarding for an EHR and their mental health screen indicates high levels of distress (addressing individual urgent need), they are flagged as such and directed immediately to a human counselor contact (eg, text-based or voice-based, depending on the preference of the client) as the first port of call for immediate triage (ie, integrity of priority care). For clients who decline contact with a human counselor, the system may provide them with targeted resources and alternate options for support (inclusivity), while flagging their profile within the ecosystem for therapeutic monitoring, follow-up, and support. Clients identified as “severe” at screening can, if required, be directed to emergency services, or referred to community or private mental health services via a contact number or URL for their closest service.

Step 4: Online Support

Once level of care requirements are established, the DMHP-Y model can offer a personalized (ie, individual and inclusive) menu of text- and voice-based choices from a suite of digital therapies and supports. Low-risk first-time clients may be directed to multimedia psychoeducation, or virtual counselor sessions can be scheduled in an environment of choice (eg, an initial telecounseling session via phone, web, or metaverse avatar counseling). This initial entrance level of support could also include evidence-based AI chatbots for mental health information [56,58] or asynchronous text-based guidance from counselors [59]. In turn, this would aid them to learn more about their mental health needs and options for individual care, both within the online service and via offline services. With appropriate technology and client consent in place, there is also potential for digital phenotyping to play a role in informing clinical support needs, by integrating data from a client’s smart devices with their engagement trends with the DMHP-Y model over time [30].

For clients who understand their mental health needs (ie, have previously received diagnosis or therapies), the option of self-guided psychoeducation and digital programs, such as cognitive behavioral therapy or acceptance and commitment therapy, may be offered. These forms of digital therapies are evidence-based for clients experiencing depression, anxiety, and stress symptoms and wanting to self-manage [60,61].

Clients who are seeking continuous social support, especially for chronic mental health conditions, may benefit from community engagement via counselor-moderated 24/7 peer-to-peer social networks. This level of ongoing social support has been found to benefit young people in managing their mood and learning new coping strategies, and offers social support outside of specific therapies [26,44-46]. Connecting for peer-to-peer support in counselor-moderated social networks such as Kids Helpline’s My Circle has also been shown to reduce mental health stigma and increase help seeking for mental health issues in the future [44-46]. Kids Helpline’s My Circle is an example of how peer-to-peer social networks can integrate and link directly to other text-based resources such as self-directed online psychoeducation modules, website content, and text-based one-to-one counseling in an easy-to-navigate digital ecosystem. The integration of digital mental health resources and interventions into an ecosystem further allows them to be used strategically and dynamically by digital counselors, by conceptualizing them on dichotomies from low to high intensity (eg, website vs one-to-one text-based counseling) or solo to assisted interventions. For example, clients presenting with “severe” symptomatology who are struggling to see improvements by using self-directed online psychoeducation may conceptually benefit from using higher-intensity and assisted interventions, which may involve online counselors directing them to peer-to-peer support or one-to-one interventions that match their willingness, motivation, and clinical needs.

Offline Support

Though not a “step” in the DMHP-Y model, referral to offline face-to-face support may be deemed appropriate or necessary at any stage of a client’s engagement with a digital mental health service. In this situation, digital counselors should provide “warm referrals” for clients; that is, they should discuss options and gain client consent to introduce them to an offline service. A system for sharing of EHRs must therefore be established between a DMHP-Y model and offline services for seamless
recordkeeping and best-practice care for clients in need of outpatient or inpatient care. This, of course, goes both ways, as clients may transition to digital services having already engaged with offline services. An EHR system for mental health should therefore ideally be implemented nationally across all community, private mental health, and digital mental health services and allow clients who may have an existing EHR established with offline services to connect it to digital mental health services (as many seek digital services themselves rather than via a referral from an offline service).

At the time of writing, there are multiple health record software programs on offer in Australia, but none provide seamless recordkeeping between online and offline services. There are, however, digital mental health services in Australia that seek to connect young people to and from offline services via direct referrals, albeit without fully transferable EHR systems. For example, Kids Helpline connects young people to offline services via warm referrals to community and specialist mental health services and emergency and child protection services where clients are at risk of significant harm. Conversely, as discussed, following a referral from participating offline services, young people can connect to Orygen’s MOST platform [26,27] to access online support either while they are waiting for face-to-face care in order to augment their face-to-face sessions or after discharge. It is these types of systems of care that need to be further developed and tested at the national level.

**Conclusion**

Australia presently overrelies on traditional triage models of care for mental health (eg, hospital emergency departments and crisis support telephone services), which came close to systemic failure during the peak of the COVID-19 pandemic as they struggled to address increased demand for mental health support during the health crisis [62]. Given that the demand for online services is growing despite many digital mental health services currently having a poor evidence base [63], it is now even more necessary to develop a road map, testing phase, and rollout plan for a DMHP-Y model of care for service provision.

User-preferred models for mental health that provide EHRs to enable seamless cross-referrals between online and offline allied health and psychiatric care will reduce pressures on existing mental health providers in Australia and facilitate continuity of care for clients. This will be achieved by maintaining detailed records of the young person’s mental health journey—tracking engagement with both clinical treatment and digital support tools—which will allow seamless repeat entrance to both online and offline support when needed. The goal should be for young people to be able to enter services for support without having to retell their story wherever possible.

The primary challenge is the “start up” phase for a DMHP-Y model of care, which must address four significant hurdles: (1) synchronization of existing systems and models of mental health care, online and offline, across different provider organizations (NGOs, government, private sectors, GPs, and hospitals); (2) establishing a transparent, secure, and agile EHR system for mental health; (3) tracking and managing clients that present within digital systems under multiple identities (eg, due to stigma or privacy concerns); and (4) upskilling mental health workers and professionals in digital health and digital therapy skills to extend evidence-based practices across existing (eg, videoconference counseling) and future technologies (eg, metaverse counseling and AI). Developing a dynamic and future-looking DMHP-Y model of care that incorporates the introduction of new automated technologies such as AI does not imply the loss of quality mental health professional care. On the contrary, a DMHP-Y model of care that incorporates AI facilitates the economic and professional release of clinicians to dedicate time and expertise to clients who are in urgent need of help or require long-term care across hybrid offerings of traditional community face-to-face resources and digital resources.

The investment and development in DMHP-Y models of care will improve services in Australia from “treatment only” to “treatment and user-choice support” systems. In turn, this will (1) alleviate pressure on emergency rooms (which have seen a dramatic increase in mental health presentations over the past decade) [64]; (2) improve online to offline referral practices (and vice versa); (3) better identify and prioritize those who need immediate face-to-face help from those who need ongoing social and psychoeducation support; (4) greatly improve the cost efficiency of the current mental health system in Australia; and (5) pioneer a national model of digital mental health care that is replicable globally.

The most significant risk to our mental health system is inaction. If we do not start implementing changes in digital systems in mental health care at the national level, the knock-on effects of poor mental health care management for young people will continue into adulthood. This will contribute to significant long-term economic and biopsychosocial problems for Australia’s population in the future, compounding national health burdens of disability and disease. Moreover, the longer the delay to develop, test, and refine such digital models for mental health services, the more complex the task will become as new technologies and platforms such as AI and the metaverse are introduced at an increasingly rapid pace.

**Conflicts of Interest**

None declared.

**References**

1. Duggan M, Harris B, Chislett WK, Calder R. Nowhere Else to Go: Why Australia’s Health System Results in People With Mental Illness Getting ‘Stuck’ in Emergency Departments: Mitchell Institute, Victoria University; 2020.

https://mental.jmir.org/2024/1/e48441

JMIR Ment Health 2024 | vol. 11 | e48441 | p.144

(page number not for citation purposes)


Abbreviations

AI: artificial intelligence
DMHP-Y: Digital Mental Health Practice model of care for youth
EHR: electronic health record
GP: general practitioner
HIT: health IT
mindLAMP: Learn, Assess, Manage, Prevent
MOST: Moderated Online Social Therapy
NGO: nongovernment organization
SEO: search engine optimization

Edited by J Torous; submitted 24.04.23; peer-reviewed by D Rickwood, J Matar; revised version received 07.01.24; accepted 25.01.24; published 22.03.24.

Please cite as:
The Urgent Need for an Evidence-Based Digital Mental Health Practice Model of Care for Youth
JMIR Ment Health 2024;11:e48441
URL: https://mental.jmir.org/2024/1/e48441
doi: 10.2196/48441
The Implementation of Recommender Systems for Mental Health Recovery Narratives: Evaluation of Use and Performance

Emily Slade¹, BA; Stefan Rennick-Egglestone¹, PhD; Fiona Ng¹, PhD; Yasuhiro Kotera¹,³, PhD; Joy Llewellyn-Beardsley¹, PhD; Chris Newby¹, PhD; Tony Glover⁵, PhD; Jeroen Keppens⁶, PhD; Mike Slade¹,⁷, PhD

¹School of Health Sciences, Institute of Mental Health, University of Nottingham, Nottingham, United Kingdom
²Department of Computer Science, University of Oxford, Oxford, United Kingdom
³Center for Infectious Disease Education and Research, Osaka University, Osaka, Japan
⁴School of Medicine, University of Nottingham, Nottingham, United Kingdom
⁵DRT Software Ltd., Nottingham, United Kingdom
⁶Department of Informatics, King’s College London, London, United Kingdom
⁷Nord University, Faculty of Nursing and Health Sciences, Namsos, Norway

Corresponding Author:
Emily Slade, BA
School of Health Sciences
Institute of Mental Health
University of Nottingham
University of Nottingham Innovation Park
Triumph Road
Nottingham, NG7 2TU
United Kingdom
Phone: 44 0115 823 0812
Email: emily_slade@outlook.com

Abstract

Background: Recommender systems help narrow down a large range of items to a smaller, personalized set. NarraGive is a first-in-field hybrid recommender system for mental health recovery narratives, recommending narratives based on their content and narrator characteristics (using content-based filtering) and on narratives beneficially impacting other similar users (using collaborative filtering). NarraGive is integrated into the Narrative Experiences Online (NEON) intervention, a web application providing access to the NEON Collection of recovery narratives.

Objective: This study aims to analyze the 3 recommender system algorithms used in NarraGive to inform future interventions using recommender systems for lived experience narratives.

Methods: Using a recently published framework for evaluating recommender systems to structure the analysis, we compared the content-based filtering algorithm and collaborative filtering algorithms by evaluating the accuracy (how close the predicted ratings are to the true ratings), precision (the proportion of the recommended narratives that are relevant), diversity (how diverse the recommended narratives are), coverage (the proportion of all available narratives that can be recommended), and unfairness (whether the algorithms produce less accurate predictions for disadvantaged participants) across gender and ethnicity. We used data from all participants in 2 parallel-group, waitlist control clinical trials of the NEON intervention (NEON trial: N=739; NEON for other [eg, nonpsychosis] mental health problems [NEON-O] trial: N=1023). Both trials included people with self-reported mental health problems who had and had not used statutory mental health services. In addition, NEON trial participants had experienced self-reported psychosis in the previous 5 years. Our evaluation used a database of Likert-scale narrative ratings provided by trial participants in response to validated narrative feedback questions.

Results: Participants from the NEON and NEON-O trials provided 2288 and 1896 narrative ratings, respectively. Each rated narrative had a median of 3 ratings and 2 ratings, respectively. For the NEON trial, the content-based filtering algorithm performed better for coverage; the collaborative filtering algorithms performed better for accuracy, diversity, and unfairness across both gender and ethnicity; and neither algorithm performed better for precision. For the NEON-O trial, the content-based filtering algorithm did not perform better on any metric; the collaborative filtering algorithms performed better on accuracy and unfairness across both gender and ethnicity; and neither algorithm performed better for precision, diversity, or coverage.
**Conclusions:** Clinical population may be associated with recommender system performance. Recommender systems are susceptible to a wide range of undesirable biases. Approaches to mitigating these include providing enough initial data for the recommender system (to prevent overfitting), ensuring that items can be accessed outside the recommender system (to prevent a feedback loop between accessed items and recommended items), and encouraging participants to provide feedback on every narrative they interact with (to prevent participants from only providing feedback when they have strong opinions).

(JMIR Ment Health 2024;11:e45754) doi:10.2196/45754

**KEYWORDS**

recommender system; mean absolute error; precision; intralist diversity; item space coverage; fairness across users; psychosis; Narrative Experiences Online trial; NEON trial; lived experience narrative; recovery story

**Introduction**

**Background**

Recommender systems create personalized recommendations within a specific domain, suggesting items that may be of use to a user and helping quickly narrow down a potentially overwhelming number of options [1]. Recommender systems are used on global platforms such as Netflix—a movie streaming service—which uses other people’s movie ratings to recommend movies, Amazon—an e-commerce company—which uses frequently-bought-together items to recommend purchases, and Pandora—a music streaming service—which uses 450 musical attributes to recommend songs [2].

A range of health care applications for recommender systems have been examined, including the use of recommender systems to suggest prompts for counselors in a suicide prevention helpline chat [3], tailor care preference assessments in nursing homes [4], and identify expert physicians for specific diseases [5].

In this paper, we present an evaluation of NarraGive, the first recommender system for providing web-based recommendations from a collection of mental health recovery narratives.

**Lived Experience Narratives**

Mental health recovery narratives are a subset of lived experience narratives, which are representations of a person’s experiences of physical or mental health and how that person has lived through and responded to those experiences [6]. The uses of lived experience narratives in health care have been extensively studied but with little focus on which narratives people engage with.

Studies have explored the use of lived experiences to encourage people to seek and sustain treatment, such as using narratives to improve health care participation in patients with breast cancer [7], promote smoking cessation in the African American community [8], and promote diabetes self-management [9] and diabetes medication adherence [10]. The use of lived experiences in support groups has also been studied, such as sharing stories in diabetes education in minority ethnic groups [11]. Some studies have provided medical students with narratives to facilitate learning and improve subsequent medical practices, such as using patient stories during practice placements [12] and learning about cancer pathology using narratives of patients who have experienced cancer [13].

Other studies have explored the use of lived experiences as a therapeutic tool for individuals, such as student nurses creating digital stories to challenge the “reality shock” of beginning clinical practice [14], young women telling their stories to reduce stress [15], women with eating disorders accessing recovery stories [16], service users with psychosis watching lived experience videos [17], incarcerated women telling their stories [18], patients with dementia using storytelling as a therapeutic tool [19], adults with diabetes engaging in lived experience support groups to reduce diabetes-related distress [20], painting trees to symbolize periods of one’s life as a starting point for telling a life story to treat depression and anxiety [21], and young people watching digital stories to reduce the prevalence of binge drinking [22].

Lived experience narratives have the potential to be used for a wide variety of purposes and, as a result—as documented previously—are frequently used in interventions. However, so far, the focus of health lived experience–based interventions has been solely on examining the effects of engaging with these narratives, with less focus on which specific narratives the participants are exposed to (though a few studies have placed emphasis on providing representative narratives [23] or particularly engaging and high-quality narratives [8]). Thus, while there have been studies evaluating the use of recommender systems in health care settings and, separately, evaluating the use of lived experience narratives, there have not been any lived experience narrative recommender systems developed before this study.

**The Problem Being Addressed**

This is the first evaluation of a lived experience narrative recommender system. The design of such a recommender system has distinct challenges. For example, narratives are sensitive types of data that impose ethical requirements to protect both the narrator and the recipient. Therefore, the use of recommender systems needs to be informed by considerations about the curation and use of narratives [24-26]. The goal of our evaluation was to develop preliminary evidence to inform the future use and evaluation of recommender systems with lived experience narratives.

**The Narrative Experiences Online Intervention**

**Overview**

The Narrative Experiences Online (NEON) study [27,28] evaluated whether having web-based access to people’s real-life stories of recovery from mental illness can be helpful for people...
who are experiencing psychosis or other mental health problems. This builds on the evidence base that indicates that receiving recovery narratives can support mental health [27]. In the NEON intervention, participants interact with a web application through which they can access a web-based collection of mental health recovery narratives (henceforth, narratives)—the NEON Collection.

**Narrative Characterization**

The development of the NEON Collection, including the narrative inclusion criteria, has been reported elsewhere [29]. In brief, recorded recovery narratives were obtained, always with consent, from existing collections and individual donations to the study. Only narratives that could be presented on the web in a single electronic file (e.g., PDF, JPEG, and WAV) were included. Within these files, narratives were presented in a range of forms, including prose, poetry, audio recordings, video recordings, individual images, and sequential art. Each was presented by a single narrator only—there were no composite narratives. The narratives were deliberately chosen to be diverse [30]. All narratives in the NEON Collection were characterized using the Inventory of Characteristics of Recovery Stories (INCRESE) [31] to capture 77 different features of the narratives related to narrator characteristics, narrative content, and turning points. While we used selected INCRESE characteristics in our recommender system, the greater breadth of characteristics collected will support future secondary analyses. The trials opened with 348 narratives and closed with 659 narratives available.

**Narrative Request Routes**

There are 6 ways for participants to request narratives through the NEON intervention, which are internally documented as 1 of 8 request methods.

Textbox 1 summarizes the external and internal narrative request routes.

The NEON intervention home page has buttons corresponding to 4 of the 6 external narrative request routes: “Match me to a story (recommended),” “Get me a random story,” “Browse stories,” and “My stories.”

The first option uses NarraGive to recommend a single narrative that the participant has not seen before. NarraGive is a hybrid recommender system (meaning that it uses a combination of recommendation strategies [32]) that uses both content-based filtering (recommending narratives based on their content) and collaborative filtering (recommending narratives based on how other participants have rated them) to recommend narratives to participants.

The second option presents a randomly selected narrative that the participant has not seen before.

The third option allows participants to browse narratives grouped into categories based on the narratives’ INCRESE characteristics (Figures S1 and S2 in Multimedia Appendix 1)—some categories are based on the value of a single characteristic (e.g., the narrator’s gender is “female”), and some are based on the value of multiple characteristics (e.g., a positive narrative, defined as having an “upbeat” tone and an “escape” or “enlightenment” genre; Table S1 in Multimedia Appendix 1). Not all narratives are accessible through the category option.

The fourth option allows participants to access narratives that they have previously bookmarked or rated highly.

In addition, the internal request routes include whether NarraGive produced the recommendation using content-based filtering or collaborative filtering and whether a narrative selected from the “My stories” page was previously rated highly for hopefulness or manually bookmarked by the participant. One important benefit of having different narrative request routes is to prevent exposure bias, a well-known issue in recommender systems where participants are only presented with a subset of the available items, so they only provide ratings for that subset, with recommender systems unable to distinguish between disliked and unrated items and unknown and unrated items [33]. For example, the “Get me a random story” button might allow participants to access narratives that they would not otherwise be exposed to but that nonetheless may be beneficial.
Textbox 1. Narrative request mechanisms that participants use to access narratives (external routes) and the corresponding logs made by the intervention (internal routes).

**External and internal narrative request routes**

- Participant clicks on the “Match me to a story (recommended)” button
  - Participant accesses a narrative recommended via content-based filtering.
  - Participant accesses a narrative recommended via collaborative filtering.
- Participant clicks on the “Get me a random story” button
  - Participant requests a random narrative.
- Participant clicks on the “Browse stories” button and selects a narrative
  - Participant makes a category-based request for a narrative.
- Participant clicks on the “My stories” button and selects a narrative
  - Participant requests a narrative that they have rated as hopeful.
  - Participant requests a narrative that they have marked as a favorite.
- Participant uses the intervention for the first time and is presented with their first narrative
  - Participant accesses their “first” narrative.
- Participant clicks on a narrative from a Narrative Experiences Online (NEON) communication
  - Participant accesses the suggested narrative in a reminder message aimed at prompting them to use the NEON intervention.

**Narrative Feedback**

After a participant has accessed a narrative through any request route, they are presented with 5 feedback questions (Table 1), and their responses to these questions are time-stamped and logged. The focus (hope, similarity, learning, and empathy) is based on the NEON Impact Model [29] developed through a systematic review [34] and qualitative [35] and experimental studies [36]. The measurement approach has been previously validated [29]. To maximize response rates, the first question is marked as compulsory. The other 4 questions are marked as optional, and the participant has the choice to answer either all or none of the optional questions. A set of 5 response values (for the 1 compulsory and 4 optional questions) forms a single rating, as does a single response value for the compulsory question. Ratings with optional questions answered are also referred to as optional ratings. Table 1 shows the questions, answer options, and numerical ranges (not visible to participants) of the questions and whether they are mandatory.

If a narrative is rerated, this overrides the previous rating (but the time-stamped logs of previous ratings are not deleted).

One benefit of recommender systems requiring a rating for each narrative is that this helps minimize selection bias, which occurs when participants are allowed to choose whether to rate the items, leading to ratings that are typically biased toward higher or more homogeneous ratings [33,37]. Selection bias is a well-known problem in recommender systems relying on explicit data.

<table>
<thead>
<tr>
<th>Question</th>
<th>Answer options</th>
<th>Range</th>
<th>Mandatory</th>
</tr>
</thead>
<tbody>
<tr>
<td>How hopeful did the story leave you feeling? (hopefulness)</td>
<td>“Less hopeful than before,” “no change,” “a bit more hopeful,” and “much more hopeful”</td>
<td>−1 to 2</td>
<td>Yes</td>
</tr>
<tr>
<td>How similar was the storyteller to you? (similarity to the narrator)</td>
<td>“Not at all,” “a bit,” “quite a lot,” and “very much”</td>
<td>0 to 3</td>
<td>No</td>
</tr>
<tr>
<td>How similar was the storyteller’s life to your life? (similarity to the narrative)</td>
<td>“Not at all,” “a bit,” “quite a lot,” and “very much”</td>
<td>0 to 3</td>
<td>No</td>
</tr>
<tr>
<td>How much did you learn from the story? (learning)</td>
<td>“Not at all,” “a bit,” “quite a lot,” and “very much”</td>
<td>0 to 3</td>
<td>No</td>
</tr>
<tr>
<td>How emotionally connected did you feel with the story? (empathy)</td>
<td>“Not at all,” “a bit,” “quite a lot,” and “very much”</td>
<td>0 to 3</td>
<td>No</td>
</tr>
</tbody>
</table>
The NarraGive Recommender System

NarraGive is a hybrid recommender system. It uses one content-based and 2 collaborative filtering algorithms to allow for comparison of performance of the 3 algorithms using 2 distinct approaches to inform this new field of lived experience narrative recommendation. NarraGive was assembled using the Simple Python Recommendation System Engine library (SurPRISE; version 1.1.1; Nicolas Hug) for Python (version 3.6 and above), integrating implementations of filtering algorithms provided in these libraries [38]. NarraGive does not recommend previously requested narratives, types of narratives that a user has previously blocked, or individual narratives that a user has blocked.

The content-based filtering algorithm is based on the SurPRISE implementation of the k-nearest neighbor (kNN) algorithm. Although kNN is traditionally used as a collaborative filtering algorithm, NarraGive used an adapted version to measure the similarity between narratives, in which it uses their INCRESE characteristics to cluster together narratives in “neighborhoods” and recommend to participants unseen narratives that are similar to their other highly rated narratives. Narrative similarity is assessed using selected INCRESE characteristics, consisting of the INCRESE sections on narrator characteristics, narrative characteristics, narrative content, and turning points.

The selected collaborative filtering algorithms are the SurPRISE implementations of the singular value decomposition (SVD) and, to support comparison, SVD++. A broad introduction to these 2 algorithms is provided in the work by Hug [39]. These aim to capture the latent factors that determine how much a participant likes a narrative. NarraGive ran these 2 algorithms and selected the narrative with the highest predicted rating. Thus, the 2 algorithms served as distinct subsystems, so this evaluation will analyze the 2 subsystems separately to compare them. For the purposes of collaborative filtering, similarity between users is assessed using the demographic items collected in a “personal profile” created at first use and containing items describing participant demographics and format preferences. Multimedia Appendix 2 provides details on all items in the profile.

When making a narrative recommendation, narrative feedback ratings are weighted (with a hopefulness rating twice as influential as each of the individual optional ratings) and combined. This was due to the underlying theory that we developed on narratives making an impact on recipients, which emphasized hope creation as the most critical mechanism. When a participant requests a narrative from NarraGive, it internally generates 1 list per algorithm of the 10 narratives with the highest predicted rating. It then presents the highest-scoring narrative of these 30 to the participant. The participant is not shown the predicted rating, other internally generated narratives, or which of the 2 filtering mechanisms was used to generate the recommendation.

The NEON Trials

The NEON intervention has been evaluated in 3 pragmatic randomized controlled trials with different populations. The NEON trial (ISRCTN11152837; N=739) is a definitive trial for people with experience of psychosis. The NEON for other (eg, nonpsychosis) mental health problems (NEON-O) trial (ISRCTN63197153; N=1023) is a definitive trial for people experiencing any other type of mental health problem. The NEON-C trial (ISRCTN76355273; N=54) is a feasibility trial with people who informally care for people experiencing mental health problems, which is not within the scope of this study. The NEON intervention was identical in all 3 trials. A separate instance of NarraGive was used for each trial, and there was no pooling of narrative feedback or recommendations among the 3 trials.

Aims and Objectives

The aim of this study was to analyze the 3 recommender system algorithms used in NarraGive to inform future interventions using recommender systems in this new field of lived experience narrative recommendations. An evaluation of the impact of the NEON intervention using NarraGive has been reported elsewhere [40]. This study did not aim to provide an indication of NarraGive’s viability but rather to inform the development of future lived experience narrative recommender systems and guide design choices on collaborative versus content-based filtering algorithms.

The objectives were as follows:

1. To describe participant characteristics and patterns of narrative requests and feedback.
2. To evaluate the algorithms used in NarraGive by comparing collaborative-based and content-based narrative recommendations to inform future implementation approaches.

Objective 1 was addressed using data from the intervention version of NarraGive, and objective 2 was addressed using data from the final evaluated version.

Methods

Overview

An evaluation of NarraGive was conducted using data from the NEON and NEON-O trials, structured using the framework for evaluating recommender systems (FEVR), which was developed through a review of recommender system evaluation work [41]. The FEVR defines a set of components intended to guide the design of a recommender system evaluation.

After the NEON trials closed, logging files describing interactions with trial procedures and the NEON intervention were downloaded for analysis. These files included trial allocation, baseline demographic characteristics, personal profiles, narrative characteristics, narratives that the participants requested and the corresponding internal narrative request route, and participants’ ratings. All log entries were time-stamped.

Ethical Considerations

The NEON study trial protocol and an update have been published elsewhere [27,28]. Ethics approval was obtained in advance of trial start from a UK National Health Service Research Ethics Committee (Leicester Central Research Ethics Committee; 19/EM/0326). All participants provided web-based data on their consent to participate in the study.
informed consent for the use of their data for research purposes, and all study data were pseudonymous, with each participant’s data linked by a unique ID. Some participants were compensated (£20 [US $25.59] vouchers) for some data collection rounds, as described in our trial protocol.

Participants

The NEON trial included participants who (1) had experience of psychosis in the previous 5 years, (2) had experience of mental health–related distress in the previous 6 months, (3) resided in England, (4) were aged ≥18 years, (5) were capable of accessing or being supported to access the internet on a PC or mobile device or at a community venue, (6) were able to understand written and spoken English, and (7) were capable of providing web-based informed consent.

The NEON-O trial included participants who (1) had experience of mental health problems other than psychosis in the previous 5 years, (2) had experience of mental health–related distress in the previous 6 months, (3) resided in England, (4) were aged ≥18 years, (5) were capable of accessing or being supported to access the internet on a PC or mobile device or at a community venue, (6) were able to understand written and spoken English, and (7) were capable of providing web-based informed consent. It excluded participants eligible for the NEON trial.

Our study included participants from the NEON trials’ intention-to-treat samples [27].

Sample Size

Both trials were powered on the mean item score for the 12 subjective items in the Manchester Short Assessment of Quality of Life (MANSA) as collected at baseline and the 52-week follow-up [42], and hence, the sample size was chosen on this basis.

For the NEON trial, a total sample size of 684 was chosen to provide 90% power to detect a minimal clinically important effect size (Cohen $d$) of 0.27 (SD 0.9 [43]; power=90%; $P=0.05$), allowing for 20% attrition. The planned analyzable sample size was 546.

For the NEON-O trial, the SD of the MANSA scores for the study population was estimated from baseline data provided by the first 350 enrolled participants (see the study by Rennick-Egglestone et al [27] for the rationale). A total sample size of 994 was selected to provide 90% power to detect a minimal clinically important effect size (Cohen $d$) of 0.27 (SD 0.94; power=90%; $P=0.05$), allowing for 40% attrition, which was estimated from the completion rates for interim data. The planned analyzable sample size was 596.

Both trials recruited their planned samples and were allowed to overrecruit (N=739 for the NEON trial and N=1023 for the NEON-O trial). The final attrition rates were 23.5% (NEON trial) and 44.8% (NEON-O trial).

Evaluation Framework

Table 2 describes the FEVR components that were selected to define the evaluation.
Table 2. Framework for evaluating recommender systems (FEVR) components defining the NarraGive evaluation.

<table>
<thead>
<tr>
<th>FEVR component</th>
<th>Brief description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Evaluation objectives</strong></td>
<td></td>
</tr>
<tr>
<td>Overall goal</td>
<td>To evaluate whether the recommender system NarraGive supported participants in finding helpful narratives</td>
</tr>
<tr>
<td>Stakeholders</td>
<td>Participants in the NEON(^a) and NEON-O(^b) trials' ITT(^c) samples</td>
</tr>
<tr>
<td>Properties</td>
<td>Prediction accuracy, usage prediction, diversity, coverage, and unfairness across participants</td>
</tr>
<tr>
<td><strong>Evaluation principles</strong></td>
<td></td>
</tr>
</tbody>
</table>
| Hypothesis or research question | Objective 1: To describe participant characteristics and patterns of narrative requests and feedback  
Objective 2: To evaluate the NarraGive recommender system by comparing collaborative-based and content-based narrative recommendations |
| Control variables | Randomized data set that is split 75:25 between the training set (to train the algorithms) and the testing set (to evaluate the metrics) |
| Generalization power | Use of real-world data from participants with mental health problems; limited due to variation in system use |
| Reliability | Cross-validation with repeated initialization of collaborative filtering algorithms |
| Experiment type | Offline evaluation |
| **Evaluation aspects** | |
| Types of data | Explicit ratings |
| Data collection | Participant ratings (prompted after every narrative access) |
| Data quality and biases | Platform bias from suggested narratives |
| Evaluation metrics | Normalized mean absolute error (for prediction accuracy)  
Mean average precision per participant (for usage prediction)  
Intralist diversity (for diversity)  
Item space coverage (for coverage)  
Overestimation of unfairness (for unfairness across participants) |
| Evaluation system | NEON intervention web application |

\(^a\)NEON: Narrative Experiences Online.  
\(^b\)NEON-O trial: NEON for other [eg, nonpsychosis] mental health problems) trial.  
\(^c\)ITT: intention to treat.

**Recruitment**
Participants were recruited across England from March 9, 2020 (both trials), to March 1, 2021 (NEON trial), or March 26, 2021 (NEON-O trial). The trials used a mixed web-based and offline approach to recruit participants. Recruitment was through paid web-based advertising on mental health websites; promotional messaging distributed by a range of community groups and health care practices; promotional messaging distributed on Facebook, Twitter (subsequently rebranded as X), and Google (with the reach of messages enhanced through payments); media appearances by the central study team; and the work of clinical research officers in 11 secondary care mental health trusts.

Clinical research officers approached participants in person and distributed promotional messaging through local authorized channels such as mailing lists of service users who had consented to be contacted about research studies. All promotional advertising and messaging conformed to principles approved in advance by the supervising research ethics committee [44].

**Registration**
All recruitment approaches directed potential participants to a web-based eligibility checking interface that requested responses to a series of questions specified in the trial protocol. All responses were self-rated. No formal diagnosis of a mental health condition was required for participation. Trial allocation was determined through responses, and eligible potential participants were provided with access to a tailored web-based participant information sheet. Participants subsequently completed a web-based informed consent form by providing an email address and optional telephone number.
The consent process was concluded by clicking on a link in an auto-generated email to validate the email address. After confirming consent, participants completed web-based forms to collect baseline demographic and clinical data and were then randomized using a web-based system validated by a clinical trial unit to the intervention or control arm. Demographic details were age (in years), gender (female, male, or other), ethnicity, region of residence, highest educational qualification, lifetime use of primary care mental health services, lifetime use of specialist mental health services, current use of mental health services in relation to psychosis (NEON trial only), main mental health problem in the last month, best description of recovery status, residential status, and employment status.

Intervention arm users gained immediate access to the NEON intervention until trial end (September 22, 2022), whereas control arm users gained access after completing the 52-week follow-up questionnaires and until trial end. Data on NEON intervention use by both intervention and control group users are within the scope of this study.

Analysis

**Objective 1: Describe Participant Characteristics and Patterns of Narrative Requests and Feedback**

**Participant Characteristics**

The demographic and clinical characteristics of participants randomized to each trial were described using means and SDs for normally distributed data and counts with percentages for categorical data. Descriptive statistics were calculated for all baseline demographic data items.

Following UK Data Service guidance on statistical disclosure [45], ethnicity responses were grouped into 2 categories (White British and other ethnicity) due to the small number of participants in most ethnicity categories, although recognizing that this could be perceived as a reductive approach to ethnicity data. “Current mental health problem” also comprised categories with low numbers of participants, so relevant rows were shown as “<5” with no percentage, and other rows were shown as “<10” with no percentage to avoid being able to infer other values.

**Patterns of Narrative Requests and Feedback**

Data on participant narrative requests and narrative feedback were taken from log files and used to calculate per-trial summary statistics for the number of participants, number of narratives who requested at least one narrative, number of narratives at the start and end of the trial, number of narratives given at least one rating, number of narrative requests, number of narrative ratings, number of optional ratings, number of ratings per narrative, number of ratings per rated narrative, length of intervention use by participants, and narrative access routes.

While providing feedback on narratives was encouraged, it was possible for the participant to navigate away from the page and not submit any feedback; therefore, the number of narrative ratings may be smaller than the number of narrative requests, so these figures were reported separately.

Statistics for the number of ratings per narrative present 2 sets of figures with different selection criteria: those including only data for narratives that received at least one rating and those including data for all narratives. This breakdown shows how many ratings NarraGive had access to as it could only access rated narratives.

Nonparametric data were presented as medians and IQRs. Category data were presented as counts with percentages.

**Overview**

The 3 algorithms (kNN, SVD, and SVD++) were trained and tested using all the available data, representing the point in time at which the trials closed. Training an algorithm involves providing it with a set of data that it can use to create predictions for missing data points. Testing an algorithm involves obtaining these predictions and measuring a feature of those predictions.

The results for objective 2 were obtained using the SurPRISE library (version 1.1.3) for Python (version 3.10.7). Only participants who provided at least one rating and narratives that were given at least one rating were included (as SurPRISE uses participant-item rating pairs as the basis for its predictions), which mirrors the information that NarraGive had access to during the intervention.

This study evaluated NarraGive using the metrics outlined in Textbox 2, applied separately to the content-based algorithm (kNN) and the collaborative filtering–based algorithms (SVD and SVD++)

There are 2 types of metrics: metrics that compare predicted ratings with actual ratings (prediction-based metrics) and metrics that measure a feature of the top-n predicted items (feature-based metrics). Prediction-based metrics include prediction accuracy, usage prediction, and unfairness across participants. Feature-based metrics include diversity and coverage. For prediction-based metrics, there is no standard data-splitting strategy [46], so the data set is split into a training set (75%) and a testing set (25%). For feature-based metrics, the entire data set is used as the training set.

NarraGive only used the first 3 sets of ratings (hopefulness, similarity to the narrator, and similarity to the narrative) to inform its recommendations as these 3 questions had been validated in a feasibility study [29] and the remaining 2 questions were added after the feasibility study. Therefore, only the first 3 sets of ratings were used in the evaluation.

The hopefulness ratings were normalized, which in this case involved shifting the ratings to use the same rating scale as that of the 4 optional questions.

The evaluated version of NarraGive presented in this paper used the same training data as the intervention version of NarraGive with 3 minor modifications. First, where the narratives’ INCREASE characteristics were updated during the trials (eg, to correct human error in inputting characteristics), this evaluation only used the final set of uploaded characteristics. Second,
during the intervention, NarraGive filtered out previously requested and blocked narratives. This evaluation included these narratives as the predictions themselves were not influenced by whether a narrative was blocked or previously requested (i.e., blocked and previously requested narratives were filtered out after the prediction process in the trial implementation), which could affect, for example, coverage metrics. Third, during the NEON trials, some accounts were removed due to suspected repeat registrations [27]; this evaluation removed all ratings from those participants even though NarraGive may have initially used those ratings.

The logs that were recorded during the intervention did not include NarraGive’s internal recommendation lists and instead only recorded the single narrative that was selected to show to the participants. Therefore, using the intervention version of NarraGive would have prevented any comparison of its subsystems and would have allowed for only a limited analysis of its performance as a whole.

The results from objective 1 (about participants and their use of the system) used the data collected from the live intervention, whereas the results from objective 2 (about NarraGive and its subsystems) used the evaluation version of NarraGive.

During a previous feasibility study of NEON (N=25 mental health service users), 465 ratings were collected for the initial set of narratives in the NEON Collection [29]. NarraGive had access to these ratings in the NEON and NEON-O trials to reduce the “cold start” problem, where recommender systems perform poorly for new items and participants [1]. The evaluation excluded these ratings to ensure that NarraGive was only evaluated on data collected live during the NEON intervention.

The SVD and SVD++ algorithms were both randomly initialized according to a normal distribution [47], and the 75:25 split between training and testing sets was also random and calculated using NumPy (a package for scientific computing with Python) [48], where “fresh, unpredictable entropy will be pulled from the OS” [48]. To account for the randomness, cross-validation was performed. The data set was split into 4 folds, with a different fold used as the testing set each time, and the SVD and SVD++ algorithms were reinitialized each time. Medians and IQRs were reported.

An additional exploratory analysis was conducted to determine how the accuracy changed over time. For each month between June 2020 and July 2022 inclusive, data up to but not including the first day of each month were used for training and testing, and the accuracy was measured (using the same accuracy metric as for the main NarraGive evaluation).

Textbox 2. The 5 metrics for evaluating NarraGive.

<table>
<thead>
<tr>
<th>Metric and metric category</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prediction accuracy</td>
</tr>
<tr>
<td>• Normalized mean absolute error</td>
</tr>
<tr>
<td>Usage prediction</td>
</tr>
<tr>
<td>• Mean average precision per participant</td>
</tr>
<tr>
<td>Diversity</td>
</tr>
<tr>
<td>• Intralist diversity</td>
</tr>
<tr>
<td>Coverage</td>
</tr>
<tr>
<td>• Item space coverage</td>
</tr>
<tr>
<td>Unfairness across participants</td>
</tr>
<tr>
<td>• Overestimation of unfairness</td>
</tr>
</tbody>
</table>

**Prediction Accuracy**
Prediction accuracy is the extent to which a recommender system can predict participant ratings [41]. The root-mean-square error (RMSE) and mean absolute error (MAE) [49] are 2 of the most commonly used metrics for evaluating rating prediction accuracy. The MAE uses the absolute difference between the predicted and true ratings, whereas the RMSE squares this difference, which results in the RMSE penalizing inaccurate predictions more [1].

The intervention was designed to be used over time rather than as a one-off, so the accuracy metric should primarily capture the overall accuracy rather than emphasizing occasional large inaccuracies (i.e., an inaccurate prediction off by 2 points followed by a completely accurate prediction should be treated as no worse than 2 inaccurate predictions off by 1 point each), and this is better achieved using the MAE. Because the hopefulness ratings were normalized, the prediction accuracy metric was the normalized MAE (NMAE).

Different variations in the MAE have been reported in the literature. In particular, some versions square root the averaged summation [1], whereas others do not [47,50]. This evaluation uses SurPRISE’s in-built MAE calculation, which does not use a square root.
A lower NMAE indicates greater prediction accuracy. For NarraGive, the scale ranges from 0 (greatest prediction accuracy) to 4 (equation 1 in Multimedia Appendix 3).

Usage prediction
Usage prediction is the rate of correct recommendations in a setting where recommendations are classified as 1 of 2 options: relevant or nonrelevant [41]. An item is relevant to a participant when the participant’s rating for it meets a predefined numerical threshold (where the threshold is participant independent and defined per question).

There are 2 common metrics for measuring usage prediction: precision and recall. Precision measures how likely it is that a recommended item is relevant and is defined as the ratio of relevant selected items to the total number of selected items [49]. Recall, conversely, measures how likely it is that a relevant item is selected and is defined as the ratio of relevant selected items to the total number of relevant items [49].

As the length of the recommendation list increases, recall improves, whereas precision worsens [1,49]. The length of NarraGive’s internal recommendation list is 10, which is relatively short (compared to, for example, a search engine that recommends tens or hundreds of web pages), meaning that it is impossible to achieve a meaningfully high recall score, so the metric for usage prediction was precision.

As usage prediction is usually used for measuring how relevant a list of recommendations is, this evaluation used NarraGive’s internal recommendation list (consisting of a 10-narrative list produced using content-based filtering and two 10-narrative lists produced using collaborative filtering). As the participants do not see this list, only metrics that focus on the characteristics of the list as a whole—rather than focusing on the order within the list—were used (ie, where the list is treated more like a mathematical set than an ordered list as the ordering beyond the first item does not affect participants), and metrics that exclusively evaluate ranking order were not used.

The analysis of recommender system evaluations by Herlocker et al [49] showed that accuracy metrics can be divided into equivalence classes. One of these classes comprises all metrics that are averaged overall, and one of these classes comprises per-user correlation metrics and the mean average precision per-user metric. To ensure that this analysis of NarraGive captured its performance as widely as possible, a variation of precision that falls into a different equivalence class from that of the NMAE was used, namely, the mean average precision per participant (hereafter, precision).

As the ratings are on a 4-point scale, they need to be converted to a binary scale that classifies recommendations as either relevant or nonrelevant. For optional questions, relevance was defined as “a bit,” “quite a lot,” or “very much.” For helpfulness, relevance was defined as “no change,” “a bit more hopeful,” or “much more hopeful.”

Higher precision indicates a greater proportion of relevant narratives. The scale ranges from 0 (least precision) to 1 (equation 2 in Multimedia Appendix 3).

Diversity
Diversity measures how varied the recommended items are [41]. The current metrics for diversity [41,50] are intralist diversity (ILD) and variations thereof. ILD was developed by Ziegler et al [51], and variations include the rank-sensitive ILD metric by Vargas and Castells [52]. Similar to usage prediction, because the lists used to calculate diversity came from NarraGive’s internal recommendation list and the ILD by Ziegler et al [51] is permutation insensitive (ie, the position of recommendations on the list does not affect the diversity score), this metric was used, with cosine similarity as the distance metric calculated using the narratives’ INCRESE characteristics.

The original study defined ILD on a per-list basis (ie, for the recommendation list of one participant). This metric has been expanded in this study to be averaged over all participants’ lists to produce an overall ILD value.

The lower the ILD value, the greater the diversity among the recommended items. The scale ranges from −1 (most diverse) to 1 (equation 3 in Multimedia Appendix 3).

Coverage
Coverage can be split into participant space coverage and item space coverage [41]. Participant space coverage is the proportion of participants who can be provided with recommendations by the recommender system [1]. The threshold for being provided recommendations is low—a participant needs to have rated at least one narrative (which is achieved when they first access the intervention as it is compulsory to provide a response for the first narrative); thus, participant space coverage was not used. A variation of participant space coverage assesses the proportion of participants that can be recommended high-quality items (ie, items with a predicted rating above a predefined threshold). This notion of variable quality among participants is addressed more thoroughly using an unfairness across participants metric instead.

Item space coverage is the proportion of items that the recommender system can recommend [1]. Ge et al [53] further split item space coverage into prediction coverage and catalog coverage. They defined prediction coverage as the proportion of items for which the recommender system can produce a predicted rating and catalog coverage as the proportion of items that are recommended in a series of recommendation lists. Because there is no predefined limit to when NarraGive can produce a predicted rating for a narrative, prediction coverage was used.

The definition of catalog coverage by Ge et al [53] captures the set of recommended items produced over time for a single participant (ie, the items that would have been recommended to the participant if they had asked for recommendations at that time; this is different from the set of recommended items that the participant requested and was actually presented with over time).

To capture the overall coverage, the proportion of narratives that are recommendable is measured, where a narrative is recommendable if, for at least one participant, the narrative appears in NarraGive’s internal recommendation list.
Other versions of coverage use only the top recommendation, but as there are more narratives than there are participants, this would upper bound the item space coverage at approximately three-quarters for the NEON trial—total number of recommendations (which is equal to the number of participants who rated at least one narrative as there is 1 recommendation per participant) divided by the number of narratives that were rated at least once. For longer recommendation lists (such as 10), because recommender system algorithms cannot always produce a predicted rating for each item, a participant’s list may be less than the desired length. For this evaluation, a length of 10 was sufficient to ensure that the total number of recommendations being considered across all participants was greater than the number of narratives.

A higher item space coverage value indicates greater item coverage. The scale ranges from 0 (lowest item coverage) to 1 (equation 4 in Multimedia Appendix 3).

**Unfairness Across Participants**

Unfairness across participants measures whether participants are treated fairly either at the group level (participants in the same group are treated fairly) or at the individual level (participants who are similar are treated fairly) [41]. NarraGive is designed for use in a health care setting—a setting in which protected characteristics such as disability are critical to attend to. It would be crude to stipulate that, for example, all participants should have an equal probability of being recommended a narrative about wheelchair users as this would be far more relevant to some participants than others (and, indeed, a recommender system’s entire purpose is to provide personalized rather than generic recommendations). As acknowledged by Yao and Huang [54], “in tasks such as recommendation, user preferences are indeed influenced by sensitive features such as gender, race, and age. Therefore, enforcing demographic parity may significantly damage the quality of recommendations.”

Thus, they proposed 4 metrics: value unfairness, absolute unfairness, underestimation of unfairness, and overestimation of unfairness. Value unfairness “occurs when one class of user is consistently given higher or lower predictions than their true preferences.” Absolute unfairness “measures inconsistency in absolute estimation error across user types.” Underestimation of unfairness “measures inconsistency in how much the predictions understate the true ratings.” Overestimation of unfairness “measures inconsistency in how much the predictions overestimate the true ratings.”

NarraGive is implemented in a health care context in which the principle of harm avoidance is crucial. Therefore, one of the most important factors to consider is whether NarraGive is recommending potentially harmful narratives to participants. The metric used to measure this aspect is the overestimation of unfairness.

Overestimation of unfairness measures how much NarraGive consistently overestimates the predicted rating of narratives (ie, how often a participant rates a narrative lower than NarraGive expected) within a disadvantaged subset of the participants and compares this to the overestimation in the nondisadvantaged group.

Participants were divided into groups based on their demographic characteristics. The first grouping was by ethnicity as having a minority ethnicity predicts mental health problems [55], and the second grouping was by gender, informed by Sex and Gender Equity in Research guidelines [56].

The disadvantaged group for the gender comparison was defined as either “Female” or “Other.” The disadvantaged group for the ethnicity comparison was defined as “Irish,” “Gypsy or Irish Traveller,” “Any other White background,” “White and Black Caribbean,” “White and Black African,” “White and Asian,” “Any other Mixed/Multiple ethnic background,” “Indian,” “Pakistan,” “Bangladeshi,” “Chinese,” “Any other Asian background,” “African,” “Caribbean,” “Any other Black/African/Caribbean background,” “Arab,” and “Any other ethnic group.”

The baseline demographic information was used for measuring unfairness between participants as the questions were compulsory, so there was higher completeness of the baseline data than of the personal profile as well as greater granularity with the range of possible answers. The overestimation of unfairness is defined according to the study by Yao and Huang [54].

A lower overestimation of unfairness value indicates that there is less disparity between overestimation among disadvantaged participants and among nondisadvantaged participants. The scale ranges from 0 (least unfair) to 4 (equation 5 in Multimedia Appendix 3).

**Other Categories**

Zangerle and Bauer [41] detailed 10 categories of evaluation metrics that can be used in the FEVR. Of these, 5 (discussed previously) were used in evaluating NarraGive, and the other 5—ranking, novelty, serendipity, fairness across items, and business oriented—were not used for the reasons described in Multimedia Appendix 4 [33,41,57].

**Results**

**Objective 1: Describe Participant Characteristics and Patterns of Narrative Requests and Feedback**

**Participant Characteristics**

The baseline sociodemographic and clinical characteristics of participants in the NEON (N=739) and NEON-O (N=1023) trials are shown in Table 3.

An exploration of the baseline differences has been reported elsewhere [58].
Table 3. Baseline sociodemographic and clinical characteristics of Narrative Experiences Online (NEON) and NEON for other (eg, nonpsychosis) mental health problems trial (NEON-O) participants.

<table>
<thead>
<tr>
<th></th>
<th>NEON baseline (N=739)</th>
<th>NEON-O baseline (N=1023)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Gender, n (%)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>443 (59.9)</td>
<td>811 (79.3)</td>
</tr>
<tr>
<td>Male</td>
<td>274 (37.1)</td>
<td>184 (18)</td>
</tr>
<tr>
<td>Other</td>
<td>16 (2.2)</td>
<td>18 (1.8)</td>
</tr>
<tr>
<td>Age (years), mean (SD)</td>
<td>34.8 (12)</td>
<td>38.4 (13.6)</td>
</tr>
<tr>
<td><strong>Ethnicity, n (%)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White British</td>
<td>561 (75.9)</td>
<td>827 (80.8)</td>
</tr>
<tr>
<td>Other ethnicity</td>
<td>172 (23.3)</td>
<td>185 (18.1)</td>
</tr>
<tr>
<td><strong>Region of residence, n (%)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>East of England</td>
<td>53 (7.2)</td>
<td>61 (6)</td>
</tr>
<tr>
<td>London</td>
<td>166 (22.5)</td>
<td>210 (20.5)</td>
</tr>
<tr>
<td>Midlands</td>
<td>112 (15.2)</td>
<td>203 (19.8)</td>
</tr>
<tr>
<td>North East and Yorkshire</td>
<td>80 (10.8)</td>
<td>102 (10)</td>
</tr>
<tr>
<td>North West</td>
<td>66 (8.9)</td>
<td>98 (9.6)</td>
</tr>
<tr>
<td>South East</td>
<td>133 (18)</td>
<td>214 (20.9)</td>
</tr>
<tr>
<td>South West</td>
<td>123 (16.6)</td>
<td>125 (12.2)</td>
</tr>
<tr>
<td><strong>Highest educational qualification, n (%)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No qualification</td>
<td>51 (6.9)</td>
<td>30 (2.9)</td>
</tr>
<tr>
<td>O-levels or GCSE&lt;sup&gt;a&lt;/sup&gt;</td>
<td>117 (15.8)</td>
<td>116 (11.3)</td>
</tr>
<tr>
<td>A-levels or AS&lt;sup&gt;b&lt;/sup&gt;-levels or NVQ&lt;sup&gt;c&lt;/sup&gt; or equivalent</td>
<td>278 (37.6)</td>
<td>327 (32)</td>
</tr>
<tr>
<td>Degree-level qualification</td>
<td>207 (28)</td>
<td>349 (34.1)</td>
</tr>
<tr>
<td>Higher degree-level qualification</td>
<td>80 (10.8)</td>
<td>191 (18.7)</td>
</tr>
<tr>
<td><strong>Living arrangement, n (%)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alone</td>
<td>215 (29.1)</td>
<td>229 (22.4)</td>
</tr>
<tr>
<td>With others</td>
<td>524 (70.9)</td>
<td>794 (77.6)</td>
</tr>
<tr>
<td><strong>Employment status, n (%)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Employed</td>
<td>277 (37.5)</td>
<td>586 (57.3)</td>
</tr>
<tr>
<td>Sheltered employment</td>
<td>10 (1.4)</td>
<td>6 (0.6)</td>
</tr>
<tr>
<td>Training and education</td>
<td>76 (10.3)</td>
<td>106 (10.4)</td>
</tr>
<tr>
<td>Unemployed</td>
<td>356 (48.2)</td>
<td>272 (26.6)</td>
</tr>
<tr>
<td>Retired</td>
<td>20 (2.7)</td>
<td>53 (5.2)</td>
</tr>
<tr>
<td><strong>Current mental health problem, n (%)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I don’t want to say</td>
<td>20 (2.7)</td>
<td>14 (1.4)</td>
</tr>
<tr>
<td>I did not experience mental health problems</td>
<td>19 (2.6)</td>
<td>31 (3)</td>
</tr>
<tr>
<td>Developmental disorder such as learning disability</td>
<td>15 (2)</td>
<td>12 (1.2)</td>
</tr>
<tr>
<td>Eating disorder</td>
<td>15 (2)</td>
<td>45 (4.4)</td>
</tr>
<tr>
<td>Mood disorder</td>
<td>265 (35.9)</td>
<td>626 (61.2)</td>
</tr>
<tr>
<td>Personality disorder</td>
<td>138 (18.7)</td>
<td>123 (12)</td>
</tr>
<tr>
<td>Schizophrenia or other psychosis</td>
<td>154 (20.8)</td>
<td>&lt;5 (&lt;1)</td>
</tr>
<tr>
<td>Stress-related disorder</td>
<td>82 (11.1)</td>
<td>152 (14.9)</td>
</tr>
<tr>
<td>Substance-related disorder</td>
<td>25 (3.4)</td>
<td>&lt;10 (&lt;1)</td>
</tr>
</tbody>
</table>
Patterns of Narrative Requests and Feedback

Table 4 shows summary statistics on the participants, narratives, narrative requests, narrative ratings, intervention use length, and narrative request routes.

A histogram of the lengths of intervention use is shown in Figure S1 in Multimedia Appendix 5. In the NEON trial, 12.4% (92/739) of the participants used the intervention only once, whereas in the NEON-O trial, 19.45% (199/1023) of the participants used the intervention only once. Lengths of >400 days were merged to prevent participant identifiability. The lengths of intervention use for the first 30 days are shown in Figure S2 in Multimedia Appendix 5, with participants who only used the intervention once (“single-use participants”) removed to show only nonzero time lengths.

Tables 5 and 6 show the number of narrative rating values that each question received for ratings from NEON trial participants. Tables 7 and 8 show the number of narrative rating values that each question received for ratings from NEON-O trial participants.

---

### Lifetime user of primary care mental health services, n (%)

<table>
<thead>
<tr>
<th></th>
<th>NEON baseline (N=739)</th>
<th>NEON-O baseline (N=1023)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>698 (94.5)</td>
<td>949 (92.8)</td>
</tr>
<tr>
<td>No</td>
<td>35 (4.7)</td>
<td>64 (6.3)</td>
</tr>
</tbody>
</table>

### Current use of mental health services for psychosis, n (%)

<table>
<thead>
<tr>
<th>Service Type</th>
<th>NEON baseline (N=739)</th>
<th>NEON-O baseline (N=1023)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No contact with any NHS&lt;sup&gt;d&lt;/sup&gt; service</td>
<td>100 (13.5)</td>
<td>N/A&lt;sup&gt;e&lt;/sup&gt;</td>
</tr>
<tr>
<td>General practitioner</td>
<td>234 (31.7)</td>
<td>N/A</td>
</tr>
<tr>
<td>Primary care counselor</td>
<td>59 (8)</td>
<td>N/A</td>
</tr>
<tr>
<td>IAPT&lt;sup&gt;f&lt;/sup&gt;</td>
<td>56 (7.6)</td>
<td>N/A</td>
</tr>
<tr>
<td>Specialist community mental health team</td>
<td>261 (35.3)</td>
<td>N/A</td>
</tr>
<tr>
<td>Mental health inpatient in hospital</td>
<td>18 (2.4)</td>
<td>N/A</td>
</tr>
</tbody>
</table>

### How would you best describe your recovery?, n (%)

<table>
<thead>
<tr>
<th>Description</th>
<th>NEON baseline (N=739)</th>
<th>NEON-O baseline (N=1023)</th>
</tr>
</thead>
<tbody>
<tr>
<td>I don’t want to say</td>
<td>48 (6.5)</td>
<td>64 (6.3)</td>
</tr>
<tr>
<td>Not yet thinking about recovery</td>
<td>91 (12.3)</td>
<td>64 (6.3)</td>
</tr>
<tr>
<td>Working on recovery</td>
<td>510 (69)</td>
<td>784 (76.6)</td>
</tr>
<tr>
<td>Living beyond disability</td>
<td>84 (11.4)</td>
<td>101 (9.9)</td>
</tr>
</tbody>
</table>

---

<sup>a</sup>GCSE: General Certificate of Secondary Education.

<sup>b</sup>AS: Advanced Subsidiary.

<sup>c</sup>NVQ: National Vocational Qualification.

<sup>d</sup>NHS: National Health Service.

<sup>e</sup>N/A: not applicable; indicates a question that participants were not asked; in particular, only NEON trial participants were asked about their current use of mental health services.

<sup>f</sup>IAPT: Improving Access to Psychological Therapies.
### Table 4. Number of narrative requests via the content-based filtering internal access route, collaborative filtering internal access route, and all other internal access routes.

<table>
<thead>
<tr>
<th></th>
<th>NEON(^a)</th>
<th>NEON-O(^b)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participants, n (%)</td>
<td>739 (100)</td>
<td>1023 (100)</td>
</tr>
<tr>
<td>Participants who requested at least one narrative</td>
<td>365 (49.4)</td>
<td>562 (54.9)</td>
</tr>
<tr>
<td>Participants who requested and rated at least one narrative</td>
<td>284 (38.4)</td>
<td>409 (40)</td>
</tr>
<tr>
<td>Narratives at the start of the trial, n (%)</td>
<td>348 (100)</td>
<td>348 (100)</td>
</tr>
<tr>
<td>Narratives at the end of the trial, n (%)</td>
<td>657 (100)</td>
<td>657 (100)</td>
</tr>
<tr>
<td>Narratives given at least one rating</td>
<td>375 (57.1)</td>
<td>366 (55.7)</td>
</tr>
<tr>
<td>Narrative requests, n (%)</td>
<td>3762 (100)</td>
<td>3548 (100)</td>
</tr>
<tr>
<td>Optional ratings</td>
<td>2288 (100)</td>
<td>1896 (100)</td>
</tr>
<tr>
<td>Ratings per narrative, median (IQR)</td>
<td>1 (0-4)</td>
<td>1 (0-3)</td>
</tr>
<tr>
<td>Ratings per rated narrative, median (IQR)</td>
<td>3 (2-6)</td>
<td>2 (1-5)</td>
</tr>
<tr>
<td>Length of intervention use, median (IQR)</td>
<td>20 days, 22 hours, and 17 minutes (0 days, 0 minutes, and 0 minutes-251 days, 0 minutes, and 0 minutes)</td>
<td>0 days, 0 hours, and 16 minutes (0 days, 0 hours, and 0 minutes-59 days, 19 hours, and 7 minutes)</td>
</tr>
<tr>
<td>Content-based filtering narrative request route, n (%)</td>
<td>554 (14.7)</td>
<td>554 (15.6)</td>
</tr>
<tr>
<td>Collaborative filtering narrative request route, n (%)</td>
<td>1113 (29.6)</td>
<td>763 (21.5)</td>
</tr>
<tr>
<td>Other narrative request route, n (%)</td>
<td>2095 (55.7)</td>
<td>2212 (62.9)</td>
</tr>
</tbody>
</table>

\(^a\)NEON: Narrative Experiences Online.
\(^b\)NEON-O: NEON for other (eg, nonpsychosis) mental health problems trial.

### Table 5. Distribution of narrative rating values for the Narrative Experiences Online trial participants.

<table>
<thead>
<tr>
<th></th>
<th>−1, n (%)</th>
<th>0, n (%)</th>
<th>1, n (%)</th>
<th>2, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hopefulness ratings</td>
<td>202 (8.83)</td>
<td>901 (39.38)</td>
<td>838 (36.63)</td>
<td>347 (15.17)</td>
</tr>
</tbody>
</table>

### Table 6. Distribution of narrative rating values for the Narrative Experiences Online trial participants (N=538 ratings).

<table>
<thead>
<tr>
<th></th>
<th>0, n (%)</th>
<th>1, n (%)</th>
<th>2, n (%)</th>
<th>3, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Similarity to the narrator</td>
<td>132 (24.5)</td>
<td>196 (36.4)</td>
<td>152 (28.3)</td>
<td>58 (10.8)</td>
</tr>
<tr>
<td>Similarity to the narrative</td>
<td>103 (19.1)</td>
<td>144 (26.8)</td>
<td>173 (32.2)</td>
<td>118 (21.9)</td>
</tr>
<tr>
<td>Learning ratings</td>
<td>104 (19.3)</td>
<td>193 (35.9)</td>
<td>181 (33.6)</td>
<td>60 (11.2)</td>
</tr>
<tr>
<td>Empathy ratings</td>
<td>155 (28.8)</td>
<td>206 (38.3)</td>
<td>126 (23.4)</td>
<td>51 (9.5)</td>
</tr>
</tbody>
</table>

### Table 7. Distribution of narrative rating values from Narrative Experiences Online for other (eg, nonpsychosis) mental health problems trial participants (N=1896 narrative ratings).

<table>
<thead>
<tr>
<th></th>
<th>−1, n (%)</th>
<th>0, n (%)</th>
<th>1, n (%)</th>
<th>2, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hopefulness ratings</td>
<td>206 (10.86)</td>
<td>845 (44.57)</td>
<td>649 (34.23)</td>
<td>196 (10.34)</td>
</tr>
</tbody>
</table>

### Table 8. Distribution of narrative rating values from Narrative Experiences Online for other (eg, nonpsychosis) mental health problems trial participants (N=538 ratings).

<table>
<thead>
<tr>
<th></th>
<th>0, n (%)</th>
<th>1, n (%)</th>
<th>2, n (%)</th>
<th>3, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Similarity to the narrator</td>
<td>195 (36.2)</td>
<td>211 (39.2)</td>
<td>104 (19.3)</td>
<td>28 (5.2)</td>
</tr>
<tr>
<td>Similarity to the narrative</td>
<td>145 (27)</td>
<td>157 (29.2)</td>
<td>168 (31.2)</td>
<td>68 (12.6)</td>
</tr>
<tr>
<td>Learning ratings</td>
<td>134 (24.9)</td>
<td>242 (45)</td>
<td>144 (26.8)</td>
<td>18 (3.3)</td>
</tr>
<tr>
<td>Empathy ratings</td>
<td>236 (43.9)</td>
<td>193 (35.9)</td>
<td>90 (16.7)</td>
<td>19 (3.5)</td>
</tr>
</tbody>
</table>
Objective 2: Evaluate the NarraGive Recommender System

Overview

The best results (per metric per trial) are italicized. Where 2 values are equal, neither was better than the other.

For rating sets, better means that all 6 values (across both trials) were better than the 2 corresponding values for the other 2 rating sets, with N/A if no rating set was better.

For algorithms, we identified the filtering approach that was better (if any), comparing the content-based and collaborative subsystems of NarraGive per rating set across both trials. Specifically, if the kNN value was better than both SVD and SVD++ values, then we identified content-based filtering as better. If both SVD and SVD++ values were better than the kNN value, then we identified collaborative filtering as better. If neither the kNN nor SVD and SVD++ was better than the other, then the value was calculated per trial.

For trials, better means that each of the 9 values was better than the corresponding value in the other trial, with N/A if neither trial was better.

Prediction Accuracy

Tables 9 and 10 show the NMAE of the kNN, SVD, and SVD++ algorithms when trained and tested on the hopefulness, similarity to the narrator, and similarity to the narrative ratings using NEON and NEON-O trial data, respectively.

For NMAE, better means lower. Hopefulness was the better rating set, collaborative filtering was the better approach, and NEON-O was the better trial.

Table 9. Normalized mean average error (NMAE; using Narrative Experiences Online [NEON] trial data).

<table>
<thead>
<tr>
<th>NMAE (NEON trial)</th>
<th>Hopefulness, median (IQR)</th>
<th>Similarity to the narrator, median (IQR)</th>
<th>Similarity to the narrative, median (IQR)</th>
</tr>
</thead>
<tbody>
<tr>
<td>kNN(^a)</td>
<td>0.686 (0.670-0.703)</td>
<td>1.070 (1.059-1.077)</td>
<td>1.150 (1.140-1.153)</td>
</tr>
<tr>
<td>SVD(^b)</td>
<td>0.650 (0.638-0.664)</td>
<td>1.043 (1.035-1.047)</td>
<td>1.098 (1.076-1.121)</td>
</tr>
<tr>
<td>SVD++</td>
<td>0.646 (0.639-0.654) (^c)</td>
<td>1.044 (1.038-1.049)</td>
<td>1.099 (1.080-1.120)</td>
</tr>
</tbody>
</table>

\(^a\)kNN: k-nearest neighbor.
\(^b\)SVD: singular value decomposition.
\(^c\)Best result is italicized (per metric per trial).

Table 10. Normalized mean average error (NMAE; using Narrative Experiences Online for other [eg, nonpsychosis] mental health problems trial [NEON-O] data).

<table>
<thead>
<tr>
<th>NMAE (NEON-O trial)</th>
<th>Hopefulness, median (IQR)</th>
<th>Similarity to the narrator, median (IQR)</th>
<th>Similarity to the narrative, median (IQR)</th>
</tr>
</thead>
<tbody>
<tr>
<td>kNN(^a)</td>
<td>0.685 (0.677-0.697)</td>
<td>0.998 (0.992-1.006)</td>
<td>1.099 (1.097-1.106)</td>
</tr>
<tr>
<td>SVD(^b)</td>
<td>0.650 (0.641-0.659)</td>
<td>0.978 (0.972-0.986)</td>
<td>1.076 (1.066-1.093)</td>
</tr>
<tr>
<td>SVD++</td>
<td>0.644 (0.635-0.655) (^c)</td>
<td>0.983 (0.977-0.989)</td>
<td>1.079 (1.065-1.099)</td>
</tr>
</tbody>
</table>

\(^a\)kNN: k-nearest neighbor.
\(^b\)SVD: singular value decomposition.
\(^c\)Best result is italicized (per metric per trial).

Usage prediction

Tables 11 and 12 show the precision of the kNN, SVD, and SVD++ algorithms when trained and tested on the hopefulness, similarity to the narrator, and similarity to the narrative ratings using NEON and NEON-O trial data, respectively.

For precision, better means higher. Hopefulness was the better rating set, there was no better filtering approach, and NEON was the better trial.

Table 11. Precision (using Narrative Experiences Online [NEON] trial data).

<table>
<thead>
<tr>
<th>Precision (NEON trial)</th>
<th>Hopefulness, median (IQR)</th>
<th>Similarity to the narrator, median (IQR)</th>
<th>Similarity to the narrative, median (IQR)</th>
</tr>
</thead>
<tbody>
<tr>
<td>kNN(^a)</td>
<td>0.255 (0.250-0.258)</td>
<td>0.054 (0.052-0.055)</td>
<td>0.057 (0.056-0.060)</td>
</tr>
<tr>
<td>SVD(^b)</td>
<td>0.256 (0.250-0.259) (^c)</td>
<td>0.053 (0.052-0.055)</td>
<td>0.057 (0.055-0.060)</td>
</tr>
<tr>
<td>SVD++</td>
<td>0.256 (0.250-0.259)</td>
<td>0.053 (0.051-0.055)</td>
<td>0.057 (0.055-0.060)</td>
</tr>
</tbody>
</table>

\(^a\)kNN: k-nearest neighbor.
\(^b\)SVD: singular value decomposition.
\(^c\)Best result is italicized (per metric per trial).
Table 12. Precision (using Narrative Experiences Online for other [eg, nonpsychosis] mental health problems trial [NEON-O] trial data).

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Similarity to the narrator, median (IQR)</th>
<th>Similarity to the narrative, median (IQR)</th>
</tr>
</thead>
<tbody>
<tr>
<td>kNN^a</td>
<td>0.037 (0.034-0.040)</td>
<td>0.041 (0.038-0.043)</td>
</tr>
<tr>
<td>SVD^b</td>
<td>0.037 (0.035-0.039)</td>
<td>0.041 (0.039-0.043)</td>
</tr>
<tr>
<td>SVD++</td>
<td>0.036 (0.033-0.039)</td>
<td>0.040 (0.038-0.043)</td>
</tr>
</tbody>
</table>

^a kNN: k-nearest neighbor.
^b Best result is italicized (per metric per trial).
^c SVD: singular value decomposition.

Diversity

Tables 13 and 14 show the ILD of the kNN, SVD, and SVD++ algorithms when trained and tested on the hopefulness, similarity to the narrator, and similarity to the narrative ratings using NEON and NEON-O trial data, respectively.

Table 13. Intralist diversity (ILD; using Narrative Experiences Online [NEON] trial data).

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Similarity to the narrative, median (IQR)</th>
<th>Similarity to the narrator, median (IQR)</th>
</tr>
</thead>
<tbody>
<tr>
<td>kNN^a</td>
<td>0.540 (0.540-0.540)</td>
<td>0.538 (0.538-0.538)</td>
</tr>
<tr>
<td>SVD^b</td>
<td>0.539 (0.538-0.539)</td>
<td>0.538 (0.537-0.538)</td>
</tr>
<tr>
<td>SVD++</td>
<td>0.539 (0.539-0.539)</td>
<td>0.538 (0.539-0.539)</td>
</tr>
</tbody>
</table>

^a kNN: k-nearest neighbor.
^b SVD: singular value decomposition.
^c Best result is italicized (per metric per trial).


<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Similarity to the narrative, median (IQR)</th>
<th>Similarity to the narrator, median (IQR)</th>
</tr>
</thead>
<tbody>
<tr>
<td>kNN^a</td>
<td>0.499 (0.499-0.499)</td>
<td>0.500 (0.500-0.500)</td>
</tr>
<tr>
<td>SVD^b</td>
<td>0.499 (0.499-0.499)</td>
<td>0.499 (0.499-0.500)</td>
</tr>
<tr>
<td>SVD++</td>
<td>0.499 (0.499-0.499)</td>
<td>0.499 (0.499-0.499)</td>
</tr>
</tbody>
</table>

^a kNN: k-nearest neighbor.
^b Best result is italicized (per metric per trial).
^c SVD: singular value decomposition.

Coverage

Tables 15 and 16 show the ISC of the kNN, SVD, and SVD++ algorithms when trained and tested on the hopefulness, similarity to the narrator, and similarity to the narrative ratings using NEON and NEON-O trial data, respectively.

Table 15. Item space coverage (ISC; using Narrative Experiences Online [NEON] trial data).

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Similarity to the narrator, median (IQR)</th>
<th>Similarity to the narrative, median (IQR)</th>
</tr>
</thead>
<tbody>
<tr>
<td>kNN^a</td>
<td>0.811 (0.811-0.811)</td>
<td>0.800 (0.800-0.800)</td>
</tr>
<tr>
<td>SVD^b</td>
<td>0.761 (0.759-0.764)</td>
<td>0.763 (0.763-0.764)</td>
</tr>
<tr>
<td>SVD++</td>
<td>0.771 (0.769-0.772)</td>
<td></td>
</tr>
</tbody>
</table>

^a kNN: k-nearest neighbor.
^b Best results is italicized (per metric per trial).
^c SVD: singular value decomposition.

For ILD, better means lower.
There was no better rating set, collaborative filtering was the better approach for the NEON trial, there was no better approach for the NEON-O trial, and NEON-O was the better trial.

For ISC, better means higher.
There was no better rating set, content-based filtering was the better approach for the NEON trial, there was no better approach for the NEON-O trial, and NEON-O was the better trial.
Table 16. Item space coverage (ISC; using Narrative Experiences Online for other [eg, nonpsychosis] mental health problems trial [NEON-O] trial data).

<table>
<thead>
<tr>
<th>ISC (NEON-O trial)</th>
<th>Hopefulness, median (IQR)</th>
<th>Similarity to the narrator, median (IQR)</th>
<th>Similarity to the narrative, median (IQR)</th>
</tr>
</thead>
<tbody>
<tr>
<td>kNN(^a)</td>
<td>0.891 (0.891-0.891)(^b)</td>
<td>0.852 (0.852-0.852)</td>
<td>0.847 (0.847-0.847)</td>
</tr>
<tr>
<td>SVD(^c)</td>
<td>0.848 (0.844-0.852)</td>
<td>0.840 (0.838-0.842)</td>
<td>0.847 (0.846-0.847)</td>
</tr>
<tr>
<td>SVD++</td>
<td>0.850 (0.849-0.851)</td>
<td>0.842 (0.841-0.843)</td>
<td>0.848 (0.844-0.852)</td>
</tr>
</tbody>
</table>

\(^a\)kNN: k-nearest neighbor.
\(^b\)Best results is italicized (per metric per trial).
\(^c\)SVD: singular value decomposition.

Unfairness Across Participants

Tables 17 and 18 show the unfairness, based on gender, of the kNN, SVD, and SVD++ algorithms when trained and tested on the hopefulness, similarity to the narrator, and similarity to the narrative using NEON and NEON-O trial data, respectively.

For unfairness across participants based on gender, better means lower.

Hopefulness was the better rating set, collaborative filtering was the better approach, and there was no better trial.

Table 17. Unfairness across participants based on gender (using Narrative Experiences Online [NEON] trial data).

<table>
<thead>
<tr>
<th>Unfairness (gender; NEON trial)</th>
<th>Hopefulness, median (IQR)</th>
<th>Similarity to the narrator, median (IQR)</th>
<th>Similarity to the narrative, median (IQR)</th>
</tr>
</thead>
<tbody>
<tr>
<td>kNN(^a)</td>
<td>0.429 (0.407-0.455)</td>
<td>0.715 (0.687-0.738)</td>
<td>0.728 (0.715-0.738)</td>
</tr>
<tr>
<td>SVD(^b)</td>
<td>0.375 (0.347-0.422)</td>
<td>0.664 (0.637-0.695)</td>
<td>0.683 (0.669-0.688)</td>
</tr>
<tr>
<td>SVD++</td>
<td>0.371 (0.342-0.412)(^c)</td>
<td>0.673 (0.643-0.706)</td>
<td>0.695 (0.680-0.700)</td>
</tr>
</tbody>
</table>

\(^a\)kNN: k-nearest neighbor.
\(^b\)SVD: singular value decomposition.
\(^c\)Best results are italicized (per metric per trial).

Table 18. Unfairness across participants based on gender (using Narrative Experiences Online for other [eg, nonpsychosis] mental health problems trial [NEON-O] trial data).

<table>
<thead>
<tr>
<th>Unfairness (gender; NEON-O trial)</th>
<th>Hopefulness, median (IQR)</th>
<th>Similarity to the narrator, median (IQR)</th>
<th>Similarity to the narrative, median (IQR)</th>
</tr>
</thead>
<tbody>
<tr>
<td>kNN(^a)</td>
<td>0.387 (0.368-0.410)</td>
<td>0.727 (0.685-0.769)</td>
<td>0.765 (0.757-0.781)</td>
</tr>
<tr>
<td>SVD(^b)</td>
<td>0.327 (0.312-0.344)</td>
<td>0.640 (0.634-0.652)</td>
<td>0.671 (0.659-0.696)</td>
</tr>
<tr>
<td>SVD++</td>
<td>0.317 (0.303-0.338)(^c)</td>
<td>0.647 (0.639-0.662)</td>
<td>0.685 (0.678-0.705)</td>
</tr>
</tbody>
</table>

\(^a\)kNN: k-nearest neighbor.
\(^b\)SVD: singular value decomposition.
\(^c\)Best results are italicized (per metric per trial).
### Table 19. Unfairness across participants based on ethnicity (using Narrative Experiences Online [NEON] trial data).

<table>
<thead>
<tr>
<th>Method</th>
<th>Similarity to the narrator, median (IQR)</th>
<th>Similarity to the narrative, median (IQR)</th>
</tr>
</thead>
<tbody>
<tr>
<td>kNN(^a)</td>
<td>0.769 (0.757-0.783)</td>
<td>0.795 (0.776-0.808)</td>
</tr>
<tr>
<td>SVD(^b)</td>
<td>0.732 (0.712-0.742)</td>
<td>0.727 (0.709-0.745)</td>
</tr>
<tr>
<td>SVD++</td>
<td>0.744 (0.722-0.755)</td>
<td>0.739 (0.716-0.763)</td>
</tr>
</tbody>
</table>

\(^a\)kNN: k-nearest neighbor.
\(^b\)SVD: singular value decomposition.
\(^c\)Best results are italicized (per metric per trial).

### Table 20. Unfairness across participants based on ethnicity (using Narrative Experiences Online for other [eg, nonpsychosis] mental health problems trial [NEON-O] trial data).

<table>
<thead>
<tr>
<th>Method</th>
<th>Similarity to the narrator, median (IQR)</th>
<th>Similarity to the narrative, median (IQR)</th>
</tr>
</thead>
<tbody>
<tr>
<td>kNN(^a)</td>
<td>0.717 (0.687-0.754)</td>
<td>0.751 (0.724-0.787)</td>
</tr>
<tr>
<td>SVD(^b)</td>
<td>0.652 (0.642-0.657)</td>
<td>0.667 (0.661-0.690)</td>
</tr>
<tr>
<td>SVD++</td>
<td>0.658 (0.649-0.664)</td>
<td>0.688 (0.680-0.708)</td>
</tr>
</tbody>
</table>

\(^a\)kNN: k-nearest neighbor.
\(^b\)SVD: singular value decomposition.
\(^c\)Best results are italicized (per metric per trial).

### MAE Over Time

Multimedia Appendix 6 shows how the median NMAE values changed over time (with an interval of 1 month) for the kNN, SVD, and SVD++ algorithms using “Hopefulness” ratings from NEON trial participants.

Figure 1 shows that the 2 collaborative filtering algorithms were more accurate than the content-based filtering algorithm. As the number of ratings increases (and the IQR decreases), the NMAE stabilizes, which happens for all 3 algorithms at approximately 2000 ratings.

Figure 1. Comparison of the accuracy of the k-nearest neighbor (kNN), singular value decomposition (SVD), and SVD++ algorithms over time. NMAE: normalized mean absolute error.
**Other Results**

Further analysis of the coverage metric showed that certain narratives were not routinely recommended by NarraGive, as described in Multimedia Appendix 7.

**Discussion**

**Principal Findings**

**Overview**

For the NEON trial, the content-based filtering algorithm performed better for coverage; the collaborative filtering algorithms performed better for accuracy, diversity, and unfairness across both gender and ethnicity; and neither algorithm performed better for precision. For the NEON-O trial, the content-based filtering algorithm did not perform better on any metric; the collaborative filtering algorithms performed better on accuracy and unfairness across both gender and ethnicity; and neither algorithm performed better for precision.

**Table 21.** Summary of Narrage evaluation metrics.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Prediction based or feature based</th>
<th>Better filtering method</th>
<th>Filtering method of best value</th>
<th>Better trial</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>NEON&lt;sup&gt;a&lt;/sup&gt;</td>
<td>NEON-O&lt;sup&gt;b&lt;/sup&gt;</td>
<td></td>
</tr>
<tr>
<td>NMAE&lt;sup&gt;c&lt;/sup&gt;</td>
<td>Prediction</td>
<td>Collaborative</td>
<td>Collaborative</td>
<td>Collaborative</td>
</tr>
<tr>
<td>Precision</td>
<td>Prediction</td>
<td>N/A&lt;sup&gt;d&lt;/sup&gt;</td>
<td>N/A</td>
<td>Collaborative</td>
</tr>
<tr>
<td>ILD&lt;sup&gt;e&lt;/sup&gt;</td>
<td>Feature</td>
<td>Collaborative</td>
<td>N/A</td>
<td>Content</td>
</tr>
<tr>
<td>ISC&lt;sup&gt;f&lt;/sup&gt;</td>
<td>Feature</td>
<td>Content</td>
<td>N/A</td>
<td>Content</td>
</tr>
<tr>
<td>Unfairness (gender)</td>
<td>Prediction</td>
<td>Collaborative</td>
<td>Collaborative</td>
<td>Collaborative</td>
</tr>
<tr>
<td>Unfairness (ethnicity)</td>
<td>Prediction</td>
<td>Collaborative</td>
<td>Collaborative</td>
<td>N/A</td>
</tr>
</tbody>
</table>

<sup>a</sup>NEON: Narrative Experiences Online.

<sup>b</sup>NEON-O: Narrative Experiences Online for other (eg, nonpsychosis) mental health problems trial.

<sup>c</sup>NMAE: normalized mean absolute error.

<sup>d</sup>N/A: not applicable.

<sup>e</sup>ILD: intralist diversity.

<sup>f</sup>ISC: item space coverage.

**Unrecommendable Narratives**

Each trial comprised between 1 and 3 ratings for each narrative. This preliminary evaluation only used data from NEON and NEON-O participants (to mirror the data used in the metrics). No rating given by a NEON or NEON-O participant was the highest, and only 1 rating had the optional questions answered (and none of these values were the highest value). The low number of ratings and the low scores given could contribute to the unrecommendableness of these narratives.

NarraGive, and recommender systems in general, often requires that there is a minimum amount of information about a participant or an item before being able to produce recommendations for or about them. In this case, due to SurPRISE’s implementation of recommender system algorithms, the kNN, SVD, and SVD++ algorithms require at least one rating for an item for it to be recommended and at least one rating by a participant for them to be recommended a narrative (to the extent that the recommender system never sees these unrated items and nonrating participants as they are filtered out before being passed to the recommender system).

This means that newly added narratives cannot be recommended immediately—they need at least one participant to access the narrative through another method (such as browsing to it) and then rate it. Consequently, there are other unrecommendable narratives (where unrecommendable in this case means that the recommender system does not have access to it in the first place rather than having access to it but not producing it as a
recommending a similar diversity to any participant). This is known as the cold start problem.

**Interpretation of Metrics**

In the context of NarraGive, diversity is not necessarily better or worse. An earlier substudy of the NEON study showed that there is unlikely to be a universally hopeful narrative [29], so a skew in recommended narratives is not necessarily a flaw.

Similarly, having greater coverage (i.e., being able to recommend a greater proportion of available narratives) may not be useful if some narratives are only hope inspiring for a very small subset of participants.

The 2 unfairness metrics (unfairness across participants based on ethnicity and gender) capture 2 types of unfairness but not all. A previous NEON study [29] identified 7 harm minimization strategies for the NEON intervention, which provided the basis for the unfairness metric (i.e., that being recommended narratives with a predicted rating that is higher than the resulting rating may be harmful). However, other types of unfairness, such as those based on disability, were not explored.

**Relationship to Prior Work**

**Recommender Systems for Nonnarrative Texts**

Several book recommendation systems exist [59-61], but the focus has mostly been on novel-like books rather than on health narratives or recovery narratives. In addition, many health recommender systems exist or have been proposed [62-72], but the focus has largely been on physical health and behavior changes rather than on providing desired content, such as enjoyable, useful, or hope-inspiring recovery narratives.

**Dimensionality Reduction**

The per-trial values for ILD were very similar. One explanation for this is a known phenomenon called the curse of dimensionality, where the increase in the number of dimensions (where the number of INCREASE characteristics represents the number of dimensions) causes a rapid increase in the “volume” that samples can occupy, which increases the data sparsity exponentially [73]. This increase in dimensionality produces effects such as the concentration of measure [74], where distance values converge and the difference between the furthest and the nearest point tends toward 0 [75], effectively making distance-based similarity comparisons meaningless.

High-dimensionality problems can occur with as few as 10 dimensions [75], making the 77-item INCREASE measure susceptible to these issues. This may explain why the results for ILD are so similar—they are based on cosine distance. Other distance measures such as the Euclidean distance and correlation are also susceptible to this challenge [73].

One solution is to reduce the dimensionality of the characteristics before analysis through dimension reduction techniques [75] such as principal component analysis or matrix factorization, which retain as much of the original meaning of the data as possible while reducing the number of dimensions to a practical number.

**Strengths and Limitations**

There are several strengths to this study. First, the NEON and NEON-O trials produced a unique data set of participant ratings, comprising ratings from both mental health service users and non–service users. This data set was suitable for analysis over time and for comparison of content-based and collaborative filtering algorithms.

Second, the numerous narrative request routes helped prevent exposure bias, and requiring a rating for each narrative helped prevent selection bias.

There are also some limitations to this study. This analysis did not consider individual participants’ rating patterns. Further analysis could add participant and item biases, which take into account items’ and participants’ average ratings to find the deviation from this average [76], or weight high-data participants (who have provided many ratings) as more informative than low-data participants.

The decision to include “no change” in hopefulness as an indicator of relevance was made to distinguish from actively hope-reducing narratives, but an alternative approach would be to only include those narratives rated as “a bit more hopeful” or “much more hopeful.”

The unfairness across participants metric (overestimation of unfairness) was based on the assumption that overestimated narratives are more likely to be harmful, but it is possible that a participant could rate a narrative highly and still find it harmful and, similarly, rate a narrative as lower than predicted but not find it harmful. The unfairness metrics also did not cover all aspects of unfairness.

Finally, this study is the first evaluation of a recommender system application to lived experience narratives. This is a complex area involving both technical challenges such as the choice of algorithm and ethical challenges such as managing narratives with respect and not just as another form of data. This complexity means that there are no existing standards against which NarraGive can be currently judged, and hence, the comprehensive evaluation presented in this study is primarily intended to be formative for the field rather than evaluative of NarraGive.

**Recommendations**

There are 6 recommendations for researchers, intervention developers, recommender system developers, and health care professionals.

First, recommender systems with a focus on providing the greatest variety and widest range of content may benefit from using a content-based kNN algorithm, whereas recommender systems with a focus on predicting participants’ ratings most accurately may benefit more from the SVD or SVD++ algorithm. Recommender systems with a focus on both should implement a hybrid model with suitably weighted filtering algorithms.

Second, health care professionals should be aware of the unrecommendability of some items and not rely on recommendations to cover the entire search space.
Third, researchers and intervention developers should carefully decide which feature of the recommender system (ie, variety or accuracy) is most important and optimize the recommender system for a specific feature. Depending on what aspect of a recommender system is most important, different methods exist for optimizing for a specific metric, such as for diversity [43] and unfairness [46].

Fourth, platforms containing a recommender system should include other item access mechanisms (such as being able to browse through items) to prevent feedback loops where participants can only rate items that already have many high ratings while unrated items remain unrated and unrecommended [28]. This helps reduce the number of inaccessible narratives (because if the recommender system is the only access route, any unrecommendable narratives will be entirely inaccessible to participants), and developers could include a “random” access route (in addition to the recommender system access route) that is weighted toward these inaccessible narratives.

Fifth, recommender system developers should actively encourage the rating of new items, such as by suggesting them to participants or having a random button that is weighted toward new narratives. Alternatively, the narratives could be given an initial set of ratings before being published.

Sixth, initial studies should be conducted on proposed recommender systems to find the number of ratings required for the accuracy to stabilize; for NarraGive, this was approximately 2000 ratings. Because clinical population may be associated with recommender system performance, initial studies should also be used to inform clinical population selection.

Implications for Future Work
Key future questions include whether a single or hybrid recommender system is optimal, a wider consideration of available algorithms and clarification of the rationale for selection, the rationale and timing of training and retraining the algorithm, and the identification of the most important metrics through which algorithmic performance should be evaluated.

For example, future studies investigating the use of recommender systems for recommending narratives could incorporate the similarity between participants and narratives by training filtering algorithms to recommend narratives with narrators that are either similar to or different from participants depending on (either implicit or explicit) participant preference.

Conclusions
Clinical population may be associated with recommender system performance. The collaborative filtering algorithms were more accurate and less unfair than the content-based filtering algorithm. Recommender systems are susceptible to a wide range of biases, and it is important to mitigate these by providing enough data for the recommender system to start with (to prevent overfitting), ensuring that there are other ways of accessing items besides through the recommender system (to prevent a feedback loop between accessed items and recommended items), and encouraging participants to provide feedback on every item they interact with (to prevent participants from only providing feedback when they have strong opinions).

Acknowledgments
This study was funded by the National Institute for Health and Care Research (NIHR; personal experience as a recovery resource in psychosis: Narrative Experiences Online Programme [RP-PG-0615-20016]). MS and SRE acknowledge the support of the NIHR Nottingham Biomedical Research Centre (NIHR203310). The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care. FN acknowledges support from an NIHR Advanced Fellowship (NIHR302218) and an Anne McLaren Fellowship.

Data Availability
Data will be available upon reasonable request, supervised by the study sponsor. Only anonymous and pseudonymous elements of the data sets used or analyzed during the study will be available. Research data will be available from the study sponsor until the end of the retention period. After the retention period, availability through the study sponsor or chief investigator may be at their discretion. Contact the study sponsor through Research@nottshc.nhs.uk citing Integrated Research Application System ID249015.

Conflicts of Interest
None declared.

Multimedia Appendix 1
Definitions of the categories within the “Browse Stories” page of the Narrative Experiences Online intervention and screenshots of the web application.

[PDF File (Adobe PDF File), 190 KB - mental_v11i1e45754_app1.pdf ]

Multimedia Appendix 2
Available options in the Personal Profile in the Narrative Experiences Online intervention.
Multimedia Appendix 3
Equations used for the normalized mean absolute error, precision, intralist diversity, item space coverage, and unfairness across participants.

Multimedia Appendix 4
Rationale for not using 5 of the evaluation categories.

Multimedia Appendix 5
Graphs showing the length of use of the Narrative Experiences Online (NEON) intervention comparing NEON and NEON for other (e.g., nonpsychosis) mental health problems (NEON-O) trial participants.

Multimedia Appendix 6
Graphs showing the normalized mean absolute error over time for the k-nearest neighbor, singular value decomposition (SVD), and SVD++ algorithms.

Multimedia Appendix 7
Nonrecommendable narratives.

References


Abbreviations

FEVR: framework for evaluating recommender systems
ILD: intralist diversity
INCRESE: Inventory of Characteristics of Recovery Stories
kNN: k-nearest neighbor
MAE: mean absolute error
NEON: Narrative Experiences Online
NEON-O: Narrative Experiences Online for other (e.g., nonpsychosis) mental health problems
NMAE: normalized mean absolute error
RMSE: root-mean-square error
SurPRISE: Simple Python Recommendation System Engine
SVD: singular value decomposition
A Comparison of ChatGPT and Fine-Tuned Open Pre-Trained Transformers (OPT) Against Widely Used Sentiment Analysis Tools: Sentiment Analysis of COVID-19 Survey Data

Juan Antonio Lossio-Ventura¹, PhD; Rachel Weger², BA; Angela Y Lee³, MA; Emily P Guinee¹, BA; Joyce Chung¹, MD; Lauren Atlas⁴, PhD; Eleni Linos⁵, MD; Francisco Pereira¹, PhD

¹National Institute of Mental Health, National Institutes of Health, Bethesda, MD, United States
²School of Medicine, University of Pittsburgh, Pittsburgh, PA, United States
³Department of Communication, Stanford University, Stanford, CA, United States
⁴National Center For Complementary and Alternative Medicine, National Institutes of Health, Bethesda, MD, United States
⁵School of Medicine, Stanford University, Stanford, CA, United States

Corresponding Author:
Juan Antonio Lossio-Ventura, PhD
National Institute of Mental Health
National Institutes of Health
3D41
10 Center Dr
Bethesda, MD, 20814
United States
Phone: 1 3018272632
Email: juan.lossio@nih.gov

Abstract

Background: Health care providers and health-related researchers face significant challenges when applying sentiment analysis tools to health-related free-text survey data. Most state-of-the-art applications were developed in domains such as social media, and their performance in the health care context remains relatively unknown. Moreover, existing studies indicate that these tools often lack accuracy and produce inconsistent results.

Objective: This study aims to address the lack of comparative analysis on sentiment analysis tools applied to health-related free-text survey data in the context of COVID-19. The objective was to automatically predict sentence sentiment for 2 independent COVID-19 survey data sets from the National Institutes of Health and Stanford University.

Methods: Gold standard labels were created for a subset of each data set using a panel of human raters. We compared 8 state-of-the-art sentiment analysis tools on both data sets to evaluate variability and disagreement across tools. In addition, few-shot learning was explored by fine-tuning Open Pre-Trained Transformers (OPT; a large language model [LLM] with publicly available weights) using a small annotated subset and zero-shot learning using ChatGPT (an LLM without available weights).

Results: The comparison of sentiment analysis tools revealed high variability and disagreement across the evaluated tools when applied to health-related survey data. OPT and ChatGPT demonstrated superior performance, outperforming all other sentiment analysis tools. Moreover, ChatGPT outperformed OPT, exhibited higher accuracy by 6% and higher F-measure by 4% to 7%.

Conclusions: This study demonstrates the effectiveness of LLMs, particularly the few-shot learning and zero-shot learning approaches, in the sentiment analysis of health-related survey data. These results have implications for saving human labor and improving efficiency in sentiment analysis tasks, contributing to advancements in the field of automated sentiment analysis.

(JMIR Ment Health 2024;11:e50150) doi:10.2196/50150

KEYWORDS
sentiment analysis; COVID-19 survey; large language model; few-shot learning; zero-shot learning; ChatGPT; COVID-19
Introduction

Background
Sentiment analysis is a field within natural language processing (NLP) that aims to extract sentiments and opinions from text related to specific entities and topics [1], such as people, organizations, events, and places [2]. Specifically, we consider the task of classifying texts as positive, neutral, or negative. Research in this area can occur at different levels of granularity, ranging from a single sentiment for an entire document to exploring various aspects associated with each entity, which can be associated with different sentiments [1,3].

Recently, we have witnessed an increase in the use of sentiment analysis to computationally evaluate the attitudes, perceptions, and emotions of social media users regarding the COVID-19 pandemic [4,5]. Most of these works study content from social media platforms such as Twitter, Reddit, and Facebook [6], as social media has been a main platform to express opinions related to COVID-19 in a public manner. Simultaneously, surveys, which refer to data collected from a group of people regarding their opinions, behavior, or knowledge through specifically designed questions, have also been used to investigate the impact of the COVID-19 pandemic. In particular, surveys conducted during the lockdown period in 2020 examined the effects on people’s lives, behaviors, and mental health, among other topics [7-9]. Web-based surveys are often semistructured, that is, composed of closed-answer components (eg, different clinical questionnaires) and open-ended questions that allow a free-text answer. Sentiment analysis tools have been applied to the latter to help monitor the attitudes, sentiments, and perceptions of the participants during the pandemic to assist health decision-making [10].

The application of sentiment analysis tools on free-text data obtained from surveys poses challenges for health care providers and researchers in the health domain. This is partly attributed to the fact that most state-of-the-art applications are designed for different domains, such as social media, and there is limited knowledge regarding their performance in survey data. In addition, recent studies have applied the most well-known sentiment analysis tools, including TextBlob [11], VADER (Valence Aware Dictionary and Sentiment Reasoner) [12], and Stanza [13], to analyze health-related content on social media platforms [14-16] and, more recently, in the context of COVID-19 [6,17]. These studies highlighted the need for a more comprehensive evaluation of sentiment analysis tools, as the initial results exhibited a lack of accuracy and yielded inconsistent outcomes [15,16]. The main reason for this discrepancy was the disparity in data sets and the potential sensitivity of the tools to the composition of the data set [16]. Consequently, researchers trained new algorithms tailored to their specific data set.

Two COVID-19 survey data sets were used in this study, both collected by teams from the National Institutes of Health (NIH) and Stanford University. The collected data were used to assess the general topics experienced by the participants during the pandemic lockdown.

Researchers from both institutions aimed to comprehend the general sentiment patterns over time and identify an overall sentiment for events during that period, such as vaccines and the 2020 presidential elections. In both data sets, it was often the case that a complete response contained multiple topics, with many sentences referring to distinct subjects. Thus, this study is focused on the analysis of sentiment at the sentence level. By assessing each sentence independently, subtle shifts in sentiment could be captured, which could potentially be neglected at the document level. Moreover, we thought that an analysis based on sentence level, rather than aspect-based level, was more appropriate, given that our focus was not on the granularity of the various aspects of an entity. For instance, when evaluating different features of an intensive care unit, aspects might encompass ventilators, rooms, staff, nurses, and others. Therefore, the decision to focus on sentence-level sentiment analysis is influenced by practical considerations, our research objectives, and the nature of the survey responses.

In this study, as the first contribution, we analyzed 2 independent survey data sets containing free-text data collected during the lockdown period of the COVID-19 pandemic, with accompanying ground-truth sentiment labels generated by human raters for hundreds of responses. The second contribution involves a comparison of 8 widely used state-of-the-art sentiment analysis tools, which have been frequently and recently used in the health domain [16], on COVID-19 surveys at the sentence level. We demonstrate that performance across tools varies and that there is a complex correlation structure between their predicted polarity scores. The third contribution of this paper is to investigate whether the polarity prediction performance can be improved through few-shot learning on a small labeled data set or zero-shot learning with ChatGPT [18].

Related Work
There are 2 main approaches to performing sentiment analysis: lexicon based and machine learning based. Initial lexicon methods are the simplest rule-based methods and seek to classify the sentiment of a sentence as a score function of the word polarities existing in a dictionary [19-23]. Lexicon-based techniques use mostly adjectives and adverbs to compute the overall sentiment score of a text, for instance, Linguistic Inquiry and Word Count (LIWC) [24], Affective Norms for English Words [25], and SentiWordNet [26]. Dictionaries of lexicons are created either manually or automatically [27,28]. First, a list is generated from a specific domain. Then synonyms and antonyms are added from other existing dictionaries such as WordNet [29]. More sophisticated lexicon-based methods focus on complex rules, such as regular expressions [30,31], instead of simply computing a sentiment score based on word polarities.

Machine learning–based techniques use statistical methods to compute sentiment polarity. The process involves training a classifier on a labeled data set, such as movie reviews or social media posts, and then using the model to predict the sentiment of new, unlabeled data. Obtaining labeled data to train the classifiers is a time-consuming task. Machine learning–based methods often face challenges when processing negative and intensifying statements and can have low performance when applied to different domains, as they rely mainly on the data set.
Sentiment analysis has become an increasingly popular technique in the health domain, as noted in the study by Rodríguez-Ibáñez et al [63]. A recent study [64] also found that the main data source for studies on health is social media, such as Twitter and Facebook. This is attributed to advancements in mobile technology and their use as a source in health-related topics, such as finding treatments, sharing experiences and opinions, and addressing public health surveillance issues [65-67]. During the pandemic, we witnessed social media becoming the main forum to express opinions related to COVID-19, which helped authorities to understand and monitor sentiments toward topics related to the pandemic [68-73].

Various studies have proposed new sentiment analysis methods and compared existing tools (e.g., TextBlob [74], VADER [12], and Stanza [13]) on topics related to COVID-19, mainly extracted from social media [6,16,17,75-78]. However, to the best of our knowledge, there are no studies that have compared several sentiment analysis tools on health-related surveys—a more structured type of text data than social media posts—that collected knowledge, beliefs, and habits during the COVID-19 pandemic [79-84]. The only study we are aware of that evaluates ChatGPT on various sentiment analysis tasks, comparing it with fine-tuned BERT, is the study by Wang at al [85]. The results demonstrated that ChatGPT exhibited promising zero-shot sentiment analysis ability, achieving performance on par with fine-tuned BERT and state-of-the-art models. However, it fell slightly behind domain-specific fully supervised state-of-the-art models.

**Methods**

This section presents the data sets used in this study along with our evaluation of sentence sentiment analysis methods, as illustrated in Figure 1. Specifically, we describe the (1) survey data sets, (2) state-of-the-art sentiment analysis tools, (3) few-shot learning with an LLM, and (4) zero-shot learning with ChatGPT.

Figure 1. Workflow of our study for evaluating sentence sentiment analysis using state-of-the-art sentiment analysis tools, few-shot learning with a large language model, and zero-shot learning with ChatGPT over health-related surveys. GPT: Generative Pre-trained Transformers; LIWC2015: Linguistic Inquiry and Word Count 2015; NIH: National Institutes of Health; OPT: Open Pre-Trained Transformers; VADER: Valence Aware Dictionary and Sentiment Reasoner.
Data

**NIH Data Set**

This data set was collected as part of a web-based survey assessing mental health during the pandemic, which started from April 2020 to May 2021. This was a sample of convenience, as participants were recruited from a poll of previous participants in the National Institute of Mental Health and National Center for Complementary and Alternative Medicine studies by advertising on social media and by flyers within the Washington metropolitan area. Participants who signed up completed various questionnaires at baseline, assessing demographics, clinical history, and psychological state [86]. The participants were then sent emails every 2 weeks for 6 months, inviting them to complete 3 of those questionnaires at that time. This latter survey consisted of 45 questions assessing various attitudes, behaviors, and impacts surrounding the pandemic and a single free-response question (“Is there anything else you would like to tell us that might be important that we did not ask about?”). There was a maximum of 13 potential survey (and free) responses per participant. Of the 3655 participants who enrolled in the study, 2497 (68.31%) responded at least once to the free-response item, yielding a total of 9738 item responses. These were composed of 26,411 sentences, which were the data used in this study. The semantic content of these responses (eg, main topics of concern over time) is available in the study by Weger et al [87].

**Stanford Data Set**

This data set was collected as part of a web-based survey conducted from March to September 2020 by a Stanford University team. The survey was conducted using a sample of convenience recruited through 3 social media platforms: Twitter, Facebook, and Nextdoor. They could participate by clicking on a survey link in the social media post upon seeing the recruitment materials. The survey comprised 21 questions including demographics and the impact of COVID-19 on individuals’ lives [88]. In this study, we focus on the evaluation of 3 free-text responses to the following questions: (1) “Although this is a challenging time, can you tell us about any positive effects or ‘silver linings’ you have experienced during this crisis?” (2) “What are the reasons you are not self-isolating more?” and (3) “Have you experienced any difficulties due to the coronavirus crisis?” Of the 4582 participants recruited, 3349 (73.09%) responded to at least 1 of the 3 free-text questions, resulting in a total of 7182 item responses. These were composed of approximately 21,266 sentences, which were the data used in this study. The topics and sentiments in these responses are reported in the study by Lossio-Ventura et al [10]. Table 1 presents additional details regarding the NIH and Stanford data sets.

| Table 1. Details of the National Institutes of Health (NIH) and Stanford data sets. |
|---------------------------------|------------|-------------|
|                          | NIH        | Stanford    |
| Start of the collection period | April 2020 | March 2020  |
| End of the collection period  | May 2021   | September 2020 |
| Responders, n/N (%)           | 2497/3655 (68.31) | 3349/4582 (73.09) |
| Response items, n             | 9738       | 7182        |
| Sentences before processing, n | 26,411     | 21,266      |
| Sentences after processing, n/N (%) | 26,188/26,411 (99.16) | 21,035/21,266 (98.91) |
| Tokens after processing, n    | 462,518    | 299,735     |
| Tokens per sentence, mean (SD)| 17.66 (11.11) | 14.25 (9.74) |

**Annotation**

We created training and test sets for both the NIH and Stanford data sets. These sets were derived from the surveys after completing the preprocessing steps and were used for training, tuning, and the official evaluation.

**Training Data Set**

We randomly selected 260 sentences, with 130 sentences from each data set. Each subset of 130 sentences was annotated by a different annotator. The annotators were instructed to assign a polarity value of −1 (negative), 0 (neutral), or 1 (positive) to each sentence.

**Test Data Set**

A total of 1000 sentences were randomly chosen, with 500 sentences selected from each data set [89]. Each set was annotated by 3 separate and independent annotators: A.1, A.2, and A.3 for NIH and A.4, A.5, and A.6 for Stanford. The annotators were instructed to assess the polarity of each sentence on a scale of −1 (negative), 0 (neutral), or 1 (positive). We used a 3-point scale to annotate the data. We then followed a 3-step procedure to determine the final labels, similar to that described in the studies by Nakov et al [90] and Rosenthal et al [91]. First, if all 3 annotators agreed on a label (full agreement), that label was accepted. Second, if 2 of the 3 agreed on a label (partial agreement), that label was also accepted. Third, if there was no agreement, the label was set as neutral (no agreement). Fleiss κ measure was calculated to assess the agreement between the 3 annotators of each test data set. The associated P values were computed to test if the agreement between annotators was substantially better than what would be expected by chance. Further details of the training and test data sets are provided in Table 2. Pearson correlation coefficients were also calculated to evaluate the degree of agreement between each pair of annotators, as shown in Figure 2.
Table 2. Details of the National Institutes of Health (NIH) and Stanford data sets.

<table>
<thead>
<tr>
<th></th>
<th>Training (n=130)</th>
<th>Test (n=500)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NIH</td>
<td>Stanford</td>
</tr>
<tr>
<td>Sentences, n (%)</td>
<td>130 (100)</td>
<td>130 (100)</td>
</tr>
<tr>
<td>Negative sentences, n (%)</td>
<td>71 (54.6)</td>
<td>45 (34.6)</td>
</tr>
<tr>
<td>Neutral sentences, n (%)</td>
<td>51 (39.2)</td>
<td>41 (31.6)</td>
</tr>
<tr>
<td>Positive sentences, n (%)</td>
<td>8 (6.2)</td>
<td>44 (33.8)</td>
</tr>
<tr>
<td>Full agreement, n (%)</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Partial agreement, n (%)</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>No agreement, n (%)</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Fleiss κ</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>P value</td>
<td>N/A</td>
<td>N/A</td>
</tr>
</tbody>
</table>

*aN/A: not applicable.

Figure 2. Correlation of annotators on the National Institutes of Health (NIH) and Stanford test data sets. A.1, A.2, A.3 represent the 3 independent NIH annotators, while A.4, A.5, A.6 represent the Stanford annotators.

Preprocessing
The survey responses contained personal identifiable information and multiple sentences covering different themes, for example, 2020 presidential elections and COVID-19 vaccines. Therefore, preprocessing steps included splitting responses into sentences, replacing people’s names, suppressing email addresses, and lemmatizing and converting text to lower case.

Sentiment Analysis Applications
We considered popular sentiment analysis applications available on the internet that use rules, machine learning, and fine-tuned LLMs.

Linguistic Inquiry and Word Count 2015
Linguistic Inquiry and Word Count 2015 (LIWC2015) [24,92,93] is a text analysis software that identifies and calculates the frequency of different categories of words in texts, such as pronouns, emotional words, cognitive words, and social words. LIWC2015 seeks to group words into categories that can be used to analyze psycholinguistic features in texts. Researchers in various fields, including psychology, sociology, and computer science, have used LIWC2015 to study a wide range of topics, such as personality, emotional expression, deception, and social interaction. LIWC2015 has also been used in various relevant studies on sentiment analysis. It provides with a summary variable “Tone” that combines positive and negative dimensions (posemo and negemo) into a single one. The higher the tone, the more positive it is. The tone ranges from 0 to 100. Numbers <50 indicate a more negative emotional tone. The default LIWC2015 Dictionary contains approximately 6400 words, word stems, and select emoticons.

SentiStrength
SentiStrength is a sentiment analysis tool that assigns scores to words and phrases based on their positive or negative sentiment [94-96]. It calculates an overall sentiment score for the text by combining these individual scores. This tool can provide dual-, binary-, trinary-, or single-scale results. In this study, a single scale ranging from −4 (extremely negative) to 4 (extremely positive) was chosen, with 0 indicating neutral sentiment. SentiStrength uses linguistic and lexicon-based methods. Linguistic methods involve rules and heuristics for identifying sentiment-bearing words and phrases, including cues such as repeated punctuation, emoticons, negations, and capital letters. The lexicon used consists of 2546 terms associated with polarity and intensity. Part of the lexicon was added from General Inquirer, including word roots such as “extrem*” to recognize variants. Training data sets included posts from various platforms such as BBC Forum, Twitter, YouTube, Digg.com, MySpace, and Runners World.

TextBlob
TextBlob is a Python library used in NLP tasks [11,74], such as part-of-speech tagging, sentiment analysis, and noun phrase extraction. TextBlob outputs a polarity score ranging from −1 to 1. A negative score signifies a negative sentiment, a positive
score indicates a positive sentiment, and a score of 0 represents a neutral sentiment. TextBlob includes 2 analysis approaches: a rule-based model and a supervised machine learning naive Bayes classifier model.

**VADER**

VADER [12,97] is a rule-based model designed for analyzing sentiment in social media text. It uses 5 rules based on grammatical and syntactical patterns to determine sentiment intensity. These rules involve punctuation, capitalization, degree modifiers, conjunctions such as “but,” and trigram evaluation to identify negations that can affect polarity. VADER was developed and validated using a gold standard list of lexical features, including LIWC, General Inquirer, and Affective Norms for English Words. The model was trained on various data sets, including tweets, New York Times opinions, movie reviews, and Amazon product reviews.

**Stanza**

Stanza is an open-source Python library that provides several methods for performing NLP tasks [13,98], including part-of-speech tagging, named entity recognition, dependency parsing, and sentiment analysis. Stanza’s sentiment analysis module assigns a positive, negative, or neutral sentiment score (0, 1, or 2, respectively) to each sentence in a given text. Stanza’s sentiment analysis tool is based on a convolutional neural network model using the vectors trained by Mikolov et al [99] on 100 billion words from Google News as well as a combination of lexical and syntactic features. It was trained on large data sets including movie reviews and the Stanford Sentiment Treebank. Unlike other methods, Stanza includes preprocessing of its own (sentence splitter and tokenizer).

**TweetEval**

TweetEval is a benchmarking platform for Twitter-specific classification tasks [100]. TweetEval consists of 7 NLP tasks: irony detection, offensive language detection, emoji prediction, emotion recognition, hate speech detection, stance detection, and sentiment analysis. Using TweetEval, a common set of evaluation metrics and data set, researchers and practitioners can compare the performance of different models on the same tasks and identify the most effective models for different NLP applications. TweetEval provides a leaderboard for ranking the performance of different models on the sentiment analysis task. The leaderboard is based on the F1-score. TweetEval returns 3 labels (positive, negative, and neutral) associated with a weight. TweetEval sentiment analysis is based on the RoBERTa model, an LLM based on BERT (trained on 58M tweets), and fine-tuned on the SemEval 2017 sentiment analysis data set (approximately 40,000 tweets) [91].

**Pysentimiento**

Pysentimiento is an open-source Python library that includes models for sentiment analysis and social NLP tasks, such as hate speech detection, irony detection, emotion analysis, named entity recognition, and part-of-speech tagging, in several languages such as English, Spanish, Portuguese, and Italian [101,102]. The English model for sentiment analysis is based on BERTweet [103], a RoBERTa model, trained on English tweets and also fine-tuned on the SemEval 2017 sentiment analysis data set [91]. Pysentimiento returns 3 polarity labels per text associated with a weight.

**NLPTown**

NLPTown [104] is a sentiment analysis application based on a BERT-base-multilingual-uncased model, fine-tuned for sentiment analysis on product reviews for 6 languages (English, Dutch, German, French, Spanish, and Italian), and predicts the sentiment of the review as the number of stars (1-5).

**Few-Shot Learning With Open Pre-Trained Transformers Language Models**

As mentioned previously, few-shot learning seeks to address the challenge of sentiment analysis when only a small amount of labeled data is available for training. In traditional supervised learning, models are trained on large data sets with many labeled examples. However, in some applications such as sentiment analysis, labeled survey data are scarce or expensive to obtain, making it difficult to train accurate models. In this study, we used the Open Pre-Trained Transformers (OPT) [105], a suite of decoder-only pre-trained transformers ranging from 125M to 175B parameters created by Meta AI. OPT has been used in several applications but has never been applied to sentiment analysis. This model has shown to perform similarly to the GPT-3 [60] on several NLP tasks. The OPT model was built using a data set of 180B tokens. This represents approximately 23% (180B/780B) of the amount of data set tokens used for the Pathways Language Model [52]. The largest OPT model has comparable number of parameters to GPT-3 (175B parameters) [60], although we used all models except for the latter given graphics processing unit limitations. The novelty of OPT is its availability as open source (albeit only for academic research).

**Zero-Shot Learning With ChatGPT**

Zero-shot learning refers to the use of a model to perform a task for which it has not been explicitly trained. Thus, zero-shot learning for sentiment analysis recognizes and classifies sentiment in text without being explicitly provided with examples of sentiment labels. Instead, the model is trained on related tasks, such as language modeling or machine translation, which enables it to understand the underlying structure of the language and the context in which it is used. In this study, we used ChatGPT (based on GPT-3.5), which has significantly improved the performance of several NLP tasks. GPT-3.5 is a model with 175B parameters created by OpenAI and trained on a vast amount of text data sourced from the internet using both reinforcement and supervised learning techniques. For this paper, we generated a polarity score for each sentence x by asking ChatGPT “What is the sentiment of the following sentence ‘x’?”

**Ethical Considerations**

The NIH survey was approved by the Institutional Review Board of the NIH (reference number 20MNI085), and all participants provided consent for the study. The Stanford survey was approved by Stanford’s Institutional Review Board (reference number 55436), and all participants provided consent for the study.
All survey data and responses in both the NIH and the Stanford data sets were anonymized and associated with a unique ID. Participants from both studies were not compensated for participating in the surveys.

Results

Evaluation Metrics
To assess the overall performance of the sentiment analysis tools, we evaluated the accuracy, macro $F$-measure, macro precision, and macro recall. Macro evaluation metrics were recommended in the NLP competition SemEval-2017 Task 4 [91].

Preparation of Applications for Evaluation
Harmonization of Applications’ Outputs
The LIWC2015, Stanza, and SentiStrength applications produce outputs that are measured on distinct scales. LIWC2015 generates a continuous value ranging from 0 to 100; SentiStrength generates an integer score ranging from $-4$ to 4; and Stanza produces a discrete whole number score of 0, 1, or 2, which correspond to negative, neutral, and positive sentiments, respectively. Therefore, it is necessary to convert these scores to a common range of $[-1, 1]$, as formally defined in equation 1.

$$
score'(x) = 2 \times \frac{score[x] - score[x]_{\text{min}}}{score[x]_{\text{max}} - score[x]_{\text{min}}} - 1\quad (1)$$

The distribution of sentiment scores across all tools is shown in Figure 3. We then classify all negative values as negative sentiment, all 0 values as neutral, and all positive values as positive sentiment. It is important to note that the VADER application uses a slightly different classification approach, considering a score $\leq 0.05$ to be negative, a score between $-0.05$ and 0.05 to be neutral, and a score $\geq 0.05$ to be positive.

Fine-Tuning for Few-Shot Learning
We used few-shot learning using our small amount of training data to fine-tune the OPT models, rather than training them from scratch. For this experiment, the training data set was split into 85% (110/130) for feeding the model and 15% (20/130) for validation. Given the memory constraints, we considered only OPT 125M, 350M, 1.3B, and 2.7B. We performed a hyperparameter search to optimize the performance of the model on sentiment analysis. We considered learning rate=$[3\times10^{-4}, 1\times10^{-4}, 3\times10^{-5},$ and $1\times10^{-5}]$, batch size=$[4, 8, 16,$ and $32]$ , number of epochs from 1 to 7, and the AdamW optimizer. The models that performed the best were OPT-1.3B and OPT-2.7B, using a learning rate of $1\times10^{-5}$, a batch size of 32, and 5 epochs. These were the models used to obtain the test set results reported in next subsections.

Experiment 1: Correlation Between the Outputs of Applications
The objective was to evaluate the agreement level among various methods for predicting the sentiments of COVID-19 survey responses. Understanding the methods’ agreement or divergence was crucial in determining the reliability and accuracy of predictions, allowing for accurate studies of the relationship between language use and mental health. The Pearson correlation coefficient was used to assess the reliability of the tools, as shown in Figure 4. Disagreement among the methods prompted us to evaluate few-shot learning to obtain high-quality predictions.
Figure 4. Pearson correlation matrix of score applications on the National Institutes of Health (NIH) and Stanford data sets. LIWC2015: Linguistic Inquiry and Word Count 2015; VADER: Valence Aware Dictionary and Sentiment Reasoner.

Experiment 2: Prediction of Sentiment Scores

Tables 3 and 4 show the performance results obtained by all applications, few-shot learning, and zero-shot learning techniques on the NIH and Stanford test data sets, respectively. Both test sets comprised 500 sentences each, as detailed in the Data section. The top 2 performance results are italicized. Of note, a perfect classifier that accurately categorizes all items obtains a value of 1, whereas a perverse classifier that misclassifies all items achieves a value of 0. However, a trivial classifier that assigns all sentences to the same category (positive, negative, or neutral) and a random classifier both have a value of 0.3333.

ChatGPT achieved a significant improvement in sentiment analysis compared with other models through zero-shot learning. On the NIH data set, ChatGPT outperformed few-shot learning (OPT-1.3B and OPT-2.7B) by 6% in accuracy and 7% in F-measure. Similarly, on the Stanford data set, ChatGPT showed better results than the OPT-1.3B and OPT-2.7B models, with 6% higher accuracy and 4% higher F-measure.

Moreover, to further evaluate the sentiment analysis tools, we used Bayesian analysis, as recommended by Benavoli et al [106], to assess the statistical significance of the performance of the methods. Specifically, we applied the Bayesian signed-rank test [107] to compare the accuracies achieved across multiple data sets. This test quantifies the likelihood of observing the signed ranks of accuracy differences under both the null hypothesis (indicating no significant difference) and alternative hypothesis (indicating a significant difference). The Bayesian signed-rank test is designed to compare performance over multiple data sets (≥2); therefore, we further partitioned the independent Stanford and NIH data sets. Each data set was partitioned into 3 subsets, based on the sentiment label assigned to them, resulting in positive, neutral, and negative subsets for each data set.

This division was influenced by insights from our prior analysis, which highlighted inherent distinctions among sentences associated with positive, neutral, and negative labels. For instance, positive sentences exhibited a preponderance of positive adjectives, whereas negative sentences featured more negative adjectives, and neutral sentences tended to emphasize facts that are characteristic of the neutral category. Therefore, we assumed a degree of independence across subsets within each data set. The heat map diagram in Figure 5 shows the results of our Bayesian analysis, with cells corresponding to row $i$ and column $j$. On the left side, “A higher than B” indicates the probability that method $i$ performs better than classifier $j$. The center indicates the probability of practical equivalence between methods $i$ and $j$. Similarly, on the right side, “B higher than A” indicates the probability that method $j$ is better than classifier $i$. These experiments confirmed that ChatGPT performed better than all the other alternatives. The OPT models showed similar performance to methods other than ChatGPT and could be considered as a viable second option.
### Table 3. Results on the National Institutes of Health (NIH) test data set.

<table>
<thead>
<tr>
<th>Application</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIWC2015</td>
<td>0.2733</td>
<td>0.5226</td>
<td>0.3587</td>
<td>0.4540</td>
</tr>
<tr>
<td>SentiStrength</td>
<td>0.5732</td>
<td>0.6006</td>
<td>0.5814</td>
<td>0.6480</td>
</tr>
<tr>
<td>TextBlob</td>
<td>0.4505</td>
<td>0.4776</td>
<td>0.4053</td>
<td>0.4340</td>
</tr>
<tr>
<td>VADER(^b)</td>
<td>0.6302</td>
<td>0.7036</td>
<td>0.6097</td>
<td>0.6580</td>
</tr>
<tr>
<td>Stanza</td>
<td>0.6178</td>
<td>0.5758</td>
<td>0.5886</td>
<td>0.6300</td>
</tr>
<tr>
<td>TweetEval</td>
<td>0.7818</td>
<td></td>
<td>0.8318</td>
<td>0.7898</td>
</tr>
<tr>
<td>Pysentimiento</td>
<td>0.7738</td>
<td>0.7780</td>
<td>0.7699</td>
<td>0.7760</td>
</tr>
<tr>
<td>NLPtown</td>
<td>0.4338</td>
<td>0.5173</td>
<td>0.4210</td>
<td>0.4520</td>
</tr>
<tr>
<td>OPT(^c) 1.3B (few-shot)</td>
<td>0.8032</td>
<td>0.8000</td>
<td>0.7992</td>
<td>0.8000</td>
</tr>
<tr>
<td>OPT 2.7B (few-shot)</td>
<td>0.8061</td>
<td>0.8040</td>
<td>0.8050</td>
<td>0.8040</td>
</tr>
<tr>
<td>ChatGPT (zero-shot)</td>
<td>0.8526</td>
<td>0.8926</td>
<td>0.8668</td>
<td>0.8600</td>
</tr>
<tr>
<td>All negative</td>
<td>0.1487</td>
<td>0.3333</td>
<td>0.2056</td>
<td>0.4460</td>
</tr>
<tr>
<td>All neutral</td>
<td>0.1547</td>
<td>0.3333</td>
<td>0.2113</td>
<td>0.4640</td>
</tr>
<tr>
<td>All positive</td>
<td>0.0300</td>
<td>0.3333</td>
<td>0.0550</td>
<td>0.0900</td>
</tr>
</tbody>
</table>

\(^a\)LIWC2015: Linguistic Inquiry and Word Count 2015.
\(^b\)VADER: Valence Aware Dictionary and Sentiment Reasoner.
\(^c\)Italicization represents the top 2 performance results.
\(^d\)OPT: Open Pre-Trained Transformers.

### Table 4. Results on Stanford test data set.

<table>
<thead>
<tr>
<th>Application</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIWC2015</td>
<td>0.3752</td>
<td>0.4391</td>
<td>0.3890</td>
<td>0.5400</td>
</tr>
<tr>
<td>SentiStrength</td>
<td>0.5738</td>
<td>0.5561</td>
<td>0.5335</td>
<td>0.5420</td>
</tr>
<tr>
<td>TextBlob</td>
<td>0.4757</td>
<td>0.4872</td>
<td>0.4527</td>
<td>0.4600</td>
</tr>
<tr>
<td>VADER(^b)</td>
<td>0.5875</td>
<td>0.5919</td>
<td>0.5755</td>
<td>0.5840</td>
</tr>
<tr>
<td>Stanza</td>
<td>0.5975</td>
<td>0.4987</td>
<td>0.4859</td>
<td>0.5040</td>
</tr>
<tr>
<td>TweetEval</td>
<td>0.7366</td>
<td>0.7178</td>
<td>0.7090</td>
<td>0.7200</td>
</tr>
<tr>
<td>Pysentimiento</td>
<td>0.6731</td>
<td>0.6362</td>
<td>0.6267</td>
<td>0.6440</td>
</tr>
<tr>
<td>NLPtown</td>
<td>0.5163</td>
<td>0.5192</td>
<td>0.5056</td>
<td>0.5420</td>
</tr>
<tr>
<td>OPT(^c) 1.3B (few-shot)</td>
<td>0.8323</td>
<td>0.8160</td>
<td>0.8211</td>
<td>0.8160</td>
</tr>
<tr>
<td>OPT 2.7B (few-shot)</td>
<td>0.8288</td>
<td>0.8100</td>
<td>0.8147</td>
<td>0.8100</td>
</tr>
<tr>
<td>ChatGPT (zero-shot)</td>
<td>0.8632</td>
<td>0.8779</td>
<td>0.8662</td>
<td>0.8740</td>
</tr>
<tr>
<td>All negative</td>
<td>0.1560</td>
<td>0.3333</td>
<td>0.2125</td>
<td>0.4680</td>
</tr>
<tr>
<td>All neutral</td>
<td>0.0780</td>
<td>0.3333</td>
<td>0.1264</td>
<td>0.2340</td>
</tr>
<tr>
<td>All positive</td>
<td>0.0993</td>
<td>0.3333</td>
<td>0.1531</td>
<td>0.2980</td>
</tr>
</tbody>
</table>

\(^a\)LIWC2015: Linguistic Inquiry and Word Count 2015.
\(^b\)VADER: Valence Aware Dictionary and Sentiment Reasoner.
\(^c\)OPT: Open Pre-Trained Transformers.
\(^d\)Italicization represents the top 2 performance results.
Discussion

Principal Findings

Our primary objective was to assess various sentiment analysis tools for the purposes of predicting the sentiments of survey responses during the COVID-19 pandemic. Obtaining a thorough understanding of the tools’ degree of agreement, as shown in Figure 4, was crucial for determining whether they could be used as surrogates for human labeling. The disagreement between tools led us to try ensemble methods to produce more reliable ratings. Fine-tuned BERT models such as TweetEval and Psentimiento outperformed other baseline methods. Fine-tuned methods have the ability to learn domain-specific patterns from text, resulting in better performance than lexicon- and rule-based methods. However, these techniques often require large training data sets to achieve optimal performance, such as the 40k tweet data set used to train TweetEval and Psentimiento.

As part of the process of determining agreement between tools, we labeled a small data set (260 sentences), which is what prompted us to consider the possibility of using few-shot and zero-shot learning techniques. We then investigated the performance of OPT, which is unexplored in sentiment analysis, for few-shot learning using a small training data set (260 sentences). The OPT-1.3B and OPT-2.7B models surpassed all the baseline methods as well as the fine-tuned BERT models. This highlighted the potential of few-shot learning in dealing with scarce annotated data and the effectiveness of few-shot learning. Although better results could have been achieved with a larger training set, these experiments primarily aimed to investigate the potential of OPT using limited annotated data. The potential is to be able to produce models tailored to specific research applications, with only a small time investment by domain experts. We believe that these models can significantly contribute to the sentiment analysis of health- and clinical-related surveys and can be further fine-tuned with additional data and optimized hyperparameters.

Our investigation also encompassed zero-shot learning with ChatGPT, which exhibited remarkable performance compared with all other models, including few-shot learning with OPT, as presented in Tables 3 and 4. Note that GPT-3.5—the model behind ChatGPT—is trained on related tasks, such as language modeling or machine translation. This enabled it to understand the underlying structure of sentiment-related language and the context in which it is used. Moreover, the necessity for manual text annotations in sentiment analysis tasks makes ChatGPT and other LLMs particularly attractive. As demonstrated by Ziems et al [108], LLMs can alleviate the workload of human annotators in a zero-shot manner, thereby enhancing the efficiency of social-science analysis. In addition, a study [109] found that ChatGPT outperformed crowd workers in various text annotation tasks, including assessing relevance, stance, topics, and frame detection. These findings suggest that there may be potential in using ChatGPT and other recent LLMs for annotation in clinical NLP and reserving human input for quality control. Sentiment analysis tools based on LLMs, such as ChatGPT, automatically identify relevant features, reducing the need for manual engineering, which is a common requirement in tools such as LIWC 2015 and VADER. In addition, LLMs enable fine-tuning, allowing for potential adaptation to different sentiment analysis tasks (eg, in new domains) without the need for complete retraining. LLM-based tools can also capture longer-range context for more accurate sentiment assessment.

Limitations

There exist several limitations and risks of ChatGPT and other non–open-source LLMs regarding protected health information (PHI). Non–open-source LLMs require sending information to an external server and do not provide transparency into how they handle PHI, making it difficult to assess how the model is processing and protecting sensitive information. They may also have security vulnerabilities that can be exploited to gain unauthorized access to PHI. Note also that LLMs are not specifically designed for sentiment analysis, which may sometimes lead to errors, for instance, subtle sarcasm such as “Oh yes, great job!,” context-dependent negation as in “The vaccine was not as bad as I thought,” and idiomatic expressions such as “It’s a piece of cake.” They may encounter difficulties with nuanced health-related terminology and concepts. Therefore, specialized health terminology may require additional adaptation beyond general text fine-tuning, for instance, medical abbreviations and acronyms such as “The patient teared up because of a significant increase in their CD4 count” and “So, my mom’s HbA1c levels have improved after insulin therapy.” In addition, although several outputs may sound plausible, they may occasionally be incorrect. In our view, the output of LLMs should not be used without a plan for human quality control (eg, via sampling) or mitigation (eg, repeated validation). This is crucial for ensuring the accuracy and reliability of the generated content, as LLMs may produce results that require refinement or correction before dissemination. Moreover, there are constraints on the ability to access ChatGPT via its application programming interface, and this may make it too
costly or time-consuming to do so. Therefore, researchers and
health care practitioners might also opt to use an open-source
language model for their NLP-related projects, such as OPT,
which can be run on site and perform well on sentiment analysis.

Finally, our study focused on using surveys to understand
people’s feelings, specifically regarding COVID-19, which was
a very important topic at the time. Thus, our conclusions apply
specifically to discussions about COVID-19 and may not be
true for other subjects. In addition, it is important to highlight
that the Stanford data set has an implicit polarity bias: it
specifically asks for positive effects (“Although this is a
challenging time, can you tell us about any positive effects or
‘silver linings’ you have experienced during this crisis?”) and
difficulties (“Have you experienced any difficulties due to the
coronavirus crisis?”). The NIH data set poses a single,
less-biased question. Therefore, it is crucial to be careful when
generalizing our findings beyond the scope of COVID-19 during
the studied time frame.

Acknowledgments
The authors would like to thank Kelsey Yeh and Nathaniel Menon, who worked as research assistants in the Social Media Lab
at Stanford University, for their assistance in annotating the Stanford test data set. The authors would also like to express gratitude
to the many respondents who shared their experiences in both studies. The research reported in this publication was supported
in part by the Intramural Research Programs of the National Institute of Mental Health (ZIAMH002922: Principal Investigator
JC and ZIC-MH002968: Principal Investigator FP), National Center for Complementary and Integrative Health (ZIA-AT000035:
Principal Investigator LA), and National Institute of Arthritis and Musculoskeletal and Skin Diseases (K24AR075060: Principal
Investigator EL). The content is solely the responsibility of the authors and does not necessarily represent the official views of
the National Institutes of Health.

Data Availability
The test data sets generated and analyzed during this study are deidentified and freely available in the FigShare repository [89].
The source code for fine-tuning the OPT models and using ChatGPT in the experiments conducted in this study is publicly
accessible on GitHub [110].

Authors’ Contributions
JALV and FP contributed to conceiving the study idea and design. LA and JC led the collection of the National Institutes of
Health (NIH) data set, whereas EL led the collection of the Stanford data set. The annotation of the training NIH and Stanford
data sets was conducted by RW and AYL, respectively. The annotation of the NIH test data set was conducted by LA, RW, and
EFG, whereas AYL and 2 research assistants annotated the Stanford test data set. JALV set up the applications and performed
the evaluation. JALV and FP wrote the initial draft and revised the subsequent versions. All authors read, revised, and approved
the final manuscript.

Conflicts of Interest
None declared.

References
   [doi: 10.1145/2436256.2436274]
   10.2196/22817] [Medline: 32897868]


56. Büyükoz B, Hürriyetoglu A, Özgür A. Analyzing ELMo and DistilBERT on socio-political news classification. 2020 Presented at: Workshop on Automated Extraction of Socio-political Events from News 2020; May 12, 2020; Marseille, France URL: https://aclanthology.org/2020.aespen-1.4


96. SentiStrength- sentiment strength detection in short texts. SentiStrength. URL: http://sentistrength.wlv.ac.uk/ [accessed 2022-03-15]


99. Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed representations of words and phrases and their compositionality. arXiv Preprint posted online on October 16, 2013. [FREE Full text]


Abbreviations

BERT: Bidirectional Encoder Representations from Transformers
DNN: deep neural network
ELMo: Embeddings from Language Models
GPT: Generative Pre-trained Transformers
LIWC: Linguistic Inquiry and Word Count
LIWC2015: Linguistic Inquiry and Word Count 2015
LLM: large language model
NIH: National Institutes of Health
NLP: natural language processing
OPT: Open Pre-Trained Transformers
PaLM: Pathways Language Model
PHI: protected health information
RoBERTa: Robustly optimized Bidirectional Encoder Representations from Transformers approach
VADER: Valence Aware Dictionary and Sentiment Reasoner

©Juan Antonio Lossio-Ventura, Rachel Weger, Angela Y Lee, Emily P Guinee, Joyce Chung, Lauren Atlas, Eleni Linos, Francisco Pereira. Originally published in JMIR Mental Health (https://mental.jmir.org), 25.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Mental Health, is properly cited. The complete bibliographic information, a link to the original publication on https://mental.jmir.org/, as well as this copyright and license information must be included.
Patient Satisfaction With a Coach-Guided, Technology-Based Mental Health Treatment: Qualitative Interview Study and Theme Analysis

Background: Technology-based mental health interventions address barriers rural veterans face in accessing care, including provider scarcity and distance from the hospital or clinic. webSTAIR is a 10-module, web-based treatment based on Skills Training in Affective and Interpersonal Regulation, designed to treat posttraumatic stress disorder and depression in individuals exposed to trauma. Previous work has demonstrated that webSTAIR is acceptable to participants and effective at reducing symptoms of posttraumatic stress disorder and depression when delivered synchronously or asynchronously (over 5 or 10 sessions).

Objective: This study explored factors that lead to greater patient satisfaction with webSTAIR, a web-based, coach-guided intervention.

Methods: We analyzed qualitative interview data to identify themes related to patient satisfaction with webSTAIR delivered with synchronous video-based coaching.

Results: Four themes emerged from the data: (1) coaching provides accountability and support, (2) self-pacing offers value that meets individual needs, (3) participants like the comfort and convenience of the web-based format, and (4) technical issues were common but not insurmountable.

Conclusions: We conclude that participants valued the accountability, flexibility, and convenience of tech-based interventions with video-delivered coaching.

(JMIR Ment Health 2024;11:e50977) doi:10.2196/50977

KEYWORDS
coaching; digital treatment; interview; mental health; patient satisfaction; PTSD; qualitative assessment; qualitative methods; sentiment analysis; technology-based; telehealth; trauma; veterans; video telehealth; web-based treatment

Corresponding Author: Ashley Helm Smith, MA
Houston Veteran Affairs Health Services Research and Development Center for Innovations in Quality, Effectiveness and Safety
Veteran Affairs Medical Center, Houston, TX, United States

Phone: 1 713 794 8601
Email: ashley.smith30@va.gov

( JMIR Ment Health 2024;11:e50977 )
(page number not for citation purposes)
Introduction

More than 1.7 million US veterans receive mental health care through the Veterans Health Administration (VHA) annually [1]. However, US veterans underestimate mental health care and often experience barriers that impact access, such as provider scarcity [2], distance to clinic [3], inability to take off from work or school, and dependent care responsibilities [4-6]. Veterans living in rural areas make up about 24% of all veterans in the United States [7], and they have similar mental health care needs as urban veterans; however, rural veterans experience access barriers at higher rates [8,9]. Technology-based mental health interventions are a potential solution to address some of the barriers that rural veterans in the United States face.

Technology-based mental health interventions can be accessed from a smartphone or tablet. They range from fully automated to therapist-guided in person or by video, telephone, or SMS text messages [10-13]. Technology-based mental health interventions have demonstrated effectiveness in clinical trials, with moderate-to-high effect sizes [14-17]. However, satisfaction with a treatment plays a significant role in treatment effectiveness. Patient satisfaction with treatment is linked to greater adherence [18,19] and greater perceived improvement in clinical status [20,21]. Even when adherence to treatment is controlled for, satisfied patients benefit more from care than less satisfied patients [22,23]. In mental health settings, a strong therapeutic alliance is linked to greater patient satisfaction [24]. Further understanding of factors that contribute to patient satisfaction is critical to designing future tech-based interventions that demonstrate better engagement and retention and greater clinical benefit [25,26].

Studying satisfaction with mental health treatments is complex. There are a number of variables that can influence patient satisfaction with mental health treatment, such as patient, disease, provider, therapy, and environmental factors (eg, technology-mediated treatments) [22,27,28]. Although many teams evaluate overall satisfaction by collecting survey data and compiling ratings to quantify patients’ levels of satisfaction, qualitative data provides a nuanced and contextualized understanding of satisfaction [22]. In 2 recent studies, our team found that having therapist support and the convenience of being able to access treatment on the web contributed to patient satisfaction with technology-based mental health interventions [29,30]. This study uses qualitative interviews with patients who completed an internet-based mental health treatment to explore factors that lead to greater patient satisfaction in such a program. This study builds on previous findings by exploring the benefits of using a coach and the ways web-based formats provide comfort and convenience, as well as increased access in some cases.

Methods

Study Design

webSTAIR is an internet-based intervention adapted from Skills Training in Affective and Interpersonal Regulation (STAIR), a treatment for trauma-exposed individuals with symptoms of posttraumatic stress disorder (PTSD) and depression [31]. The program was developed to engage rural women veterans, who have been found to be underrepresented in mental health services [32]. However, male veterans were also enrolled if they were interested and met the study criteria. webSTAIR offers 10 web-based modules of self-directed skills training on emotion regulation (eg, emotional awareness, emotional management, and distress tolerance) and interpersonal skills (eg, assertiveness, flexibility, and compassion for oneself and others).

In addition to the asynchronously delivered web-based content, participants were offered 5 or 10 synchronous video coaching sessions, which took place using video-telehealth, after the completion of 1 or 2 modules. Coaches were licensed Veterans Affairs (VA) mental health providers who underwent training in the intervention and attended weekly supervision with a certified STAIR trainer. Sessions with the coach lasted approximately 45-50 minutes and involved reviewing module content, discussing the applicability of concepts in the veteran’s life, and strategizing on how to integrate skills into daily practice. The webSTAIR program was delivered as routine care in mental health outpatient clinics at 9 sites serving rural patients within the VHA. The qualitative data presented were collected as part of a naturalistic evaluation of the program.

Recruitment and Sample

Participants in the webSTAIR program were US veterans recruited from 9 VHA facilities across the country that serve rural patients. Referrals were either clinician- or self-initiated. Participants completed the program between September 2018 and March 2020. Eligibility for the study was determined based on an initial telephone screening. Patients were considered eligible if they reported a history of trauma exposure and screened positive on the Primary Care PTSD Screen (PC-PTSD; positive on 3 of 5 items [33]) or the 2-item Patient Health Questionnaire (PHQ-2; positive on 1 of 2 items [34]), indicating clinically significant symptoms of PTSD or depression. The following resulted in ineligibility for enrollment: cognitive impairment or psychosis, mania, primary drug or alcohol abuse, current domestic violence, concurrent trauma-focused treatment, residential care for PTSD within 1 year, or an inability to attend telemental health appointments by video.

Measures and Data Collection

Demographic information for all patients in the program was collected at the initial intake phase of the study. The qualitative data for this study were collected during a 1-time interview conducted following the completion of the webSTAIR program. Patients who completed the program were sent a letter explaining the purpose of the interview and study details. This letter also included a contact number that patients could call to decline participation. All patients who did not actively call to decline were then contacted by study staff and asked if they would like to participate. Study staff made at least 3 attempts to contact participants. Outreach resulted in a response rate of 30%, with 74 participants agreeing to participate. The demographic information of interviewees was linked using a study ID number.

The interview guide was piloted and refined during the pilot intervention phase of the study [30,35]. It contained both
categorical and open-ended questions. One member of the study staff conducted the interview, while 1-2 staff members took detailed notes during the interview. While the study team did not record or transcribe interviews, note-takers were often able to capture responses nearly verbatim. Quotes in this paper are drawn from these notes.

Participants were asked questions about their experience working through the program and about their satisfaction with the intervention. Interviews were conducted by telephone by study staff, including the first author (AHS) and second author (HT), as well as other members of the study team (MW). Interviews lasted approximately 60 minutes. At the conclusion of the interview, the interview notes were consolidated, and the finalized interview notes were entered into a database. Interview protocols were reviewed by the principal investigator’s affiliated institutions’ institutional review board, designated as “quality improvement.”

Data Analysis

Interview data were analyzed using strategies from qualitative content analysis and thematic analysis [36-39]. The interview notes for each participant were entered into a matrix in Microsoft Excel (Microsoft Corp) and organized by interview question. A list of the interview questions used in the analysis for this study can be found in Multimedia Appendix 1. Data coding and analysis were conducted by a master’s-level researcher with extensive qualitative experience who also functioned as a research assistant on the study (AHS) and a postdoctoral nurse trained in qualitative methods (HT). A codebook was developed based on interview guide questions and refined based on emergent findings. The second author (HT) reviewed the codes assigned by the first author, either agreeing or disagreeing. Coders met to resolve any disagreements until consensus was reached. The themes were analyzed and further defined to describe participants’ satisfaction with the intervention. The data are presented here by satisfaction with the web-based, coach-guided format of webSTAIR.

Ethical Considerations

Study procedures were reviewed and found to be exempt by the institutional review board for Baylor College of Medicine and Affiliated Hospitals. Data were collected under a quality improvement designation. Participant confidentiality was upheld with the use of a randomized participant ID number.

Results

Sample

The exit interview was completed by 74 participants, the majority of whom were rural (46/74, 62%), White or Caucasian (48/74, 65%), female (42/74, 57%), aged between 35 and 44 years (25/74, 34%), and had some college or a 2-year college degree (32/74, 44%). A detailed list of participant characteristics can be found in Table 1.
### Participant demographics.

<table>
<thead>
<tr>
<th>Demographics</th>
<th>Coach 5 (n=43), n (%)</th>
<th>Coach 10 (n=31), n (%)</th>
<th>Total (N=74), n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sex</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>14 (33)</td>
<td>17 (55)</td>
<td>31 (42)</td>
</tr>
<tr>
<td>Female</td>
<td>29 (67)</td>
<td>13 (42)</td>
<td>42 (57)</td>
</tr>
<tr>
<td>Transgender</td>
<td>0 (0)</td>
<td>1 (3)</td>
<td>1 (1)</td>
</tr>
<tr>
<td><strong>Age (years)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25-34</td>
<td>8 (19)</td>
<td>5 (16)</td>
<td>13 (18)</td>
</tr>
<tr>
<td>35-44</td>
<td>16 (37)</td>
<td>9 (29)</td>
<td>25 (34)</td>
</tr>
<tr>
<td>45-54</td>
<td>8 (19)</td>
<td>7 (23)</td>
<td>15 (20)</td>
</tr>
<tr>
<td>55-64</td>
<td>9 (21)</td>
<td>7 (23)</td>
<td>16 (22)</td>
</tr>
<tr>
<td>65-74</td>
<td>2 (4)</td>
<td>3 (9)</td>
<td>5 (6)</td>
</tr>
<tr>
<td><strong>Rurality</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Urban</td>
<td>16 (37)</td>
<td>12 (39)</td>
<td>28 (38)</td>
</tr>
<tr>
<td>Rural</td>
<td>27 (63)</td>
<td>19 (61)</td>
<td>46 (62)</td>
</tr>
<tr>
<td><strong>Race</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White or Caucasian</td>
<td>24 (56)</td>
<td>24 (78)</td>
<td>48 (65)</td>
</tr>
<tr>
<td>Black or African American</td>
<td>6 (14)</td>
<td>2 (6)</td>
<td>8 (11)</td>
</tr>
<tr>
<td>Hispanic or Latino</td>
<td>4 (9)</td>
<td>2 (6)</td>
<td>6 (8)</td>
</tr>
<tr>
<td>American Indian</td>
<td>1 (2)</td>
<td>1 (4)</td>
<td>2 (2)</td>
</tr>
<tr>
<td>Mixed race or ethnicity</td>
<td>8 (19)</td>
<td>2 (6)</td>
<td>10 (14)</td>
</tr>
<tr>
<td><strong>Employment status</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Full-time</td>
<td>12 (28)</td>
<td>13 (42)</td>
<td>25 (34)</td>
</tr>
<tr>
<td>Part-time</td>
<td>6 (13)</td>
<td>4 (13)</td>
<td>10 (13)</td>
</tr>
<tr>
<td>Not currently working for pay</td>
<td>11 (26)</td>
<td>5 (16)</td>
<td>16 (22)</td>
</tr>
<tr>
<td>Retired</td>
<td>14 (33)</td>
<td>9 (29)</td>
<td>23 (31)</td>
</tr>
<tr>
<td><strong>Educational level</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Some high school</td>
<td>1 (2)</td>
<td>0 (0)</td>
<td>1 (1)</td>
</tr>
<tr>
<td>Earned high school degree</td>
<td>4 (9)</td>
<td>4 (13)</td>
<td>8 (11)</td>
</tr>
<tr>
<td>Some college or 2-year degree</td>
<td>20 (47)</td>
<td>12 (39)</td>
<td>32 (44)</td>
</tr>
<tr>
<td>Earned 4-year degree</td>
<td>14 (33)</td>
<td>10 (32)</td>
<td>24 (32)</td>
</tr>
<tr>
<td>Postgraduate</td>
<td>3 (7)</td>
<td>5 (16)</td>
<td>8 (11)</td>
</tr>
<tr>
<td>Missing</td>
<td>1 (2)</td>
<td>0 (0)</td>
<td>1 (1)</td>
</tr>
</tbody>
</table>

### Themes

Participants responded to categorical questions that they were generally satisfied with the webSTAIR program; that is, the majority of interview participants felt that webSTAIR met their needs and that they would use a similar program again in the future. Four themes emerged from the open-ended responses and are detailed below: (1) coaching provides accountability and support, (2) self-pacing offers value that meets individual needs, (3) participants like the comfort and convenience of the web-based format, and (4) technical issues were common but not insurmountable. **Table 2** presents quotations exemplifying each theme.
Table 2. Domains and illustrative quotes.

<table>
<thead>
<tr>
<th>Themes</th>
<th>Illustrative quotes</th>
</tr>
</thead>
</table>
| Theme 1: coaching provides accountability and support       | • Because she was so good at bringing back the tools you get out of webSTAIR. Website tells you to apply it, but it's nice to hear it differently and she brings it back to my situation. That was really helpful for me. (35-44–year-old White female)  
  • Seemed like sessions were the practice/reiteration part. She really provided the application, tying it together and making it stick. Would stress this a lot. It really helped and challenged me. (35-44–year-old White male) |
| Theme 2: self-pacing offers value that meets individual needs | • It allowed me to go at my own pace. I had time to consider my answers and think about what I wanted to say. (55-64–year-old White male)  
  • When you're meeting in person, you often are bobble heading even if you're not getting what they're telling you. Reading for yourself and being able to go over it and process… gives you time to process. (55-64–year-old mixed-race female)  
  • Some lessons needed more time to practice and to mentally digest. I'm still going over them because I printed them out, but I feel like I needed more time… (45-54–year-old White female) |
| Theme 3: participants like the comfort and convenience of the web-based format | • I work, helped me not be stressed about getting appointments without taking off work. (25-34–year-old African American female)  
  • With PTSD, I don’t sleep well. To not have to get up and drive. Closest facility is 45 min-1 hour away… To not have to drive is great. I live in the country on purpose. I wouldn’t go to therapy without webSTAIR. (35-44–year-old mixed-race female)  
  • I liked that it was online and for whatever reason felt like a safer environment than sitting face to face with someone. (25-34–year-old White female)  
  • It's easy to get complacent too, catch 22 at your home. Easy to get distracted, too comfortable at home so easily distracted. (25-34–year-old White male) |
| Theme 4: technical issues are common but not insurmountable  | • If you're working on something and push the back button you delete everything you did. You have to start over. (35-44–year-old African American female)  
  • I didn't like the fact the info I wrote down in the modules, therapist couldn't see, so when we were reviewing, I would have to go back in and start from the beginning. It would be helpful if she could see what I wrote. (35-44–year-old African American female)  
  • Sometimes couldn't hear each other or it would freeze. Normal facetime issues. More issues with computer, iPad or iPhone crystal clear, perfect. (35-44–year-old American Indian or Alaskan Native female) |

**Theme 1: Coaching Provides Accountability and Support**

The presence of regular check-ins with a coach kept participants accountable and motivated them to do the work and stay on track by providing guidance as to when they should complete the next module. Only 1 participant noted that the regular check-ins with their coach and expectations about content completion each week felt like too much pressure, but this was uncommonly reported. Most respondents indicated they appreciated check-ins.

_I knew that we were going to have our sessions. So it helped with accountability – like actually doing the work because we’re going to talk about it. And she keeps me on track because she tells me this is when you’re going to do the next module._ [35-44-year-old Black or African-American female]

Coaches provided emotional support and helped participants understand and apply the webSTAIR content to their own lives. Participants often noted a good rapport with their coach, and many would have liked to talk to them more simply because they enjoyed talking with them. As a mixed-race male in his early 30s commented, “She was awesome. She always remembered what was going on with me. She helped with materials.”

The majority of participants were satisfied with the number of sessions with their coach, regardless of the number of coaching sessions they received. Those that did express a preference for more sessions cited a variety of reasons, most commonly the desire for more emotional support and clarification about the content. Participants perceived coaching sessions as a time to process their feelings and better understand how content applied to their individual situation. The coaching sessions were considered an essential component of webSTAIR and were highly valued.

**Theme 2: Self-Pacing Offers Value That Meets Individual Needs**

Participants generally liked the format of webSTAIR, which allowed them to take their time and interact with and respond to the web-based content. Having time between reviewing the material and meeting with their coach helped participants feel less pressure and less rushed than standard face-to-face psychotherapy using evidence-based psychotherapy protocols, as in “No pressure to feel like you get an answer right or wrong or in a hurry to get it done” (35-44–year-old White male). They were able to reflect on and practice the skills webSTAIR aimed to teach them. In many cases, this led to more thoughtful reactions to the content, which informed meetings with their coach and made them more productive. A White male in his mid-forties commented:
In contrast to those who liked the self-paced format, a small number of participants reported they would have liked the pace to be more personalized or more self-paced. For example, a few participants would have liked to speed through some of the content that they felt they did not need or take more time (e.g., more time in between video coaching sessions) to go deeper into the content they thought they needed most without feeling rushed. However, the desire for the pace to be more personalized did not impact their satisfaction with the program.

A minority of participants noted difficulty understanding or remembering the content between sessions.

Some participants successfully prepared for coaching sessions by taking detailed notes about the content and their questions, or by reviewing the webSTAIR content immediately before meeting with their coach. This strategy was acceptable for some webSTAIR participants, but others found it tedious or “annoying” to have to refresh between sessions to avoid forgetting the content, though this was uncommonly reported.

**Theme 3: Comfort and Convenience of the Web-Based Format Made it Easier to Access Care**

Participants liked the convenience of the coach-guided, web-based format, which allowed flexible scheduling and reduced travel time. The ability to meet on the web allowed veterans to engage in care they may not have otherwise engaged in due to distance or other logistical factors, such as difficulty scheduling around work or childcare obligations. A Hispanic male in his late 30s explained how and why the web-based format was more convenient:

> Able to do it from home. I cannot take advantage of many mental health services at the VA; VA is 30-40 minutes from my home. I was glad I was able to take advantage of this.

In addition to the logistical benefits, participants liked that they did not have to confront crowded waiting rooms at their facility. Instead, they could participate in the program from home, where they felt most comfortable and, in some cases, where they felt most safe. A minority of participants found it more difficult to find privacy in their own homes, away from the people they lived with, and others found themselves more prone to distractions at home. According to 1 participant, a White female in her mid-forties, “...the same thing that made it nice also made it not so nice.” However, privacy issues or distractions at home were not commonly reported.

**Theme 4: Technical Issues Are Common but not Insurmountable**

The majority of participants experienced some technical challenges with the video coaching sessions, and almost half experienced difficulty with the website. Participants spontaneously reported that about a third of all the technical challenges they experienced occurred only 1 or 2 times throughout the entire program. For issues that did not resolve on their own, participants were able to troubleshoot, either independently or with their coach, to complete their sessions and web-based content. In cases where coaches were able to help solve problems or troubleshoot, their guidance seemed to ameliorate the negative impact on patient satisfaction.

[Had] difficulties quite a few times, but she always made it work. [35–44-year-old Hispanic female]

Technical issues included being unable to hear the other person, issues with the link to the video session, and difficulty with the video freezing or closing unexpectedly. Sometimes the issue resolved quickly on its own and occurred infrequently. Other times, the issue could be attributed to a recent update or a setting that needed to be changed on their device. Strategies that were frequently used by coaches to help troubleshoot included pivoting to the phone for audio, sending different pieces of equipment to use, and offering emotional support that someone would help them resolve the issue. A White female in her mid-forties said:

> One time we [experienced video issues]; [we] just turned off the audio, [and] video was still there and [we] used phone for audio. Work around; it worked just fine.

Only 1 participant spontaneously reported having to reschedule a session with their coach due to connection issues. More often, participants shared that they were still able to meet on the web with their coach despite issues with the video connection, in part because of their coach’s persistence.

It didn’t connect twice but she worked hard to get it connected, so we still had our sessions. [35–44-year-old African American female]

Participants reported a wide range of issues with the website, including trouble accessing the tools or videos, glitches such as issues with the “back” or “next” button, difficulty with the website kicking them out or freezing up, and issues with the equipment that made accessing all the content from the modules difficult. Several participants spontaneously reported that they resolved the issue with the website through independent troubleshooting, usually by closing and reopening the browser: “Sometimes the role play [exercise] wasn’t working right. Closed out the browser and then it would work fine” (35-44–year-old Hispanic female). In at least 1 case, an issue with the website was resolved when the coach sent the participant a VA iPad (Apple Inc). Issues that were most easily resolved included the website or web content freezing up and issues with the “back” button not working.

More persistent issues with the website included typed responses not saving, being unable to access program materials, such as worksheets or role plays, and difficulty printing. There was only
Previous studies of technology-based mental health interventions suggest that the most frequently reported reason for dissatisfaction is related to intervention pacing that moves too quickly [45]. Participants in this study often noted the benefit of the self-paced format of webSTAIR, as it allows more time and space to reflect on, practice, and absorb the content compared to standard, synchronous, evidence-based psychotherapy sessions in which participants may feel pressure to respond and process material in the moment. Additional time to review the material between professionally guided sessions contributed to the considerable satisfaction that we have observed in this study.

Participants in this study generally appreciated the comfort and convenience of the web-based format. Approximately 62% (46/74) of our participants were rural, so decreasing drive time to their provider, in some cases, may have enabled participants to receive care they would not have otherwise been able to access. About 57% (42/74) of our participants were women, and all had a history of trauma. This may have contributed to why participants felt more comfortable receiving their care in their own homes and why they appreciated avoiding crowded VHA waiting rooms. These findings echo previous studies that found similar benefits of telehealth for patients with obsessive-compulsive disorder and veterans receiving mental health care over video telehealth [29,46].

When using video telehealth, about a third of participants noted they only encountered 1-2 technical issues, which they were able to resolve with minimal interruption. Others with more significant interruptions were able to troubleshoot on their own or with their coach to eventually get connected or access the web-based material. Since this study describes the experiences of participants who completed webSTAIR, we can presume that they were able to navigate technical issues and complete the program. Given these findings, we can infer that if a technical issue is eventually resolved, the presence of the technical issue may not impact overall satisfaction. For this reason, resilience, supportive troubleshooting, and having a backup plan (eg, using the phone for sound when audio does not work) may contribute to greater satisfaction and the successful completion of internet-based interventions.

**Limitations**

The primary limitation of our study is selection bias. Our interview sample consisted of participants who completed the webSTAIR program and were prescreened to have adequate access to video telehealth. These factors may result in more positive experiences with the program, as reflected in the data. The 37% (74/202) response rate may have further exacerbated selection bias and resulted in more positive feedback. These patients were more likely to have the time and motivation to participate in an interview. The experiences of those who may not have felt the program’s impact or did not feel inclined to participate in an interview were not captured. Additionally, the study was not randomized, and we do not have a control group to compare satisfaction among those who dropped out of the study. However, as a qualitative study, our aim was to provide a detailed description of participant experiences and explore how common themes across experiences may be relevant to
clinicians and researchers interested in web-based mental health treatments.

**Conclusions**

Our findings suggest that a combination of self-paced, web-based content and a coach-guided intervention may integrate the best of both worlds, that is, the convenience and flexibility of a web-based modality and the structure, guidance, and accountability of a traditional psychotherapy approach. Our results indicate that providers and researchers should design tech-based interventions that are sensitive to individual differences and that integrate coaching support, as participants highly value coaching support for their digital interventions. Future research should focus on the impact of satisfaction on engagement, retention, and the therapeutic benefits of web-based mental health interventions.

**Acknowledgments**

The authors would like to acknowledge the South Central Mental Illness Research, Education, and Clinical Center (MIRECC) and the Center for Innovations in Quality, Effectiveness, and Safety (IQuESt, CIN13-413), which provided facility and resource support, as well as the Veterans Affairs (VA) Palo Alto National Center for PTSD, which assisted in the design, collection of data, and preparation and editing of the manuscript. The opinions expressed are those of the authors and do not necessarily reflect those of the Department of Veterans Affairs, the US government, or Baylor College of Medicine.

**Conflicts of Interest**

None declared.

Multimedia Appendix 1

Interview questions.

[DOCX File, 14 KB - mental_v11i1e50977_app1.docx ]

**References**

1. Office of research and development. U.S. Department of Veterans Affairs. URL: [https://www.research.va.gov/topics/mental_health.cfm](https://www.research.va.gov/topics/mental_health.cfm) [accessed 2023-01-18]


Abbreviations

PC-PTSD: Primary Care PTSD Screen
PHQ-2: 2-item Patient Health Questionnaire
PTSD: posttraumatic stress disorder
STAIR: Skills Training in Affective and Interpersonal Regulation
VA: Veterans Affairs
VHA: Veterans Health Administration
Combining mHealth Technology and Pharmacotherapy to Improve Mental Health Outcomes and Reduce Human Rights Abuses in West Africa: Intervention Field Trial

Dror Ben-Zeev¹, PhD; Anna Larsen¹,², MPH, PhD; Dzifa A Attah³, MPhil, PhD; Kwadwo Obeng⁴, MBChB; Alexa Beaulieu¹, MPH; Seth M Asafo³, MPhil, PhD; Jonathan Kuma Gavi³, PhD; Arya Kadakia¹, BA; Emmanuel Quame Sottie³, BA; Sammy Ohene³, MBChB; Lola Kola⁵,⁶, PhD; Kevin Hallgren¹, PhD; Jaime Snyder¹,⁷, PhD; Pamela Y Collins⁸, MPH, MD; Angela Ofori-Atta³, MA, PhD; M-Healer Research Team³

Abstract

Background: In West Africa, healers greatly outnumber trained mental health professionals. People with serious mental illness (SMI) are often seen by healers in “prayer camps” where they may also experience human rights abuses. We developed “M&M,” an 8-week-long dual-pronged intervention involving (1) a smartphone-delivered toolkit designed to expose healers to brief psychosocial interventions and encourage them to preserve human rights (M-Healer app), and (2) a visiting nurse who provides medications to their patients (Mobile Nurse).

Objective: We examined the feasibility, acceptability, safety, and preliminary effectiveness of the M&M intervention in real-world prayer camp settings.

Methods: We conducted a single-arm field trial of M&M with people with SMI and healers at a prayer camp in Ghana. Healers were provided smartphones with M-Healer installed and were trained by practice facilitators to use the digital toolkit. In parallel, a study nurse visited their prayer camp to administer medications to their patients. Clinical assessors administered study measures to participants with SMI at pretreatment (baseline), midtreatment (4 weeks) and post treatment (8 weeks).

Results: Seventeen participants were enrolled and most (n=15, 88.3%) were retained. Participants had an average age of 44.3 (SD 13.9) years and 59% (n=10) of them were male. Fourteen (82%) participants had a diagnosis of schizophrenia and 2 (18%) were diagnosed with bipolar disorder. Four healers were trained to use M-Healer. On average, they self-initiated app use 31.9 (SD 28.9) times per week. Healers watched an average of 19.1 (SD 21.2) videos, responded to 1.5 (SD 2.4) prompts, and used the app for 5.3 (SD 2.7) days weekly. Pre-post analyses revealed a significant and clinically meaningful reduction in psychiatric symptom severity (Brief Psychiatric Rating Scale score range 52.3 to 30.9; Brief Symptom Inventory score range 76.4 to 27.9), psychological distress (Talibieh Brief Distress Inventory score range 37.7 to 16.9), shame (Other as Shamer Scale score range 41.9 to 28.5), and stigma (Brief Internalized Stigma of Mental Illness Scale score range 11.8 to 10.3). We recorded a significant reduction in days chained (1.6 to 0.5) and a promising trend for reduction in the days of forced fasting (2.6 to 0.0, \(P=.06\)). We did not identify significant pre-post changes in patient-reported working alliance with healers (Working Alliance Inventory), depressive symptom severity (Patient Health Questionnaire-9), quality of life (Lehman Quality of Life Interview for the Mentally Ill), beliefs about medication (Beliefs about Medications Questionnaire–General Harm subscale), or other human rights abuses. No major side effects, health and safety violations, or serious adverse events occurred over the course of the trial.

Conclusions: The M&M intervention proved to be feasible, acceptable, safe, and clinically promising. Preliminary findings suggest that the M-Healer toolkit may have shifted healers’ behaviors at the prayer camp so that they commit fewer human rights abuses.

(JMIR Ment Health 2024;11:e53096) doi:10.2196/53096
Introduction

Mental health systems in West Africa are constrained in their resources, infrastructure, and access to trained mental health personnel, limiting their capacity to deliver care [1,2]. In part due to these limitations, and in part due to local belief systems that frame mental health problems as spiritual rather than medical or psychological in nature [3,4], people with serious mental illness (SMI) in the region are often treated by traditional or faith healers, who are in abundance [5,6].

Healers often provide services at spiritual centers or "prayer camps"—rustic facilities where family members bring their relatives with mental illness, developmental disabilities, or substance use problems to be "healed" [7]. Pragmatically, prayer camps serve as West Africa’s de facto inpatient units and psychiatric residential homes. Once there, individuals seeking care may be retained at prayer camps for weeks, months, or even years at a time. Prayer camps are unregulated by any authorities, for the most part. Healers rarely have training in the etiology, assessment, or treatment of psychopathology, and their clients will seldom receive psychotropic medications or evidence-based psychosocial interventions. Healers provide spiritual consultation, prescribe prayer, and administer various ceremonial or herbal remedies [8,9]. In the absence of both training and resources to manage mental illness, healers may also engage in practices that have dangerous effects and constitute human rights abuses. These may include forced fasting, flogging, confinement in overcrowded or unsanitary conditions, and chaining patients to trees or concrete slabs so that they do not escape the camp’s grounds [9,10]. Chaining or similar forced mechanical restraining of people with SMI, particularly those exhibiting signs of psychosis, are common practices in many low- and middle-income countries and are psychologically and physically damaging [11-13]. Despite the well-documented harsh practices used by some healers, and the negative attention their practices receive from Western media and human rights groups, these paraprofessionals continue to receive referrals and fill a societal need.

Our multinational team works closely with traditional and faith healers and people with SMI in West Africa. Through consultation, bilateral information sharing, articulation of mutual respect, and transparency in our activities, we have fostered partnerships with healers and their prayer camp communities [14-16]. These ties have served as the building blocks for cross-sector collaborations that are designed to improve the safety and quality of care that people with mental illness receive at prayer camps.

Findings from these collaborations have led us to develop M&M, a dual-pronged intervention involving (1) a smartphone-delivered mobile health treatment support toolkit (M-Healer app) designed to expose healers to brief psychosocial interventions, encourage them to preserve human rights in their practice, and prompt them to monitor the status of the people they serve, and (2) a visiting community nurse who provides pharmacological care directly to their patients at the prayer camp (Mobile Nurse). The combination of healer-facing support technology, which can be installed on mobile devices that are widely accessible to healers throughout the region, and patient-facing pharmacotherapy delivered by medically trained personnel, which are more widely available than specialty mental health providers, may be a scalable strategy for addressing unmet mental health and quality-of-care needs in the region. Here we report on the first field trial of the integrated M&M intervention in a prayer camp in Ghana.

Methods

Study Design

The study involved a field trial of the integrated M&M intervention at a large prayer camp in Ghana. The objectives of the study were to (1) examine the safety and acceptability of the intervention among people with mental illness who are treated at the camp (hereinafter referred to as “patients”) and the prayer camp staff who provide them with services (hereinafter referred to as “healers”) and (2) evaluate the preliminary effectiveness of the intervention package.

Ethical Considerations

The study was approved by the institutional review boards of the University of Washington (00015549) and the University of Ghana (0001276). All study participants provided informed consent. Both healer and patient participants were compensated 75 Ghanaian Cedi (US $5.83) for their engagement in each assessment interview with study personnel.

Study Procedures

Initially, members of the investigative team met with the leader of the prayer camp (hereinafter referred to as “the Prophet”) to describe the project (ie, objectives, procedures, and timeline), introduce members of the team, and demonstrate the M-Healer technology on a smartphone. During this meeting, the team was introduced to the 4 healers who would participate in the study. In a second meeting, the team met with the healers separately to discuss their routine practices at the camp, describe the “enhanced” services they would be trained to deliver in the context of the study, solicit feedback, cocreate the study’s operational timeline, and address any questions or concerns.

Candidate patients for intervention were suggested by prayer camp staff and screened on-site by study team personnel. The initial screening was conducted by study doctors (psychiatry residents). Participants were evaluated for general medical and psychiatric fitness for participation. Medical assessment included measurement of temperature, pulse, BMI, blood pressure, random blood sugar, blood hemoglobin level, and malaria. The
study inclusion criteria were as follows: (1) being aged 18 years or older; (2) speaking English or Twi; (3) being a current inpatient staying at a study prayer camp; (4) having a diagnosis of schizophrenia spectrum disorder (ie, schizophrenia, schizoaffective disorder, delusional disorder, or schizoaffective disorder), bipolar disorder, or major depressive disorder, as determined by study staff administering the relevant sections of the Structured Clinical Interview for the Diagnostic and Statistical Manual of Mental Disorders (Fifth Edition) during screening. Candidates were excluded if they had a serious physical illness or needed urgent medical attention (eg, they had a high fever, serious infection, visible injury, or hemorrhage). Individuals who met the criteria and who were interested in participating were enrolled in the study. Participants were provided with a snack and meal by the study staff during assessment visits.

**Intervention Description**

M&M is an 8-week-long dual-pronged intervention package comprising psychoeducation, skills training, and treatment support scaffolding tools delivered to healers via the M-Healer toolkit app, and pharmacotherapy that is administered directly to patients at prayer camps by a visiting mobile nurse.

**Intervention Component 1: M-Healer Toolkit**

M-Healer is an Android Smartphone app that was specifically designed to serve as a digital toolkit for healers providing care to people with mental illness in West Africa [16]. The goal of the toolkit is to provide exposure to psychoeducational materials and to support basic monitoring of patient progress. Once installed on an Android device, the app does not require an active data plan or internet connectivity to operate, mitigating possible disruptions in functionality in resource-constrained environments. M-healer is available in English and Twi, the 2 most common languages spoken in Ghana. The app has 4 core functionalities. First, M-Healer provides psychoeducation on the administration of psychosocial interventions such as guided relaxation techniques, rapport building, verbal de-escalation, challenging dysfunctional beliefs about psychiatric symptoms, and preservation of human rights and dignity in practice. All content is accessible “on demand” from the home screen and delivered as brief digital animations or audio recordings (Figure 1). Second, M-Healer allows healer users to create a list of active patients currently at the camp to support basic tracking and monitoring of individual progress. New patients are assigned a number, and healers have the option to also record an audio identifier for each individual. Every day, M-Healer prompts the healer to check in with each patient and provide a rating on whether they are doing better, worse, or the same as the day before. Ratings are assigned in the form of emojis with happy, sad, or neutral facial expressions. Once entered, a daily rating is added to a series of small icons showing patient progress over several days, which enables the healer to track patient progress over time (Figure 1). Third, M-Healer has push notifications, prompting users to watch the “video of the day” drawn from the bank of psychosocial digital animation training videos. Fourth, alongside these user-facing functionalities, M-Healer also tracks the frequency and duration of use in a given time frame, allowing the study staff and the nurse to monitor healers’ interactions with the M-Healer toolkit.

Healer participants were provided Android smartphones with M-Healer installed and activated, which they were allowed to retain at the conclusion of the study. Healers were trained by 2 practice facilitators (study coinvestigators) and the study nurse on how to use M-Healer. Practice facilitation was conducted over 2 sessions in a designated space within the prayer camp, which was removed from the areas where patients resided. In the first session, the practice facilitators established rapport and trust, introduced the M-Healer toolkit, and taught some technical and clinical skills to the healer. This session was interactive and involved demonstrations of how to use the system and its features. Each healer had an opportunity to practice these features with the assistance of the facilitators and the nurse. To help the healers understand how they could use M-Healer skills in caring for their patients, facilitators led role-plays, which involved the use of preselected videos and providing real-time instruction as the healers navigated through their devices. Role-playing skills in the videos involved the participation of both cofacilitators or a cofacilitator and a volunteer who modeled as a healer and patient. Role-playing sessions on how to add patients and track their progress involved all healers. As the lead facilitator led the training, the other study team members walked around to assist healers who were experiencing navigational or technical difficulties. To conclude the session, healers were tasked to watch a video, listen to an audio module, and check in and track the progress of their patients on a daily basis.
The second facilitation session was designed to strengthen rapport and trust, refine technical skills, and strengthen clinical skill acquisition. The facilitators and study nurse reviewed the healers’ experiences of using M-Healer and their use data. Healers’ efforts in using the app were also acknowledged and praised. Healers were then asked to demonstrate how to watch a video or listen to an audio module, add a new patient for daily monitoring, and record their patients’ progress. The facilitators addressed any questions and concerns. At the end of the session, the healers were reminded to watch a video, listen to an audio module, and check in with their patients daily moving forward. After these 2 training sessions, the healers continued to receive ongoing M-Healer support from the study nurse during weekly site visits. During these visits, the nurse had opportunities to interact with the healers, inquire about and address their difficulties, check and record their user data, and encourage their use of the app.

**Intervention Component 2: Mobile Nurse**

A registered mental health nurse trained to use the World Health Organization’s Mental Health Gap Action Programme–Intervention Guide (mhGAP-IG; version 2.0) served as the intervention’s mobile nurse. Participants were prescribed medications by the screening study doctors. When possible, the prescribers opted to recommend long-acting injectable medications for participants with symptoms of psychosis to mitigate the risk of poor adherence to oral medication regimens. The use of long-acting injectable
medications was facilitated by the fact that most of the participants had received treatment with antipsychotics prior to their entry into the camp. Participants ultimately decided whether they would take medication and specified their preferred medication modality. The nurse monitored participants’ BMI, blood pressure, pulse, and temperature weekly and monitored their random blood sugar monthly. The nurse conducted weekly assessments of response to treatment and medication side effects using mhGAP-IG guidance. The nurse was supervised clinically by 2 study psychiatrists and instructed to call them if they encountered any urgent treatment-related problems that were not clearly described by the mhGAP-IG or the study protocol. The clinical team maintained active, real-time telecommunication to facilitate timely consultation and a weekly clinical team meeting to review nonurgent inquiries.

Data Collection

Study assessors (masters’-level clinical psychologists) administered a battery of measures to patient participants at baseline, midintervention (week 4), and post intervention (week 8). Assessments were administered in English or Twi using a tablet-based data collection software REDCap (Research Electronic Data Capture) to facilitate secure, password-protected, on-site digital data collection. We collected information on demographics, psychological symptoms, psychological distress, quality of life, and prayer camp experiences.

Patient participants’ psychiatric symptoms were measured at baseline, midtreatment, and post treatment using three measures: (1) the Brief Psychiatric Rating Scale (BPRS) [17], an 18-item clinical assessor–rated measure of current indicators of psychiatric illness including somatic concern, anxiety, emotional withdrawal, conceptual disorganization, feelings of guilt, tension, mannerism and posturing, grandiosity, depressive mood, hostility, suspiciousness, hallucinatory behavior, motor retardation, uncooperativeness, unusual thought content, blunted affect, excitement, and disorientation; the study assessors rated each indicator from 1 (“not present”) to 7 (“extremely severe”), with higher scores indicating more severe symptoms (range 18-126); (2) the Brief Symptom Inventory (BSI) [18], a 53-item self-report assessment measuring 9 symptom dimensions; patients rated how much each symptom has bothered them in the past 8 days from 0 (“not at all”) to 4 (“extremely”), with higher scores indicating more severe distress over the past 7 days (range 0-212); and (3) the Patient Health Questionnaire-9 [19], a 9-item depressive symptom scale in which respondents self-report how often (1=“not at all” to 4=“nearly every day”) they experience symptoms including feeling depressed and hopeless over the prior 2 weeks.

The following measures were also administered at baseline, midtreatment, and post treatment. Psychological distress was measured using the Talibieh Brief Distress Inventory [20], a 24-item self-report measure using a Likert scale for prior-month frequency rating (0=“not at all” to 4=“extremely”). Higher scores represent more severe distress. Quality of life was measured using the global quality-of-life items from the Lehman Quality of Life Interview for the Mentally Ill [21], whereby participants rate how they feel about various aspects of their life from 1 (“terrible”) to 7 (“delighted”), and higher scores indicate higher quality of life. Shame was measured using the Other as Shamer Scale [22], an 18-item self-report (range 8-32; higher scores correspond to higher experiences of guilt and shame). Patient beliefs about medication were assessed using the Beliefs about Medications Questionnaire–General Harm subscale [23], internalized stigma was measured with the Brief Internalized Stigma of Mental Illness scale [24], and patient working alliance with the healers at the camp was assessed with an adapted version of the Working Alliance Inventory [25].

Human rights abuses were measured at baseline, midtreatment, and post treatment. At each assessment, study assessors interviewed patient participants and asked them to self-report on their experiences during the preceding week, up until and including the day of the assessment. Specially, they were asked to report the number of days they were chained or shackled (ie, “days chained”), the number of days they were forced to take herbal remedies, the number of hours they were held in isolation, and the total number of times they were touched on their genitals in an uncomfortable manner in the previous week.

To characterize the sample before treatment, we collected data on individuals’ history of sexual abuse (Sexual Abuse Severity Scale) [26] substance use (Tobacco, Alcohol, Prescription Medications, and Other Substances scale) [27], and social support (Oslo Social Support Scale) [28] at baseline.

To measure the feasibility of the M-Healer intervention, we collected data post treatment on healers’ use of the M-Healer technology, as well as knowledge and skills related to engagement with M-Healer. At each visit, study staff viewed the password-protected M-Healer data analytics screen on the healers’ phones and recorded data including the number of days on which the app was used, the number of videos watched, the number of user-initiated app interactions, and the number of prompts to which the healers responded. To assess fidelity, healers participated in a 12-item verbally administered knowledge checklist whereby healers agreed or disagreed with statements regarding M-Healer knowledge domains. Healers performed 3 behavioral tests to demonstrate skills learned in M-Healer modules by interacting with the study staff acting as a standardized patients in role-played scenarios common to prayer camp settings. Study staff rated healer-demonstrated skills in a checklist.

Statistical Analysis

Descriptive statistics were used to characterize patient participants and their psychiatric and psychosocial factors at all study visits. We used paired samples t tests to compare baseline to midtreatment and baseline to posttreatment mean scores. For participants missing <50% of scale items for a psychosocial scale, item-level scores were imputed as the median score across the participant’s existing scale items (person-median imputation) [29]. Psychosocial scores were not analyzed for participants missing ≥50% of scale items. Quantitative analyses were conducted using Stata (version 17; StataCorp).
Results

In total, 29 patients were screened for study eligibility among whom 19 (65.5%) were eligible to participate and 17 (89.5%) were enrolled (Figure 2). The majority of participants (n=15, 88.3%) were retained for midtreatment (4 weeks) and posttreatment (8 weeks) follow-up visits. Participants had an average age of 44.3 (SD 13.9) years; 59% (n=10) of them were male, and 82% (n=14) of them identified as Christian (Table 1). Fourteen (82%) participants had a diagnosis of schizophrenia, 1 had a comorbid diagnosis of major depressive disorder, and 2 were diagnosed with bipolar disorder. The average number of weeks spent residing in the camp was 72.6 (SD 111.9), and this was the first experience in a prayer camp for 41% (n=7) of participants.

We found a significant and clinically meaningful reduction in psychiatric symptom severity from pretreatment to posttreatment as measured by the BPRS and the BSI (Table 2). Subjective ratings of psychological distress and shame were significantly lower post treatment than before treatment. Participants had significantly reduced internalized stigma about their mental health conditions from pretreatment to posttreatment. We recorded a significant reduction in days chained from pretreatment to posttreatment.

Figure 2. Flowchart for prayer camp participant enrollment. *Reasons for ineligibility were not mutually exclusive. **Malaria Rapid Diagnostic Test positive: n=1; hemoglobin level < 8% and symptoms: n=1; pulse > 120 bpm and symptoms: n=2. ***Participants returned for the 8-week visit.
<table>
<thead>
<tr>
<th>Demographic characteristics</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (years; n=16), mean (SD)</td>
<td>44.3 (13.9)</td>
</tr>
<tr>
<td><strong>Sex, n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>7 (41.0)</td>
</tr>
<tr>
<td>Male</td>
<td>10 (59.0)</td>
</tr>
<tr>
<td><strong>Educational attainment, n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>Primary</td>
<td>2 (11.8)</td>
</tr>
<tr>
<td>Junior high school</td>
<td>6 (35.3)</td>
</tr>
<tr>
<td>Senior high school</td>
<td>5 (29.4)</td>
</tr>
<tr>
<td>Tertiary or higher</td>
<td>4 (23.5)</td>
</tr>
<tr>
<td><strong>Marital status, n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>Married</td>
<td>6 (35.0)</td>
</tr>
<tr>
<td>Divorced or separated</td>
<td>2 (12.0)</td>
</tr>
<tr>
<td>Single</td>
<td>9 (53.0)</td>
</tr>
<tr>
<td><strong>Religion n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>Christian</td>
<td>14 (82.0)</td>
</tr>
<tr>
<td>No religion</td>
<td>1 (6.0)</td>
</tr>
<tr>
<td>Other</td>
<td>1 (6.0)</td>
</tr>
<tr>
<td>Missing</td>
<td>1 (6.0)</td>
</tr>
<tr>
<td>Weeks residing in this prayer camp (n=14), mean (SD)</td>
<td>72.6 (111.9)</td>
</tr>
<tr>
<td><strong>First experience residing in a prayer camp</strong></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>7 (41)</td>
</tr>
<tr>
<td>No</td>
<td>8 (47)</td>
</tr>
<tr>
<td>Missing data</td>
<td>2 (12)</td>
</tr>
<tr>
<td><strong>Social support</strong>&lt;sup&gt;a&lt;/sup&gt; (n=12), mean (SD)</td>
<td>7.8 (2.6)</td>
</tr>
<tr>
<td>Ever use tobacco&lt;sup&gt;b&lt;/sup&gt;, n (%)</td>
<td>2 (12)</td>
</tr>
<tr>
<td>Ever use alcohol&lt;sup&gt;b&lt;/sup&gt;, n (%)</td>
<td>2 (12)</td>
</tr>
<tr>
<td>Ever use drugs&lt;sup&gt;b&lt;/sup&gt;, n (%)</td>
<td>2 (12)</td>
</tr>
<tr>
<td>Ever use prescription drugs&lt;sup&gt;b&lt;/sup&gt;, n (%)</td>
<td>0 (0)</td>
</tr>
<tr>
<td><strong>Psychiatric diagnosis n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>Schizophrenia</td>
<td>14 (82.0)</td>
</tr>
<tr>
<td>Bipolar disorder</td>
<td>2 (12.0)</td>
</tr>
<tr>
<td>Major depressive disorder (comorbid with a primary diagnosis of schizophrenia)</td>
<td>1 (6.0)</td>
</tr>
</tbody>
</table>

<sup>a</sup>3-item Oslo Social Support Scale; absolute range 3-12; a higher score indicates a higher level of social support.

<sup>b</sup>Tobacco, alcohol, prescription medications, and other substances scale.
<table>
<thead>
<tr>
<th></th>
<th>Pretreatment baseline assessment (n=17)</th>
<th>Midtreatment 4-week assessment (n=15)</th>
<th>Posttreatment 8-week assessment (n=15)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>Mean (SD)</td>
<td>P value</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>Mean (SD)</td>
<td>N</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BPRS&lt;sup&gt;a&lt;/sup&gt;</td>
<td>16</td>
<td>52.3 (24.0)</td>
<td>.01&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>BSI&lt;sup&gt;b&lt;/sup&gt;</td>
<td>16</td>
<td>76.4 (56.6)</td>
<td>.004</td>
</tr>
<tr>
<td>TBDI&lt;sup&gt;d&lt;/sup&gt;</td>
<td>15</td>
<td>37.7 (23.5)</td>
<td>.002</td>
</tr>
<tr>
<td>OAS&lt;sup&gt;e&lt;/sup&gt;</td>
<td>15</td>
<td>41.9 (21.3)</td>
<td>.64</td>
</tr>
<tr>
<td>BMQ-Harm&lt;sup&gt;f&lt;/sup&gt;</td>
<td>14</td>
<td>11.8 (3.6)</td>
<td>.007</td>
</tr>
<tr>
<td>ISMI-10&lt;sup&gt;g&lt;/sup&gt;</td>
<td>14</td>
<td>30.4 (3.9)</td>
<td>—</td>
</tr>
<tr>
<td>PHQ-9&lt;sup&gt;h&lt;/sup&gt;</td>
<td>15</td>
<td>10.4 (6.5)</td>
<td>.35</td>
</tr>
<tr>
<td>Lehman Quality of Life Scale for the Mentally Ill&lt;sup&gt;i&lt;/sup&gt;</td>
<td>14</td>
<td>20.1 (7.8)</td>
<td>.40</td>
</tr>
<tr>
<td>Working Alliance Inventory–Bond Subscale&lt;sup&gt;j&lt;/sup&gt;</td>
<td>13</td>
<td>14.8 (5.6)</td>
<td>.80</td>
</tr>
<tr>
<td>Days chained&lt;sup&gt;k&lt;/sup&gt;</td>
<td>16</td>
<td>1.6 (2.8)</td>
<td>.09</td>
</tr>
<tr>
<td>Days forced to fast&lt;sup&gt;l&lt;/sup&gt;</td>
<td>16</td>
<td>2.6 (5.5)</td>
<td>.12</td>
</tr>
<tr>
<td>Days forced to take herbal remedies&lt;sup&gt;l&lt;/sup&gt;</td>
<td>16</td>
<td>0.06 (0.3)</td>
<td>.40</td>
</tr>
<tr>
<td>Hours kept isolated&lt;sup&gt;l&lt;/sup&gt;</td>
<td>16</td>
<td>3.0 (8.2)</td>
<td>.32</td>
</tr>
<tr>
<td>Times touched in an uncomfortable manner&lt;sup&gt;l&lt;/sup&gt;</td>
<td>16</td>
<td>0.1 (0.3)</td>
<td>.14</td>
</tr>
</tbody>
</table>

<sup>a</sup>BPRS: Brief Psychiatric Rating Scale—an 18-item scale with an absolute range of 18-126; higher scores indicate higher symptom severity.

<sup>b</sup>Italicized values are statistically significant at P<.05.

<sup>c</sup>BSI: Brief symptom inventory—a 53-item scale with an absolute range of 0-212; higher scores indicate higher symptom severity.

<sup>d</sup>TBDI: Tablieh Brief Distress Inventory—a 24-item scale with an absolute range of 0-96; higher scores indicate greater distress.

<sup>e</sup>OAS: Other as Shamer Shame Scale—an 18-item scale with an absolute range of 0-90; higher scores indicate greater shame.

<sup>f</sup>BMQ-Harm: Beliefs about Medications Questionnaire–Harm Subscale—a 4-item scale with an absolute range of 4-20; higher scores indicate unfavorable thoughts about medications.

<sup>g</sup>ISMI-10: Internalized Stigma of Mental Illness Inventory—a 10-item scale with an absolute range of 10-40; higher scores indicate greater internalized stigma.

<sup>h</sup>Not available.

<sup>i</sup>PHQ-9: Patient Health Questionnaire—9—a 9-item scale with an absolute range of 0-27; higher scores indicate higher symptom severity.

<sup>j</sup>A 6-item scale with an absolute range of 6-42; higher scores indicate better quality of life.

<sup>k</sup>A 4-item scale with an absolute range of 4-20; higher scores indicate better working alliance.

Psychiatric symptoms measured with the BPRS and BSI were reduced midtreatment compared to those at baseline. Unfavorable beliefs about medications had decreased by the midtreatment visit compared to those at baseline. We observed a trend toward statistical significance for reduction in forced fasting by the posttreatment follow-up visit (P=.06). Over the intervention period, we did not identify changes in patient-reported working alliance with healers, depressive symptom severity, quality of life, days forced to take herbal
remedies, hours retained in isolation, or number of times a participant reported being touched in an uncomfortable manner.

Results with imputed variables were not meaningfully different from those obtained from raw data. Notably, a significant decrease in unfavorable beliefs about medications from pretreatment to posttreatment was identified in the nonimputed data set; yet, this change was not detected in the imputed data set. Participants’ scales with imputed values were predominantly only missing 1-2 items, which were imputed using person-median values per scale to retain the available data.

In the initial visit to the prayer camp, 15 participants consented to receiving pharmacotherapy. For schizophrenia treatment, fluphenazine decanoate (25 mg/mL) was the most commonly prescribed pharmacotherapy (n=12, 80.0%), followed by trihexyphenidyl (5 mg tablet; n=6, 40.0%) and risperidone (2 mg; n=3, 20.0%); some participants were prescribed multiple medications. Among participants with bipolar disorder, fluphenazine decanoate (25 mg), risperidone (25 mg), and trihexyphenidyl (5 mg) were each prescribed to 6.7% (n=1) of participants. Overall, 110 medication follow-up visits were conducted by the mobile nurse.

Four healers were trained to use M-Healer and their usage of the app was monitored weekly. M-Healer feasibility measures indicated that healers had an average of 31.9 (SD 28.9) user-initiated app uses per weekly visit, watched 19.1 (SD 21.2) videos, responded to 1.5 (SD 2.4) prompts, and used the app for 5.3 (SD 2.7) days a week. Healers had posttreatment average knowledge assessment checklist scores of 11.75 (absolute range 9-14, total possible score=16). Two healers successfully demonstrated all 3 behavioral skills, 1 healer demonstrated 2 skills, and 1 healer demonstrated 1 skill.

**Discussion**

**Principal Results**

The combination treatment package deployed in this study was designed to inform and shape West African healers’ practices and to provide symptomatic relief to individuals with psychiatric illnesses receiving services at their prayer camps. The intervention proved to be feasible and acceptable to healers and their patients. Healers were able to navigate the M-Healer system successfully on the study smartphone devices that were provided to them. Healers expressed a clear understanding of M-Healer functionalities, watched psychoeducational videos, and listened to the audio lessons over the study period. Healers in the study self-initiated M-Healer use throughout the intervention period above and beyond system-promoted interactions and chose to view M-Healer videos and listen to audio lessons at a rate that far exceeded our expectations. Healers’ use of the app was very frequent—almost daily. When evaluated, healers were able to describe the content accurately and to demonstrate skills that they had learned from the app.

Healers agreed to grant the mobile nurse access to the individuals they treated at the prayer camp, the majority of whom were placed in locked dormitories, with some shackles. The nurse was able to assess, provide pharmacotherapy to, and monitor these patients weekly without difficulty or obstruction, despite these austere conditions. All patients approached by our study team expressed openness to meeting with the nurse, and all but 1 expressed interest in receiving pharmacotherapy to manage their psychiatric symptoms and improve their health.

The combined intervention proved to be clinically promising. Patient participants experienced significant reductions in the severity of their psychiatric symptoms, psychological distress, subjective feelings of shame, and internalized stigma over the course of the intervention. Moreover, patient participants reported a significant reduction in the days during which they were chained or shackled by the healers and a nonsignificant trend of reduction in the days during which they were forced to fast. These findings suggest that exposure to the training content delivered by the M-Healer toolkit may have shifted healer behaviors after the intervention was implemented at the prayer camp’s study site. One participant was discharged from the prayer camp and allowed to return to their family over the course of the intervention by prayer camp staff.

The intervention proved to be safe. The study helped promote better care in the prayer camp’s study site. Over the course of the intervention, 3 participants were identified by our study staff as requiring immediate medical attention and were referred to the district hospital where they received care. The nurse monitored medication side effects weekly. No major side effects, health and safety violations, serious adverse events, or other major complications occurred over the course of the field trial.

Poststudy debriefing interviews with healers indicated that they felt some ambivalence about M-Healer content. On one hand, they enjoyed the opportunity to learn about novel psychosocial illness management strategies and indicated that the digital toolkit informed the way they engaged with their patients. On the other, some expressed frustration that the toolkit’s human rights module seemed too critical of their practices (eg, M-Healer’s guidance to refrain from chaining patients). Given that current prayer camp practices do not prioritize preservation of human rights, this tension may have been inevitable. A more complete report on the findings from these interviews will be provided in a separate qualitative study.

**Limitations**

The study has several limitations. First, the study involved patient participants and healers at a single prayer camp, limiting study generalizability. Prayer camps vary dramatically in terms of their size, practices, and openness to research and outsider involvement in their day-to-day activities. Future multisite research in several prayer camps can help determine the generalizability of our findings to other settings. Second, we relied on study participants’ self-reports to determine whether they experienced human rights abuses at the camp. Participants may have underreported these events due to shame, fear of retaliation, or to protect camp staff. Third, the study sample was small, allowing only for preliminary examination of treatment effects. Despite the exploratory nature of our evaluation of outcomes, promising evidence of clinical effectiveness was recorded. Future research involving larger samples will facilitate more robust evaluation of the effects of the intervention on participants’ health and well-being. Finally, prayer camp staff had previous experience being involved in research with our
team. It will be important to examine the intervention in research-naive prayer camps to evaluate whether such relationships are necessary for successful intervention uptake and outcomes.

Conclusions
In West Africa, traditional and faith healers greatly outnumber trained mental health professionals. Healers hold political capital and are respected by the local population. These key stakeholders can serve as facilitators, conduits, obstructionists, or blockades to the delivery of evidence-based mental health services in their communities. Engaging healers and training them to provide compassionate psychosocial interventions themselves and opening the doors of their prayer camps to medical professionals who can also deliver pharmacological interventions have the potential to enhance the regional capacity to address unmet mental health needs.

Previous research has demonstrated that pharmacotherapy can be conducted successfully at prayer camps. While psychotropics may help effectively manage the psychiatric symptoms of patients staying at prayer camps, those improvements did not translate to reductions in the days during which patients were shackled [14]. This study demonstrated that adding a key component—a healer-facing digital health training toolkit—can impact how healers interact with their patients, including a reduction in the use of mechanical restraints. These findings are encouraging and suggest that the M&M intervention holds promise as a dual-pronged model for both improving mental health outcomes and reducing human rights abuses in West African prayer camps. If future research finds similarly robust treatment effects on a larger scale, this will provide important information for Ghanaian policy makers. Such findings would demonstrate that a combination of accessible technologies coupled with individuals trained to deliver interventions in the field may help reconfigure how prayer camps serve people with SMI. Ghanaian government mandates, monitoring, and enforcement over the last few years are beginning to produce positive effects on the reduction of human rights abuses in prayer camps. In addition to barring harmful practices, the government can also play a key role in supporting the adoption of new models of care. Through targeted trainings and appropriate oversight, individuals who are already working in the field (eg, government-employed district nurses or members of the Psych Corps—psychology graduates who are posted to health facilities across the country as part of their national service) could be leveraged to train healers and other paraprofessionals in the use of evidence-based digital mental health technologies. Government investment in such force-multiplying activities and technologies can prove to be a pragmatic and scalable approach for addressing significant professional workforce shortages in the region.

Acknowledgments
The study was supported by a grant award from the US National Institute of Mental Health (R01MH127531-01A1; DBZ and AOA). The authors thank the camp leader for permitting the study to be conducted at his faith healing camp and for the collaboration he fostered between camp staff and the study team. The authors acknowledge and thank the head of the district health management team for sending a district nurse to the study site to engage with the study nurse to facilitate continuation of pharmacological care, the district health management team that facilitated the acquisition of national health insurance cards for the study, and the study participants.

Conflicts of Interest
DBZ has financial interests in Merlin LLC and FOCUS technology. He has had an intervention content licensing agreement with Pear Therapeutics and has provided consultation services to Trusst Health, K Health, Boehringer Ingelheim, eQuility, Deep Valley Labs, and Otsuka Pharmaceuticals. KO and SO have provided consultation services to Janssen pharmaceuticals.

References

**Abbreviations**

- **BPRS**: Brief Psychiatric Rating Scale
- **BSI**: Brief Symptom Inventory

https://mental.jmir.org/2024/1/e53096

JMIR Ment Health 2024 | vol. 11 | e53096 | p.211

(page number not for citation purposes)
Efficacy of BrighterSide, a Self-Guided App for Suicidal Ideation: Randomized Controlled Trial

Natasha Josifovski, PhD; Michelle Torok, PhD; Philip Batterham, PhD; Quincy Wong, PhD; Joanne R Beames, PhD; Adam Theobald, BPsych (Hons), MPsych (Clin); Sarah Holland, BComm; Kit Huckvale, PhD; Jo Riley, BSc; Nicole Cockayne, MPH, BSc (Hons); Helen Christensen, PhD; Mark Larsen, DPhil

Corresponding Author: Natasha Josifovski, PhD

Abstract

Background: Self-guided digital interventions can reduce the severity of suicidal ideation, although there remain relatively few rigorously evaluated smartphone apps targeting suicidality.

Objective: This trial evaluated whether the BrighterSide smartphone app intervention was superior to a waitlist control group at reducing the severity of suicidal ideation.

Methods: A total of 550 adults aged 18 to 65 years with recent suicidal ideation were recruited from the Australian community. In this randomized controlled trial, participants were randomly assigned to receive either the BrighterSide app or to a waitlist control group that received treatment as usual. The app was self-guided, and participants could use the app at their own pace for the duration of the study period. Self-report measures were collected at baseline, 6 weeks, and 12 weeks. The primary outcome was severity and frequency of suicidal ideation, and secondary outcomes included psychological distress and functioning and recovery. Additional data were collected on app engagement and participant feedback.

Results: Suicidal ideation reduced over time for all participants, but there was no significant interaction between group and time. Similar improvements were observed for self-harm, functioning and recovery, days out of role, and coping. Psychological distress was significantly lower in the intervention group at the 6-week follow-up, but this was not maintained at 12 weeks.

Conclusions: The BrighterSide app did not lead to a significant improvement in suicidal ideation relative to a waitlist control group. Possible reasons for this null finding are discussed.

Trial Registration: Australian New Zealand Clinical Trials Registry ACTRN12621000712808; https://trialsearch.who.int/Trial2.aspx?TrialID=ACTRN12621000712808

(JMIR Ment Health 2024;11:e55528) doi:10.2196/55528

KEYWORDS

suicidal ideation; suicide prevention; digital health; clinical trial

Introduction

One in 6 adults experience serious thoughts of suicide during their lives, with an estimated prevalence of 3.4% of adults experiencing suicidal ideation in a 12-month period [1]. Data from the World Health Organization suggest that those with suicidal ideation are 10 times more likely to make a suicide attempt across their lifetime than those without suicidal ideation [2]. In recent years, there has been a shift in how people seek mental health support for a suicidal crisis. For example, during the 2021-2022 financial year, Lifeline—an Australian crisis service—answered 1,142,234 calls, a 56% increase since 2019 [3]. At the same time, hospitals in Australia experienced a 14.3% reduction in mental health–related presentations [4]. One potential way to offset the high demand for crisis support services is to make self-guided digital health interventions publicly available for those in distress. These interventions have the potential to increase access to mental health care and enhance the capacity of mental health systems to respond to persons in crisis by offering high-fidelity, evidence-based therapeutic support, anonymously and at low to no cost, which can be readily accessed anywhere [5,6].

Meta-analytic evidence suggests that digital health interventions can effectively reduce the severity of suicidal thoughts [7].
Those interventions that specifically target suicidality are more effective than generalized mental health apps for reducing suicide-related outcomes [7]. However, despite the huge potential for self-guided smartphone interventions to address service access gaps for those experiencing suicidal distress, there are currently few digital interventions targeting suicidal ideation that have been rigorously tested in randomized controlled trials (RCTs). Even fewer of these have been in general adult populations.

One exception to the above is the web-based self-help program Living with Deadly Thoughts, which was adapted from the Dutch program Living Under Control [8]. Drawing from principles of cognitive behavioral therapy and dialectical behavior therapy, this RCT involved access to 6 online modules for community-recruited adults experiencing suicidal ideation. While the Dutch program found a small but significant effect in reducing suicidal thoughts, the English-adapted Living with Deadly Thoughts program found no difference between intervention and control groups in an Australian study [8]. One possibility for this discrepancy may be that the English study was underpowered to detect an effect size that would be comparable with the Dutch version. However, there may be additional nuances given the difference in the recruited population—the Dutch study included only participants who experienced mild to moderate suicidal thoughts (defined as a score between 1 and 26 on the Beck Scale for Suicidal Ideation) and who were not severely depressed (defined as a scores greater than 39 on the Beck Depression Inventory). Conversely, Living with Deadly Thoughts did not include any cutoffs for either depression or suicidal thoughts as exclusion criteria. An effect on suicidal ideation could therefore be dependent on the severity of suicidal thoughts, in that those with thoughts that are more severe may not benefit from the modules involved. Alternatively, or as well as this, the program may involve other factors that are not generalizable, and an effective program to reduce suicidal ideation may require greater insight from those with lived experience.

In the context of these gaps, we developed BrighterSide. BrighterSide is a self-guided app based on a mix of cognitive behavioral therapy and dialectical behavior therapy, alongside elements of acceptance and commitment therapy and positive psychology. These therapeutic approaches have previously been demonstrated to be efficacious in reducing suicidal ideation when delivered through web-based programs [9,10]. The app involves 5 modules that each contain activities to encourage the user to develop and practice strategies to manage their suicidal thoughts. None of the activities last more than 5 minutes and users have complete control over which modules and activities they wish to complete. Furthermore, the app was co-designed with lived experience advisors in order to consider how best to maximize engagement and how best to support those with suicidal ideation or behaviors (see Torous et al [11] for a discussion on how co-design with consumers may enhance app engagement).

This study aimed to determine if those using the BrighterSide app would have a greater reduction in suicidal thoughts relative to the waitlist control group in a community trial in the adult population. We hypothesized that those in the intervention group would demonstrate significantly lower suicidal ideation, the primary outcome, at 6 and 12 weeks after baseline compared to a waitlist control group. We also hypothesized that the intervention group would report fewer incidents of self-harm and suicide attempts at 6 and 12 weeks after baseline and that the intervention group would report greater improvements in their ability to cope at 6 and 12 weeks after baseline.

Methods

Ethical Considerations

This study is reported as per the Consolidated Standards of Reporting Trials (CONSORT; Checklist 1) guideline (complete supplementary information is provided in Multimedia Appendix 1). The trial protocol was approved by the University of New South Wales Human Research Ethics committee (HC210196) and prospectively registered on the Australian New Zealand Clinical Trials Registry (ACTRN12621000712808).

Trial Design

The trial was a single-blind, 2-arm parallel RCT. Participants were randomized 1:1 between intervention and waitlist-control groups. Researchers were blind to group allocation. Those allocated to the intervention group received immediate access to the BrighterSide app upon completing the baseline survey. Those allocated to the control group received access to the app at the end of the trial.

Participants

Individuals were eligible for the study if they were (1) aged 18 to 65 years, (2) had experienced suicidal ideation within the previous 2 weeks, (3) owned an iPhone (with iOS 13.0 or higher) or Android (with Android 6.0 or higher) smartphone, (4) were fluent in English, and (5) currently lived in Australia. There were no specific exclusion criteria as this enabled a more heterogeneous sample.

Recruitment took place in 2 waves: the first was from June 30, 2021, to July 23, 2021, and the second was from September 20, 2021, to September 24, 2021 (see the Sample Size section). All data collection was completed by December 24, 2021. Participants were recruited via multiple channels. First, the trial was advertised on the Black Dog Institute (BDI) website. Second, recent visitors to the Black Dog Institute Online Clinic (a free mental health assessment tool) who had indicated recent suicidal ideation and consented to be contacted for future research opportunities were sent an invitation email by the clinic team. Third, the study was advertised on Facebook and Instagram via the BDI and Lifeline social media channels and paid advertising.

All individuals who responded to a study advertisement first completed an online screening questionnaire to determine eligibility. Those who screened as ineligible were redirected to a web page with information on crisis services. Eligible participants were presented with the participant information statement and digital consent form. Consenting participants then completed baseline questionnaires online via BDI’s bespoke trial software, upon completion of which participants were randomized. Randomization was performed by the trial...
management software using a block size of 4. Participants were notified of group allocation, but investigators remained blinded.

Participants in the intervention group were sent an email on baseline completion with a link to download the BrighterSide app from the Apple App Store or Google Play Store, along with a unique link that provided access to the app content. This prevented use of the app by users who were not registered to the trial. There was no specific timeframe to download the app, nor were any reminders sent if participants did not download it. Participants in the control group were sent an email with the link to access the app on completion of the final 12-week questionnaire (or, if not completed, at 13 weeks). Participants did not receive reimbursement or incentives to participate in the study. Participants were permitted to engage in other treatment, preexisting or new, while participating in the study.

**Intervention**

BrighterSide is a self-guided smartphone app for adults experiencing suicidal ideation. The app was designed to help users develop and practice strategies to manage suicidal thoughts and was derived from content from the Living with Deadly Thoughts online program [8]. A multidisciplinary team of clinicians, researchers, lived experience advisors, designers, and developers engaged in a human-centered design process to update the original content into a more engaging form, ensuring the language, design, and user journey within the app were clear, simple, and supportive for adults experiencing suicidal thinking. This multidisciplinary collaboration was engaged across all features of the app, including the safety planning and check-in features.

The app contained 5 modules: Understand your Thoughts, Prevent a Crisis, Navigate your Emotions, Navigate your Thoughts, and Plan for the Future (see Table S1 in Multimedia Appendix 1 for brief details of each module, and Multimedia Appendix 2 for selected screenshots of each module). Each module contained interactive activities (e.g., guided breathing). The content was based on cognitive behavioral therapy and dialectical behavior therapy, with elements of acceptance and commitment therapy and positive psychology. Users could access content in any order they wished or could choose a guided option that progressed through the modules in a specific order. The app included a safety planning function where users could list warning signs, helpful techniques, and supportive contact details and share them with others via email. The app also included a daily “check-in" feature that asked whether the user was safe and linked them to their safety plan and crisis contact numbers. Finally, in addition to the modules, the app also included “distraction activities," such as Bubble Pop, a built-in game simulating the motion of popping bubble wrap, and “calming activities," such as guided mindfulness recordings. Participants were free to engage with the app in whichever way they chose, including after the final data collection at 12 weeks, and the trial did not mandate frequency or patterns of use. A guided option was available if participants preferred, which allowed them to prioritise modules in a specific order depending on their main concern. The intervention did not include reminders to use the app. Participants in the waitlist control group received an email with details of crisis support services, and access to the app was granted after the 12-week study period.

**Outcomes**

Standard demographic data were collected at baseline, with outcome measures collected at baseline, 6 weeks, and 12 weeks. Self-report questionnaires were administered online via the BDI’s Research Engine platform, and participants were sent a link via email at 6 and 12 weeks to complete the follow-up time points.

The primary outcome was the frequency and severity of suicidal ideation, as measured by the Suicidal Ideation Attributes Scale (SIDAS) [12]. This comprises 5 items measuring frequency of ideation, controllability, closeness to attempt, level of distress, and the impact on daily functioning. The total score is in the range 0 to 50, with higher scores indicating higher levels of suicidal ideation. Scores of 21 or higher are indicative of high risk for suicidal behavior [12]. The Cronbach α for this study was α=.803.

Secondary outcomes included self-harm behavior, coping strategies (abbreviated Coping Orientation to Problems Experienced Inventory; Brief-COPE [13]) functioning and recovery (Functioning and Recovery Scale; FRS [14]), psychological distress (Distress Questionnaire-5; DQ5 [15]), help-seeking (modified Actual Health-Seeking Questionnaire; ASHQ [16]), and days out of role (WHO Disability Assessment Schedule; WHODAS–1 item [17]). The distress, functioning and recovery, and days out of role measures were added following lived experience consultation—these measures were added following initial registration of the trial, but prior to recruitment opening.

At 6 weeks, participants in the intervention group completed additional measures in relation to the app: appropriateness of intervention (Implementation Appropriateness Measure; IAM [18]), the Digital Working Alliance Inventory (DWAI [19]) and a bespoke questionnaire seeking feedback on the BrighterSide app. Participants in the intervention group were also invited on completion of the 12-week measures to participate in a semistructured interview to provide detailed feedback on their experience using BrighterSide.

**Safety Monitoring**

Data on adverse events and serious adverse events related to suicidal ideation and suicide attempts, respectively, were routinely collected at all time points though the self-report outcome measures already described. Specifically, severe suicidal ideation was indicated by a total score of 21 or higher on the SIDAS, and a recent suicide attempt was marked by either a self-report of 1 or more suicide attempts in the past 6 weeks (in the self-harm behavior questionnaire), or as indicated by a score of 10 (“I have made an attempt”) on question 3 of the SIDAS.

If a participant indicated they had recently attempted suicide (as described above), an email was automatically sent to them with support contact details and to arrange for a follow-up phone call by the research team. If a response was not received within 1 business day, a second email was sent to the participant.
Follow-up phone calls ensured that the participant was safe and offered a referral to Lifeline, a telephone crisis support service who would be able to call the participant and offer specialized support. In the case of the intervention group, participants were also routinely asked during questionnaires if their suicidal ideation and/or suicide attempt was related to use of the BrighterSide app. If they indicated yes, the follow-up call would be identical to that described above, with an addition to seek clarity on if and how use of the app contributed to their suicidal thoughts or behaviours.

**Patient and Public Involvement**

The BDI’s lived experience advisory team collaborated on the content and design of the app. Furthermore, the lived experience advisors were consulted on the trial design and recommended including outcome measures related to functioning and recovery (measured with the FRS), psychological distress (measured with the DQ5), and days out of role (measured with WHODAS-1).

**Sample Size**

The initial recruitment target for the trial was 394 participants, with 197 participants per arm. This allowed for detection of a small to medium effect size ($d=0.3$) in the primary outcome (severity of suicidal ideation) between the intervention and control arms with 80% power ($\alpha=0.05$), allowing for 40% attrition at postintervention follow-up. This was informed by 3 previous trials that incorporated the underlying intervention content in a web-based program, reporting a pooled effect size of 0.31 and average attrition rate of 35% [8-10]. After observing a higher than estimated attrition rate at the 6-week follow-up ($n=223; 56.6\%$), the recruitment target was raised to 546 ($n=273$ per arm) to maintain statistical power.

**Statistical Methods**

Demographic and clinical characteristics were compared between study arms using 2-sided independent sample $t$ tests, $\chi^2$ tests, or the Fisher exact test. Mixed model repeated measures (MRRM) analyses with maximum likelihood estimation and an unstructured covariance matrix were used to evaluate the efficacy of the BrighterSide app relative to the control condition. The primary outcome was severity of suicidal ideation as assessed by the SIDAS over time (baseline to 6 weeks; baseline to 12 weeks). The mixed model approach incorporates all available data, including participants with missing follow-up data points, under the missing-at-random assumption that is robust to data that are missing contingent on observed variables. Analyses were performed under the intention-to-treat principle by a statistician who was blinded to group allocation. Analyses of secondary outcomes used the same methods for continuous outcomes. The frequency of adverse events, including severe suicidal ideation, recent self-harm, and recent suicide attempts, were compared between groups using a $\chi^2$ test of independence.

Descriptive statistics were used to evaluate use (indicated by the number of modules completed, recorded using app analytics) of the BrighterSide app by participants in the intervention condition. Interview data collected from a subset of participants in the intervention group were analyzed thematically by one of the authors [20]. An inductive approach, independent of a theoretical confirmative method, was used to identify and group themes. Themes were refined to determine the final coding framework.

**Post Hoc Analysis**

A post hoc analysis using logistic regression was performed to assess whether follow-up attrition rates at 6 and 12 weeks after baseline could be predicted by any factors measured at baseline.

**Results**

**Recruitment and Baseline Characteristics**

A total of 795 participants were assessed for eligibility, of whom 550 were randomized (see Figure S1 in Multimedia Appendix 1). A total of 275 participants were randomized to each group. All participants were analyzed under the intention-to-treat principle, except 1 (in the intervention group) who withdrew and requested that their data not be retained. Baseline characteristics for participants are presented in Table 1; the groups did not differ across any measure.
Table . Baseline characteristics and clinical outcomes for each group. Significance values refer to comparisons of the 2 groups using a 2-sided independent-sample t test, except where footnoted.

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Total (n=549)</th>
<th>BrighterSide (n=274)</th>
<th>Control (n=275)</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Characteristics</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female, n (%)</td>
<td>399 (72.7)</td>
<td>203 (74.1)</td>
<td>196 (71.3)</td>
<td>.84a</td>
</tr>
<tr>
<td>Age (years), mean (SD)</td>
<td>39.1 (13.5)</td>
<td>39.1 (13.5)</td>
<td>39.1 (13.5)</td>
<td>.98</td>
</tr>
<tr>
<td><strong>Actual Help-Seeking Questionnaire</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mental health help sought in past 3 months, n (%)</td>
<td>453 (82.5)</td>
<td>226 (82.5)</td>
<td>227 (82.6)</td>
<td>.54a</td>
</tr>
<tr>
<td><strong>Suicidal Ideation Attributes Scale</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean score (SD)</td>
<td>25.8 (10.1)</td>
<td>25.7 (10.3)</td>
<td>25.8 (10)</td>
<td>.89</td>
</tr>
<tr>
<td>Score ≥21, n (%)</td>
<td>366 (66.7)</td>
<td>178 (64.9)</td>
<td>188 (68.4)</td>
<td>.26a</td>
</tr>
<tr>
<td><strong>Self-harm (%)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Have you ever harmed yourself on purpose? (score &gt;0)</td>
<td>416 (75.8)</td>
<td>208 (75.9)</td>
<td>208 (75.6)</td>
<td>.51a</td>
</tr>
<tr>
<td>In the last six weeks have you harmed yourself on purpose? (score &gt;0)</td>
<td>202 (36.8)</td>
<td>101 (36.9)</td>
<td>101 (36.7)</td>
<td>.54a</td>
</tr>
<tr>
<td>Have you ever attempted to take your own life? (score &gt;0)</td>
<td>291 (53)</td>
<td>152 (55.5)</td>
<td>139 (50.6)</td>
<td>.14a</td>
</tr>
<tr>
<td>In the last six weeks have you attempted to take your own life? (score &gt;0)</td>
<td>38 (6.9)</td>
<td>19 (6.9)</td>
<td>19 (6.9)</td>
<td>.45a</td>
</tr>
<tr>
<td>Functioning and Recovery Scale score, mean (SD)</td>
<td>13.3 (2.4)</td>
<td>13.3 (2.4)</td>
<td>13.4 (2.3)</td>
<td>.55</td>
</tr>
<tr>
<td>Distress Questionnaire-5 score, mean (SD)</td>
<td>19.2 (2.8)</td>
<td>19.3 (2.7)</td>
<td>19.2 (2.9)</td>
<td>.67</td>
</tr>
<tr>
<td>WHO Disability Assessment Schedule (1 item for days out of role) score, mean (SD)</td>
<td>10.1 (8.9)</td>
<td>10.4 (9.2)</td>
<td>9.9 (8.7)</td>
<td>.45</td>
</tr>
<tr>
<td>Abbreviated Coping Orientation to Problems Experienced Inventory score, mean (SD)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Problem-focused score</td>
<td>18.4 (4.9)</td>
<td>18.3 (5.1)</td>
<td>18.4 (4.8)</td>
<td>.79</td>
</tr>
<tr>
<td>Emotion-focused score</td>
<td>27.6 (4.9)</td>
<td>27.7 (5.2)</td>
<td>27.5 (4.7)</td>
<td>.53</td>
</tr>
<tr>
<td>Avoidant-focused score</td>
<td>18.2 (3.6)</td>
<td>18.3 (3.6)</td>
<td>18.1 (3.7)</td>
<td>.69</td>
</tr>
</tbody>
</table>

P value refers to the Χ² or Fisher exact test.

**Primary and Secondary Outcomes**

The mixed effects models for the primary and secondary outcomes are shown in Table 2. The main effect of time was significant for suicidal ideation, functioning and recovery, days out of role, psychological distress, problem-focused coping, and avoidant coping. There were no main effects for condition for any of the measures, and the time by group interaction was only significant for psychological distress. Residuals for SIDAS scores were nonnormal; however, results were identical under a negative binomial MMRM. Therefore, while the primary outcome, suicidal ideation, demonstrated a significant reduction over time, this did not significantly differ between groups (Figure 1). The Cohen d effect size for suicidal ideation between the intervention and control groups from baseline to 6 weeks was d=−0.03, and d=−0.15 from baseline to 12 weeks. Negative effect sizes favor the control group. See Table S2 in Multimedia Appendix 1 for summary of Cohen d effect sizes.
<table>
<thead>
<tr>
<th>Table</th>
<th>Omnibus mixed model repeated measures ANOVA time (baseline; 6 weeks; 12 weeks) x group (BrighterSide; control).</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$F$ test $(df)$</td>
</tr>
<tr>
<td>Suicidal Ideation Attributes Scale</td>
<td></td>
</tr>
<tr>
<td>Time</td>
<td>124.11 (2,223.1)</td>
</tr>
<tr>
<td>Group</td>
<td>0.01 (1,404.3)</td>
</tr>
<tr>
<td>Time x group</td>
<td>0.00 (2,223.1)</td>
</tr>
<tr>
<td>Functioning and Recovery Scale</td>
<td></td>
</tr>
<tr>
<td>Time</td>
<td>25.89 (2,242.7)</td>
</tr>
<tr>
<td>Group</td>
<td>0.02 (1,384.1)</td>
</tr>
<tr>
<td>Time x group</td>
<td>1.18 (2,247.7)</td>
</tr>
<tr>
<td>WHO Disability Assessment Schedule (1 item for days out of role)</td>
<td></td>
</tr>
<tr>
<td>Time</td>
<td>13.69 (2,198.6)</td>
</tr>
<tr>
<td>Group</td>
<td>0.04 (1,375.9)</td>
</tr>
<tr>
<td>Time x group</td>
<td>1.05 (2,198.6)</td>
</tr>
<tr>
<td>Distress Questionnaire-5</td>
<td></td>
</tr>
<tr>
<td>Time</td>
<td>47.04 (2,234.0)</td>
</tr>
<tr>
<td>Group</td>
<td>1.74 (1,396.3)</td>
</tr>
<tr>
<td>Time x group</td>
<td>3.62 (2,234.0)</td>
</tr>
<tr>
<td>Self-harm</td>
<td></td>
</tr>
<tr>
<td>Time</td>
<td>37.29 (2,214.8)</td>
</tr>
<tr>
<td>Group</td>
<td>0.00 (1,411.6)</td>
</tr>
<tr>
<td>Time x group</td>
<td>0.35 (2,214.8)</td>
</tr>
<tr>
<td>Brief-COPE b (problem)</td>
<td></td>
</tr>
<tr>
<td>Time</td>
<td>3.38 (2,221.4)</td>
</tr>
<tr>
<td>Group</td>
<td>0.71 (1,380.3)</td>
</tr>
<tr>
<td>Time x group</td>
<td>1.03 (2,221.4)</td>
</tr>
<tr>
<td>Brief-COPE (emotional)</td>
<td></td>
</tr>
<tr>
<td>Time</td>
<td>0.04 (2,225.7)</td>
</tr>
<tr>
<td>Group</td>
<td>2.11 (1,370.2)</td>
</tr>
<tr>
<td>Time x group</td>
<td>1.01 (2,225.7)</td>
</tr>
<tr>
<td>Brief-COPE (avoidant)</td>
<td></td>
</tr>
<tr>
<td>Time</td>
<td>18.35 (2,225.6)</td>
</tr>
<tr>
<td>Group</td>
<td>1.24 (1,381.4)</td>
</tr>
<tr>
<td>Time x group</td>
<td>2.45 (2,225.6)</td>
</tr>
</tbody>
</table>

<sup>a</sup>Significant at $\alpha=.05$.

<sup>b</sup>Brief-COPE: abbreviated Coping Orientation to Problems Experienced Inventory.
Follow-up analysis of the interaction between time and group in the DQ5 demonstrated a significant difference between the intervention and control groups at 6 weeks ($t_{244.60}=2.68$, 95% CI 0.25-1.67; $P=0.01$), where the BrighterSide group reported lower scores (mean 17.36, SE 0.27) than the control group (mean 18.22, SE 0.25; Cohen $d=0.26$). However, this difference was no longer significant at 12 weeks ($t_{189.61}=1.05$, 95% CI –0.41 to 1.35; $P=0.29$; Cohen $d=-0.01$) (Figure 2).
Adverse Events
Comparison of the frequency of severe suicidal ideation, self-harm, and suicide attempts between the control and intervention groups yielded no significant differences at either 6 weeks or 12 weeks (Table S3 in Multimedia Appendix 1). Given the target population were people experiencing suicidal ideation at baseline, the frequency of adverse events is not remarkable at either follow-up time point.

Attrition
We used a post hoc logistic regression analysis modeling predictors of attrition at 6 weeks and 12 weeks. SIDAS and FRS were chosen as variables to investigate for any potential moderating effects, rather than all surveys, as doing so may
have violated the assumption of no multicollinearity (full results available in Table S4 in Multimedia Appendix 1). The only significant predictors of attrition were group at 12 weeks and the SIDAS \times \text{group} interaction at 12 weeks. The interaction reflects a greater proportion of participants with lower levels of suicidal ideation (SIDAS scores of 0-20) dropping out in the intervention group (77/96, 80\%) compared to the control group (48/87, 59\%).

User Engagement Outcomes
Analytic information regarding the frequency of use of the app was obtained. Overall, 188 intervention participants (n=275, 68.4\% of the baseline intervention group) enrolled into the BrighterSide app, which is the initial onboarding task after
downloading and opening the app. Table 3 presents descriptive statistics on module engagement across the course of the trial.

Table 3. Engagement with activities by module in the BrighterSide app (n=275).

<table>
<thead>
<tr>
<th>Module</th>
<th>Understand Your Thoughts</th>
<th>Prevent a Crisis</th>
<th>Navigate Your Emotions</th>
<th>Navigate Your Thoughts</th>
<th>Plan for the Future</th>
</tr>
</thead>
<tbody>
<tr>
<td>Engaged with ≥1 activity in module</td>
<td>100</td>
<td>36.4</td>
<td>58</td>
<td>21.1</td>
<td>57</td>
</tr>
<tr>
<td>Engaged with ≥50% of activities in module</td>
<td>95</td>
<td>34.5</td>
<td>57</td>
<td>20.7</td>
<td>33</td>
</tr>
<tr>
<td>Completed all activities in module</td>
<td>19</td>
<td>6.9</td>
<td>22</td>
<td>8</td>
<td>8</td>
</tr>
</tbody>
</table>

*a*: number of users.  
*b*: percentage of baseline sample randomised to the intervention group.

In addition to module engagement, we also reviewed engagement with distraction and calming activities. We found that 69 users (25.1% of the intervention group) engaged with at least 1 calming activity and 43 users (15.6% of the intervention group) engaged with at least 1 distraction activity. Given the low percentage of engagement with these activities, they were not considered further.

Use and Acceptability Outcomes

Tables S5 and S6 in Multimedia Appendix 1 present the results of the appropriateness (IAM) and therapeutic alliance (DWAI) measures collected at the 6-week time point. The average of the appropriateness scales was 3.63 (SD 0.89) on a range of 1 to 5, indicating that the participants’ response to the app was positively skewed. This result is consistent with another mental health–oriented app using this same measure, which returned an average of 3.6 [21]; these authors regarded this as an indication of appropriateness. Each question on the DWAI averaged a response between 2.11 and 3.28 across all participants, on a scale of 1 (not at all) to 4 (completely), suggesting that typically, participants indicated they somewhat to mostly agreed with each statement.

Participants were also given a study-specific acceptability survey, where only 44 of 86 responders (51%) agreed the app met their needs, but 73 of those responders (85%) agreed it was easy to use (agreement was determined by a score of 5 or higher on a 7-point scale; see Table S7 in Multimedia Appendix 1 for full results). This survey also included questions regarding the use of the safety plan feature. This showed that 31 of the 86 (36%) responders to this question had filled out the safety plan, and 100% of these did so alone; 81% (25/31) did not share the safety plan with someone else, and the remaining 19% (6/31) did so with one other person. Finally, 58% (18) of those who completed their safety plan looked back on it. For those who did not fill out the safety plan, the most frequent responses when asked why were that they did not feel they had the time; they found it too confronting or overwhelming; they had completed a safety plan elsewhere; or they simply did not know how to navigate to it in the app.

Semistructured Interviews

Of the 69 intervention participants who completed the 12-week follow-up, 6 agreed to be interviewed about their experiences of using the BrighterSide app. Themes and subthemes that emerged throughout the interviews are described in Table S8 in Multimedia Appendix 1. The most notable patterns were that participants found the distraction and mindfulness activities particularly useful, and that the check-in function alleviated the burdensomeness often associated with reaching out to support networks. At the same time, some participants said that while not all of the information in the modules was helpful to them, it may be useful for people who have little or no experience with professional mental health care. Participants felt that the safety plan was useful as a reference tool for when they may be entering a crisis, and that the ability to share this plan with others was helpful since they often are unable to recognize their own warning signs.

Discussion

Principal Findings

The primary objective of this study was to determine if use of the BrighterSide app significantly decreased suicidal ideation at 6 weeks and 12 weeks after baseline compared to a waitlist control group. While there was a significant overall decrease in suicidal ideation from baseline to 12 weeks, there was no difference between the 2 groups. However, the intervention group did report significantly lower psychological distress at 6 weeks compared to the control group, although this difference was no longer significant at 12 weeks. There was no difference
in rates of adverse events (severe suicidal ideation, recent self-harm, or recent suicide attempts), nor were there any significant moderators of demographic variables (age and gender) on attrition.

There may be several possible reasons for the null effect on suicidal ideation. First, there was little engagement with the app. While 68% downloaded and completed onboarding of the app, only 36% of the intervention group engaged with at least one activity in the first module (Understand Your Thoughts; see Table 3). This reflects that participants in the intervention group were not exposed to the full anticipated benefit of BrighterSide. Indeed, only 51% of participants found the app had met their needs. This may be an artifact of the nonprescriptive approach to the intervention. Participants were able to access content in any order they wished, or, if they preferred, use a guided option. If participants were initially exposed to the first module and did not find it helpful, they may have been less inclined to engage with the remaining modules, which may have been more relevant to them. In a similar vein, it could be argued that, given the identified relationship between thwarted belongingness and suicidal ideation [22,23], placing the onus on the participant to choose their method of interaction without the possibility of social connection (ie, no contact with another person) may have been counterintuitive to the aim of reducing suicidal ideation. This may reflect the finding that only half of participants felt the app met their needs, and an investigation into the usefulness of an app with or without a social connection aspect, and the impact of this on app engagement, should be investigated in future research. In any case, the lack of engagement with the app would anticipate a null effect.

Second, a high proportion of participants (453 of 550 participants, over 80%) had recently sought professional help for their mental health, and the strategies provided in the app may therefore have been already known to participants. Third, the trial recruitment period occurred during a period of public health protections in the COVID-19 pandemic. While rates of suicidal ideation were high but stable during the period of this study [24], broader public mental health support during this period may have had a confounding effect on the trial. Finally, despite the extensive co-design process to engage people with lived experience and clinicians, the final intervention may not have achieved an optimal balance between therapeutic content and user engagement. Given the null effect of the BrighterSide app on suicidal ideation, the remainder of this discussion will evaluate the differences between BrighterSide and other digital interventions that have also aimed to reduce suicidal ideation.

Psychological Distress

Participants in the BrighterSide group reported significantly lower distress at 6 weeks than did those in the control group. While this difference was not maintained at 12 weeks, it might indicate that use of the app provided useful tools to navigate psychological distress in the short term. This is noteworthy given that high psychological distress is evidenced to be related to high reports of suicidal ideation and suicide attempts [25]. Ameliorating psychological distress may therefore act as a protective factor, although this is not directly captured by the data. Given that this study indicates the two are not comorbid, further investigation into the relationship between psychological distress and suicidal ideation, with greater power, is warranted.

Comparison of Therapeutic Models

The previous finding that the digital intervention Living with Deadly Thoughts did not significantly reduce suicidal ideation between groups should be discussed in the context of this study [8]. While BrighterSide is modeled on this previous intervention, one of the key differences was the co-design process involved in developing the content for BrighterSide. Incorporating consumers in the design process of these interventions was one method proposed by Torous et al [11] to enhance engagement. Despite this, BrighterSide did not see substantial app engagement, and was rated by almost half of participants to have not met their needs. However, it is known that suicide prevention apps are useful in reducing suicide ideation—for example, the LifeBuoy app saw a significant difference in suicide ideation for the intervention group compared to an active control group [26]. We therefore consider the differences in the therapeutic content involved in these 2 models.

BrighterSide contains modules based on a mix of cognitive behavioral therapy, dialectical behavior therapy, acceptance and commitment therapy, and positive psychology. The amalgamation of different therapeutic elements for a brief intervention such as BrighterSide may have lacked enough adherence to a particular model to see any benefit. Instead, a greater effect on suicidal ideation might be achieved with modules that adhere to one therapeutic model, such as dialectical behavior therapy, since it has been shown to have a great effect on reducing suicidal behavior (see Ougrin et al [27] for a review). The LifeBuoy study implemented an intervention that followed a dialectical behavior therapy model and demonstrated a significant reduction in suicidal ideation when compared to a control group. They also allowed flexibility with module use, but their implementation was prescriptive, so that one module had to be completed in order to unlock the next. Additionally, if there were greater coherence within the modules, perhaps participants would be more inclined to engage meaningfully. Given these results, and the null findings for BrighterSide, it may be beneficial in future research to adhere to a single therapeutic model, such as dialectical behavior therapy, to both enhance delivery of skills and to enhance engagement among participants.

Limitations

The most considerable limitation in this study is the low engagement with the app itself, which inhibits the capacity to adequately assess the primary and secondary outcomes. While maintenance engagement with app use is a recognized issue in digital mental health [11], this study did take into consideration some factors to enhance usability via the co-design method. Additionally, the way the app conveyed information may not have been conducive to participants actually implementing the learned knowledge and skills from the modules, particularly given the brief nature of the app and the amalgamation of components from different therapeutic models. Future research should consider assessing the learnability of skills portrayed in
the modules, to determine whether apps are able to implement behavioral change.

While the interviews with participants were mostly positive, interviews were conducted with participants who self-selected to participate in an interview after completing the 12-week follow-up. Therefore, the small number who self-selected were more likely to have actively engaged in the app, generally had greater motivation for improving mental health research, and may have already felt more positively about the app. Regardless, the interview outcomes were consistent with the results from the surveys on acceptability, functionality, and perceptions of BrighterSide.

**Conclusion**

This study aimed to investigate the ability of an app, BrighterSide, to reduce suicide ideation. While there were no between-condition effects for suicidal ideation, the severity of psychological distress was significantly reduced in the intervention condition after having access to the app for 6 weeks, relative to the control group. Further work may be required to optimally incorporate effective therapeutic content with engaging user design.

**Acknowledgments**

This research was supported by the Australian Government Department of Health–funded National Suicide Prevention Research Fund, managed by Suicide Prevention Australia. The funder had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. We acknowledge that BrighterSide is a registered trademark, even though the ® symbol has been omitted from this paper.

**Authors' Contributions**

NJ, AT, NC, and ML contributed to project management; NJ, PB, and ML contributed to data analysis; NJ, PB, and ML contributed to data interpretation; NJ contributed to draft manuscript preparation and final manuscript preparation; MT, PB, QW, JRB, AT, SH, KH, JR, NC, HC, and ML contributed to study conception and design; MT, SH, KH, JR, NC, HC, and ML contributed to study resources; MT, PB, QW, JRB, AT, HC, and ML contributed to draft revisions; and AT, SH, and ML contributed to data curation. All authors reviewed the final manuscript.

**Conflicts of Interest**

None declared.

**Multimedia Appendix 1**

Complete set of appendices.

[DOCX File, 42 KB - mental_v11i1e55528_app1.docx]

**Multimedia Appendix 2**

BrighterSide mobile app screenshots.

[DOCX File, 1089 KB - mental_v11i1e55528_app2.docx]

**Checklist 1**

CONSORT-EHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile Health Application and online Tele Health) checklist.

[PDF File, 7360 KB - mental_v11i1e55528_app3.pdf]

**References**


Abbreviations

ASHQ: Actual Health-Seeking Questionnaire
BDI: Black Dog Institute
Brief-COPE: Coping Orientation to Problems Experienced Inventory
CONSORT: Consolidated Standards of Reporting Trials
DQ5: Distress Questionnaire-5
DWAI: Digital Working Alliance Inventory
FRS: Functioning and Recovery Scale
IAM: Implementation Appropriateness Measure
MMRM: mixed model repeated measures
RCT: randomized controlled trial
SIDAS: Suicidal Ideation Attributes Scale
WHODAS: World Health Organization Disability Assessment Schedule

---

Edited by J Torous; submitted 23.01.24; this is a non–peer-reviewed article; accepted 31.01.24; published 18.03.24.

Please cite as:
Efficacy of BrighterSide, a Self-Guided App for Suicidal Ideation: Randomized Controlled Trial
JMIR Ment Health 2024;11:e55528
URL: https://mental.jmir.org/2024/1/e55528
doi:10.2196/55528

© Natasha Josifovski, Michelle Torok, Philip Batterham, Quincy Wong, Joanne Beames, Adam Theobald, Sarah Holland, Kit Huckvale, Jo Riley, Nicole Cockayne, Helen Christensen, Mark Larsen. Originally published in JMIR Mental Health (https://mental.jmir.org/), 18.3.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Mental Health, is properly cited. The complete bibliographic information, a link to the original publication on https://mental.jmir.org/, as well as this copyright and license information must be included.
Review

Health Care Professionals’ Views on the Use of Passive Sensing, AI, and Machine Learning in Mental Health Care: Systematic Review With Meta-Synthesis

Jessica Rogan1,2, BSc, MSc, ClinPsy; Sandra Bucci1,2, BSc, ClinPsy; Joseph Firth1, BSc, PhD

1Division of Psychology and Mental Health, School of Health Sciences, Faculty of Biology, Medicine and Health, Manchester Academic Health Sciences, The University of Manchester, Manchester, United Kingdom
2Greater Manchester Mental Health NHS Foundation Trust, Manchester, United Kingdom

Corresponding Author:
Sandra Bucci, BSc, ClinPsy
Division of Psychology and Mental Health, School of Health Sciences, Faculty of Biology, Medicine and Health, Manchester Academic Health Sciences
The University of Manchester
2nd Floor, Zochonis Building, Brunswick Street
Manchester, M13 9PL
United Kingdom
Phone: 44 161 306 0422
Email: sandra.bucci@manchester.ac.uk

Abstract

Background: Mental health difficulties are highly prevalent worldwide. Passive sensing technologies and applied artificial intelligence (AI) methods can provide an innovative means of supporting the management of mental health problems and enhancing the quality of care. However, the views of stakeholders are important in understanding the potential barriers to and facilitators of their implementation.

Objective: This study aims to review, critically appraise, and synthesize qualitative findings relating to the views of mental health care professionals on the use of passive sensing and AI in mental health care.

Methods: A systematic search of qualitative studies was performed using 4 databases. A meta-synthesis approach was used, whereby studies were analyzed using an inductive thematic analysis approach within a critical realist epistemological framework.

Results: Overall, 10 studies met the eligibility criteria. The 3 main themes were uses of passive sensing and AI in clinical practice, barriers to and facilitators of use in practice, and consequences for service users. A total of 5 subthemes were identified: barriers, facilitators, empowerment, risk to well-being, and data privacy and protection issues.

Conclusions: Although clinicians are open-minded about the use of passive sensing and AI in mental health care, important factors to consider are service user well-being, clinician workloads, and therapeutic relationships. Service users and clinicians must be involved in the development of digital technologies and systems to ensure ease of use. The development of, and training in, clear policies and guidelines on the use of passive sensing and AI in mental health care, including risk management and data security procedures, will also be key to facilitating clinician engagement. The means for clinicians and service users to provide feedback on how the use of passive sensing and AI in practice is being received should also be considered.

Trial Registration: PROSPERO International Prospective Register of Systematic Reviews CRD42022331698; https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=331698

(JMIR Ment Health 2024;11:e49577) doi:10.2196/49577

KEYWORDS
artificial intelligence; machine learning; passive sensing; mental health care; clinicians; views; meta-synthesis; review; mental health; health care; health care professionals; psychology; psychiatry; mental health professionals; mobile phone
Introduction

Background

Mental health problems are highly prevalent globally, with approximately 1 in 8 people experiencing mental health difficulties, which can have significant personal and economic consequences [1]. Rapid growth in digital technology innovation has led to an increased interest in digital mental health interventions [2]. Digital tools with built-in sensors, such as smartphones, smartwatches, and other wearable devices, allow for the unobtrusive and continuous collection of objective data, providing insight into user behavior and physiology [3]. Machine learning, which is a branch of artificial intelligence (AI), can be applied to these data to learn from it and generate clinically actionable insights and predictions [4]. It has therefore been suggested that passive sensing data and applied machine learning methods could overcome what some describe as trial-and-error–driven approaches used in mental health care by supporting precise diagnoses and prognoses [5]. Indeed, mental health remains one of the only domains in health care that relies only on service users’ self-report of cognitive and emotional states and symptoms and on clinicians to accurately recognize and map these states to make diagnostic, prognostic, and therapeutic decisions [6]. Passive sensing data and AI may offer a means to overcome the pitfalls of current clinical measures by presenting a more complete picture of a person’s difficulties [7]. For example, raw sensor data captured regarding speech characteristics, location, and activity can be transformed to derive high-level behavioral markers, such as fatigue, sleep disruption, and mood, which can be used to identify clinical states, such as depression [8]. In addition, digital tools that allow for passive sensing can support service users’ self-management of symptoms and access to digitally delivered therapies [4]. Through self-management, service users may feel empowered [9], and service user and clinician access to digital remote data capture has the potential to identify early warning signs of deterioration, providing the opportunity to reduce the risk of relapse of mental health difficulties via early identification and intervention [10]. This may be particularly useful, as current health care systems generally rely on the delivery of treatment by scheduled appointments, which can result in warning signs of mental health relapse being missed or treated too late [11].

Using sensors from digital tools, such as smartphones and wearable devices, to identify clinical and behavioral features of worsening mental health and applying machine learning methods to identify patterns in the data could augment mental health care by delivering more precise treatment at the time it is needed [12].

Despite the potential benefits, there remains a persistent gap between the rapid developments in digitally supported mental health care and the successful adoption of these tools in clinical practice [13]. A key driver to the potential success of digitally supported health care uptake is the willingness, confidence, and capacity of clinicians to make changes to their practice [9]. Resistance to incorporating digital approaches in clinical practice can occur for various reasons, including the lack of technological literacy, fear that AI models could replace professionals, and concerns about ethical and legal issues [6]. There is trepidation that core aspects of clinician roles, such as diagnosis, assessment, formulation and treatment, may be delegated to AI models without human input [14]. This has been viewed as dehumanizing and could have negative implications for therapeutic relationship [15]. Ethical issues have also been raised, such as implications for service user privacy and data security [16]. As clinicians’ perceptions and attitudes pose a potential barrier to implementation [17], it is important that they are invited into the dialog around digitally supported AI in mental health care, to embrace any benefits there might be, as well as share their concerns and explore the limitations and risks [2]. However, it has been noted that stakeholder’s views are rarely considered in model design or evaluation in relation to machine learning approaches [18]. Indeed, professionals have felt that their knowledge and views have been disregarded in the design of digital health solutions or are only considered as an afterthought [19]. As the extent to which these methods can be successfully implemented in health care depends on their acceptability [3], research is needed to understand stakeholders’ perspectives on digital health systems [11].

Objectives

Although there have been some qualitative studies exploring mental health care professionals’ views and experiences of passive sensing and AI in mental health care, there are no published reviews that systematically aggregate these findings, specifically through examining participants’ experiences and perspectives, both deeply (because of the qualitative approach) and broadly (because of the integration of studies from different health care contexts and participants) [20]. This meta-synthesis aims to synthesize and evaluate the relative strengths of the qualitative literature regarding mental health care professionals’ views on the use of passive sensing and AI in mental health care to provide a new, comprehensive interpretation of the findings that goes beyond the depth and breadth of the original studies [21]. Although research continues to grow in this area, it is now an appropriate time to review the literature, as the COVID-19 pandemic has increased the urgency for creating digital interventions that can fulfill the full potential of digital health [22], and it is necessary to engage multiple stakeholder groups early in the design and development process [23].

Methods

Overview

Meta-synthesis is a systematic review and integration of findings from qualitative studies to facilitate the transfer of knowledge and bring together a broad range of participants and descriptions [20]. A systematic approach for identifying and assessing the quality of potential papers, followed by analysis of the data and synthesis, was used with the aim of understanding what mental health care professionals think about the use of passive sensing and AI in mental health care, to embrace any benefits there might be, as well as share their concerns and explore the limitations and risks [2]. However, it has been noted that stakeholder’s views are rarely considered in model design or evaluation in relation to machine learning approaches [18]. Indeed, professionals have felt that their knowledge and views have been disregarded in the design of digital health solutions or are only considered as an afterthought [19]. As the extent to which these methods can be successfully implemented in health care depends on their acceptability [3], research is needed to understand stakeholders’ perspectives on digital health systems [11].

Eligibility Criteria

Eligible papers for this review (1) were peer-reviewed studies published in English that used a qualitative method—mixed
methods studies were also included, but only the qualitative findings were considered; and (2) examined health care professionals’ views on hypothetical or actual use of service user-facing digital tools that use passive sensing and AI in mental health care. Studies with participants that included other stakeholders, as well as health care professionals, were not discounted; however, findings were only included if they were explicitly associated with mental health professionals. There were no limits on the publication year.

**Search Strategy**

A discussion within the research team and a review of the literature allowed for the identification of common terminology used in this research area and the selection of search terms. The search tool “SPIDER” (sample, phenomenon of interest, design, evaluation, research type) was used to ensure that all relevant areas were covered when developing the search terms. Relevant studies were identified through systematic searches of the following electronic databases: AMED, PsycINFO, Embase, and Medline. The search terms were (clinician*) OR (health care professional*) OR (staff) OR (physician*) OR (provider*) OR (practitioner*) OR (psychologist*) OR (doctor*) OR (therapist*) OR (care coordinator*) OR (mental health nurse*) OR (psychiatric nurse*) OR (support worker*) OR (counsellor *) OR (case manager*) OR (GP*) AND (view*) OR (opinion*) OR (perception*) OR (qualitative) OR (interview*) AND (remote monitoring) OR (digital phenotyping) OR (machine learning) OR (passive sens*) OR (passive monitor*) OR (passive data) OR (artificial intelligence) OR (wearables). A manual search of references and citations from eligible articles was also performed by JR to identify additional studies. Papers were initially screened according to title and abstract, followed by a full article.

**Study Selection**

The study selection and exclusion processes were conducted in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [24] in October 2022 and are outlined in Figure 1. Article titles and abstracts were screened for eligibility by JR. If the inclusion criteria were unclear, full-text articles were obtained and reviewed. Any uncertainty regarding study eligibility was resolved through discussion with a wider research team. A second independent rater screened 10% (106/1056) of titles and reviewed 10% (16/154) of full-text articles to assess the reliability of the study selection. There was an “almost perfect” level of agreement between the raters at the screening stage \((k=0.918)\) and at the full-text stage \((k=1)\) [25]. As all studies were published in recent years, the search was conducted again in February 2023. Overall, 10 studies met the eligibility criteria and were included in the review.

**Figure 1.** PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram of systematic search.
Quality Appraisal

Study quality was assessed using the Critical Appraisal Skills Programme (CASP) tool for quality appraisal in qualitative evidence synthesis (CASP, 2018, [26]), which assesses the strengths, limitations, relevance, and credibility of qualitative research. The CASP comprises 10 items that focus on different methodological aspects of qualitative studies, such as method, design, recruitment, data collection, and reflexivity. It is considered a good measure of transparency of research practice and reporting standards and is recommended for use in health-related research [27]; therefore, it was deemed appropriate for use in this review. A 3-point scale was used, with a score applied to each criterion (0=criterion not met, 1=criterion partially met, and 2=criterion totally met) [28]. Therefore, papers were given a total quality score of 20.

The scoring was completed by JR. A second independent rater assessed the quality of 50% (5/10) of the studies, and the scores were compared at the item level. Interrater reliability estimates showed good agreement between raters ($k$=0.832) [25]. Disagreements in ratings were resolved through discussion among raters until agreement was reached.

Data Extraction and Synthesis

The included studies were read and reread to ensure that they met the inclusion criteria. Key study information was then recorded, including the number and characteristics of participants, aims of the research, analysis method used, and the settings (Table 1). In addition, the original authors’ analysis of primary qualitative data was extracted (second-order constructs), and individual participants’ quotes were also noted (first-order constructs), in line with meta-synthesis principles [29]. An inductive thematic analysis approach within a critical realist epistemological framework was then taken with the aim of developing a cohesive, synthesized understanding of the data [30] and new interpretations [31]. JR completed the coding of the text and quotes using NVivo qualitative data analysis software (NVivo version 12, Lumivero). The constructs were then grouped into core themes. These themes were discussed by a broader research team, considering how each paper contributed to each core theme. The themes were then grouped into final higher-order themes, which were again reviewed and agreed upon by the research team. These themes are considered third-order constructs and allow for reflection on how each paper’s findings fit within the wider literature and for findings to extend beyond the original papers [21]. JR returned to the papers to ensure that the themes identified reflected the data and that other themes were not overlooked.
<table>
<thead>
<tr>
<th>Study</th>
<th>Participants</th>
<th>Aim</th>
<th>Data analysis</th>
<th>Setting</th>
<th>Results and themes</th>
<th>Critical Appraisal Skills Programme quality appraisal score (out of 20)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Greer et al [32] (the United Kingdom)</td>
<td>Five focus groups, made up of mental health nurses (N=25); age range 22-64 y; mean age 42.7 y; male: n=9; female: n=16</td>
<td>To explore staff views, specifically benefits and barriers to using remote monitoring to predict risk of inpatient aggression</td>
<td>Thematic analysis</td>
<td>Inpatient forensic mental health service, the United Kingdom</td>
<td>Utility in clinical practice, risk to user safety and well-being, data security and privacy, impact on staff workload, engagement, and adherence</td>
<td>17</td>
</tr>
<tr>
<td>Ng et al [33] (the United States)</td>
<td>Interviews with mental health professionals (N=17); age and gender not stated</td>
<td>Explore opportunities and barriers mental health staff perceive in applying sensor-captured patient-generated data among populations with PTSD in routine care settings and how providers perspectives inform the design of tracking systems and strategies to implement technologies</td>
<td>Thematic analysis</td>
<td>Centre for Veterans With PTSD, the United States</td>
<td>Patient-driven uses of Fitbit and its data; integrating Fitbit data into treatment protocols; challenges to the use in treatment</td>
<td>18</td>
</tr>
<tr>
<td>Blease et al [34] (global)</td>
<td>Web-based survey of psychiatrists (N=791); age range 25≥65 y; mean age group 35-44 y; male: n=550; female: n=230; other: n=11</td>
<td>To explore psychiatrists' opinions about the potential impact of innovations in AI and machine learning on psychiatric practice</td>
<td>Qualitative descriptive analysis of written responses</td>
<td>Web-based survey across 22 countries</td>
<td>Patient-psychiatrist interactions, quality of patient medical care, profession of psychiatry, health systems</td>
<td>18</td>
</tr>
<tr>
<td>Thenral and Anna-malai [35] (India)</td>
<td>Interviews with psychiatrists (n=14), patients (n=14), technology experts (n=13), and chief executive officers (n=5); overall (N=46); psychiatrist characteristics: mean age 35.5 y; male: n=7; female: n=7</td>
<td>To understand the perceived challenges in building, deploying, and using AI-enabled telepsychiatry for clinical practice</td>
<td>Grounded theory</td>
<td>Practices in urban areas of India; Chennai, Mumbai, Bangalore, and Delhi</td>
<td>Knowledge and gaps deficit; attitudes and perception; data challenges; ethical, legal accountability; AI related; health system infrastructure; human resources and skills; technology; clinical practice</td>
<td>15</td>
</tr>
<tr>
<td>Dawoodbhoy et al [36] (the United Kingdom)</td>
<td>Interviews with health care professionals (n=9) and AI experts (n=11); overall (N=20); age and gender of health care professionals not stated</td>
<td>To identify issues in patient flow on mental health units and align them with potential AI solutions, ultimately devising a model for their integration at service level</td>
<td>Thematic analysis</td>
<td>Acute mental health inpatient units, the United Kingdom</td>
<td>Current mental health inpatient service and patient flow model: patient factors; problems with social care; problems with clinical management; problems with inpatient service and system; solutions</td>
<td>16</td>
</tr>
<tr>
<td>Study</td>
<td>Participants</td>
<td>Aim</td>
<td>Data analysis</td>
<td>Setting</td>
<td>Results and themes</td>
<td></td>
</tr>
<tr>
<td>-------------------------------------------</td>
<td>-------------------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------</td>
<td>-----------------------------------------------</td>
<td>----------------------------------------------</td>
<td>---------------------------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>Rodriguez-Villa et al [23] (the United States and India)</td>
<td>Focus groups and interviews with mental health clinicians (n=53) and service users and their families (n=75); overall (N=128); clinician characteristics: age range 23-72 y; mean age 36 y; gender not stated</td>
<td>To engage clinicians and people living with schizophrenia spectrum disorders and their family members from 3 study sites distinct in culture and setting in developing new features and co-designing the mindLAMP app</td>
<td>Thematic analysis</td>
<td>Mental health services in the United States and India</td>
<td>COVID-19 led to an uptake of virtual therapy, presenting new challenges and opportunities for providers and people with schizophrenia, using technology; access to data may offer providers and people with schizophrenia new insight into illness and treatment, but too much data elicit discomfort; relevance and integrated experience increase engagement</td>
<td></td>
</tr>
<tr>
<td>Byrne et al [10] (Australia)</td>
<td>Focus groups with service users (n=12) and interviews with mental health clinicians (n=10); overall (N=22); clinician characteristics: age range 25-60 y; male: n=2; female: n=8</td>
<td>Explore patient and clinician-related acceptability of an mHealth device to monitor stress for severe mental illness</td>
<td>Thematic analysis</td>
<td>Community youth mental health service, Australia</td>
<td>Self-monitoring improves insight; clinician monitoring as a benefit to treatment; privacy and data misuse concerns; ease of use; engaging design; procedural guidelines</td>
<td></td>
</tr>
<tr>
<td>de Angel et al [3] (the United Kingdom)</td>
<td>Three focus groups: 2 focus groups with patients (n=16) and 1 focus group with mental health clinicians (n=6); overall (N=22); characteristics of clinicians: mean age 36.7 y; male: n=1; female: n=5</td>
<td>To identify clinically meaningful targets for digital health research and to explore patient and clinician attitudes toward the use of remote monitoring technologies and identify any perceived barriers to and facilitators of using these methods in psychological treatments for depression</td>
<td>Thematic analysis</td>
<td>Improving Access to Psychological Therapies services, the United Kingdom</td>
<td>Promoters of change (internal and external); markers of change (internal and external)</td>
<td></td>
</tr>
<tr>
<td>Gotzl et al [37] (Germany)</td>
<td>Two focus groups with young people (n=8) and interviews with experts, including psychologists (n=2); overall (N=13); age and gender of psychologists not stated</td>
<td>To investigate the subjective needs, attitudes, and preferences of key stakeholders toward an AI-informed mHealth app</td>
<td>Mixed methods-only qualitative component considered: content analysis</td>
<td>This study formed part of the living laboratory &quot;AI4U: Artificial Intelligence for personalized digital mental health promotion and prevention in youth&quot;</td>
<td>Young peoples’ understanding of mental health; experts understanding of mental health in youth; opportunities and risks seen by experts; experts’ recommendations</td>
<td></td>
</tr>
</tbody>
</table>
Results

Summary of Papers

A total of 10 papers were deemed eligible for inclusion in this review. A total of 3 studies were conducted in the United Kingdom, 1 in India, 1 in the United States, 1 across both the United States and India, 1 in Australia, 2 in Germany, and 1 in a global study. In total, 6 studies used thematic analysis, 1 used a grounded theory approach, 1 used qualitative descriptive analysis, and 2 used content analysis. Participants in 4 of the papers were health care professionals only, with the remaining 6 papers including health care professionals as well as other stakeholders, such as service users and their families, technology experts, and technology company owners. The findings were only included if they were explicitly associated with health care professionals. The number of health care professionals ranged from 2 to 53 (mean 17). The age of the mental health professionals where this was reported (6 papers) ranged from 22 to 72 years. Among the 5 papers that reported gender, 28 participants were male, and 42 were female. Owing to the high number of participants in the global web-based survey [34], these data are described separately, with 791 participants taking part, ranging in age from 25 to ≥65 years. Of the participants, 550 identified as male, 230 identified as female, and 11 identified as others.

Study Quality

The overall CASP quality appraisal scores are included in Table 1, and a breakdown of these scores is provided in Table 2. There was variation in the scores across the papers. Those given stronger scores tended to provide more detail as to why certain qualitative approaches were selected over others, provided details regarding the sample including why participants may have opted not to take part, and ethical considerations were reported. It should be noted that although some studies did make some reference to the relationship between researcher and participants, this was the area that scored lowest, with few studies referencing bias and considering the influence their own roles may have had on results and reporting. For papers that included other stakeholders alongside mental health professionals, higher scores were given if the results were written to clearly distinguish which themes were associated with which participant group.

Table 2. Quality ratings on each of the Critical Appraisal Skills Programme domains.

<table>
<thead>
<tr>
<th>Study</th>
<th>Aims</th>
<th>Method</th>
<th>Design</th>
<th>Recruitment</th>
<th>Data collection</th>
<th>Bias and reflexivity</th>
<th>Ethical issues</th>
<th>Data analysis</th>
<th>Clear findings</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Greer et al [32]</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Ng et al [33]</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Blease et al [34]</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Thenral and Annamalai [35]</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Dawoodbhoy et al [36]</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Rodriguez-Villa et al [23]</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Byrne et al [10]</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>de Angel et al [3]</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Gotzl et al [37]</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Reis and Maier [38]</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>
Findings

Analysis of the data revealed three distinct but interrelated themes: (1) the use of passive sensing and AI in clinical practice, (2) barriers to and facilitators of use in practice, and (3) consequences for service users. A total of 5 subthemes were identified from the data. The themes, subthemes, and relationships between them are summarized in Figure 2.

Figure 2. Themes, subthemes, and the relationships between them. AI: artificial intelligence.

Theme 1: Uses of Passive Sensing and AI in Clinical Practice

Findings across the reviewed papers included how clinicians felt they could use passive sensing and AI in their practice. Passive sensing technology has been shown to be particularly useful, as the unobtrusive collection of objective service user data could offer a new information source [23]. This could provide insight into factors such as speech, social media use, and activity levels, which could be considered alongside self-report questionnaires and assessment tools to facilitate more accurate assessment of service users' mental health difficulties [35]. In assessing service user needs, it was also felt that these data could clarify discrepancies between self-report, observation, and psychometrics and validate service users’ concerns [33]. It has been suggested that passive sensing data and AI could entirely replace some methods of assessment, such as questionnaires, to improve the clinical experience for both service users and clinicians [3,38]. Passive sensing data can reduce errors and biases in clinical decisions regarding diagnosis, treatment, and medication [34], as data-driven technologies may uncover correlations that humans cannot [36]:

*The benefits would be greater reliability in diagnosis and prognosis, being able to choose specific customized treatment plans after analysis.* [Participant in Blease et al [34]]

Further suggestions were made as to how passive sensing and AI could be useful in therapeutic work, such as guiding productive discussions [33], setting treatment goals, delivering low-intensity support [3], tracking the efficacy of brief interventions [23], and encouraging ongoing engagement and regular self-reflection [38]. Furthermore, discussions were conducted about how AI’s ability to process, connect, and make conclusions from large amounts of data could be used to risk-stratify service users according to their personal factors and needs [36] and support identification and awareness of early warning signs, thus reducing the risk of relapse of mental health difficulties [32-34,37]. As clinicians have access to these data, it was also felt that they could identify when to intervene [38], which may further reduce a service user’s risk of deterioration in mental health [10]. Indeed, clinicians have reported that seeing a change in the data regarding a service user’s speech and self-care habits would promote awareness of a decline in their well-being [3]. This was considered useful in community and ward environments, where staff may not always have eyes on service users [32], particularly for those who may lack insight into their difficulties or do not volunteer information themselves [23,32]:

*...not all our patients will be able to say, “oh well I feel agitated” or be able to come out and say it, but within themselves all the physical, you know, changes are taking place so I think it’s good, it will help us to see the covert, you know, things that are not outward that the patients cannot express.* [Participant in Greer et al [32]]

The idea that aspects of psychiatric work could gradually be replaced by AI was viewed as positive progression by some, but others were concerned that overreliance on AI and...
technology in practice may result in staff becoming deskilled [34]:

_May lead to less skilled mental health staff._ [Participant in Blease et al [34]]

**Theme 2: Barriers and Facilitators to Use in Practice**

**Overview**

Throughout the papers, participants discussed the perceived barriers to and facilitators of using passive sensing and AI in mental health care. The barriers discussed included access, concerns about clinicians’ workloads, and the potential negative impact on the therapeutic relationship. Facilitators included ease of use and training.

**Subtheme 2.1: Barriers**

**Access**

It was highlighted that technology is now readily available, and this was reported as a benefit to using passive sensing and AI in mental health practice as it may improve service user access to mental health care [34]. However, not all mental health services have sufficient access to technology [37] because of factors such as cost and the lack of the necessary infrastructure to support digital tools [3]. For example, participants reported that in India, most hospitals do not have access to the internet [35], and service users do not always have access to smartphones [23]. This would likely present significant barriers to health care professionals using such technology in mental health care. Therefore, it is important to consider the digital divide, that is, the gap between those who benefit from the digital age and those who do not [3]:

_All these devices, technology, AI, etc., require high-speed internet...Majority of the hospitals do not have internet...patients who have basic livelihood issues cannot afford a device or internet._ [Participant in Thenral and Annamalai [35]]

**Increased Workload**

A further barrier discussed was the impact of passive sensing and AI could have on clinicians’ workloads. Clinicians wondered about the amount of time and effort required to incorporate data flows into their practice and whether they would be required to review data before sessions [33], which could result in clinicians trawling through a significant amount of data to generate actionable insights [3]. Indeed, participants reported feeling “overwhelmed” when presented with passively collected data [23]:

_I feel overwhelmed with the data to begin with._ [Participant in Rodriguez-Villa et al [23]]

Queries were also raised around documentation and whether the use of these tools would increase administrative work for clinicians [33,34]. Furthermore, it was suggested that clinicians may have to spend time with service users reviewing the use of devices and verifying data, which could take time away from evidence-based practices. If clinicians are alerted to changes in behavior that are the result of inaccurate readings, this could cause unnecessary alarm and waste clinicians’ time [33]:

_There’s always so much to do...You’re already kind of preparing for sessions, and at a certain point it’s like, “How much am I treating or assessing what I’m seeing on a screen or on paper compared to just talking to someone and figuring stuff out together?”_ [Participant in Ng et al [33]]

Managing risk was another concern raised, with participants wondering about their clinical responsibility for monitoring the data for risk issues [3,10,35] because responding to constant data streams would not be possible [23]. This is important because risk aversion is cited as a potential barrier to engagement [36].

In contrast, it has been suggested that passive sensing and AI have the potential to multiply resources, in that it offers a means of support to service users when health care professionals are unavailable [38]:

_Having a system that treats people would be awesome. We cannot be everywhere, and the number of mental health professionals is too low._ [Participant in Reis and Maier [38]]

**Impact on the Therapeutic Relationship**

An issue that arose across studies was the impact that use of passive sensing and AI in practice could have on the therapeutic relationships. Some service users may prefer in-person consultations [35]; therefore, using digital tools and AI methods as a replacement of human contact could be detrimental to the therapeutic alliance [3] because of a loss of empathy and inaccurate interpretations of service users’ presentations [34]:

_Psychiatry is incompetent and incomplete without empathy. I doubt a machine could ever empathise with a live human being...I don’t think affect of patient and mood, feelings, emotions can be analysed accurately._ [Participant in Blease et al [34]]

A potential lack of meaningful interactions between service users and clinicians [33] has led some to believe that service users may become resistant to or refuse treatment [34]. It was also suggested that, as service users are less accountable to clinicians, this could negatively impact motivation [3]. Furthermore, service users may become reliant on a device during treatment, and having this subsequently removed could have negative repercussions, including service users becoming mistrusting of services [3]. It was also highlighted that clinicians may not be able to fully trust the data they receive, as participants suggested that service users may influence these data deliberately [34]:

_If a patient simulates a disease AI might not be able to determine it._ [Participant in Blease et al [34]]

Having said this, it was proposed that allowing service users to submit data to clinicians, who could then respond with recommendations, would enable remote support and continuity of care, which could strengthen the therapeutic relationship [23]. It appeared that the general consensus was that although digital tools may enhance practice, they should not replace service user or staff interactions, something which was viewed as integral to the therapeutic relationship [32].
Subtheme 2.2: Facilitators

Ease of Use
Throughout the studies, it was highlighted that to improve engagement with digital tools that allow for passive sensing and applied AI methods, the technology and systems would have to be relatively straightforward to use in terms of accessibility and convenience [3]. Health care professionals discussed that clinicians are very busy and would not have time to navigate complicated systems [10]:

I think it needs to be relatively simple...not over complicated and very easy to navigate. [Participant in Byrne et al [10]]

To ensure ease of use, suggestions were made, such as including relevant stakeholders in the development of such technologies and related systems to ensure that they use accessible language [37] and presenting the data in a simple way that is easy to understand [3]:

Being able to have a really simple, easy way to compare the progress throughout the weeks of treatment. So, you would obviously be collecting a huge amount of data but if there was a way that we could somehow get, “Okay, you did an average of X amount of steps in week one, and your average sleep was X amount with waking up X amount of times.” [Participant in de Angel et al [3]]

Training
The importance of training to support clinicians in using digital tools, passively collected data, and applied AI methods in practice has been emphasized in most studies. This was discussed in the context of being given time to access the training as well as time to consider implementation [3]:

As long as we’ve had adequate training...And it’s not just having the training, it’s then having the time to think about that afterwards and incorporate it into your practice which would require a corresponding decrease in clinical work. [Participant in de Angel et al [3]]

Clinicians may differ in technological literacy, and some may generally find technology challenging [35]. It was discussed that this technology will only be useful if clinicians understand it and feel comfortable using it [23]. For training to be adequate, it was suggested that clinicians would value clear procedures and guidance on when and how these digital technologies should be used [10], how to connect the data to their established clinical practice [33], and how to interpret data. Depending on the condition, certain markers of behavior could be interpreted positively or negatively [3]. Clinicians would also require clear guidelines regarding responsibility, interoperability, information governance, and potential risks [36]:

You’re going to need [to] train them on how to balance all of these different pieces of data they have access to and how to prioritize the data. I think it would be especially important for new therapists coming on. It would probably be pretty overwhelming for some to have access to that much data and I think we would need to do like a standard operating procedure of how to [deal with] the information. [Participant in Ng et al [33]]

Clinicians would further benefit from being informed of the evidence base around passive sensing and AI in mental health care [33], especially as the belief that there is a lack of studies to support the use of AI technology in health care could be a barrier to clinician engagement [35]:

There is also a lack of well-established trials and studies to understand the applicability of AI-related technology. They build solutions with no real-time applications. [Participant in Thenral and Annamalai [35]]

Theme 3: Consequences for Service Users

Overview
Throughout the studies, findings on the consequences that passive sensing and AI in mental health care could have for service users were discussed. There appeared to be a positive notion that this could empower service users, although it was also acknowledged that there could be risks to service users’ well-being. Concerns have also been raised regarding the protection and safety of service users’ data.

Subtheme 3.1: Empowerment
It was suggested that passive sensing and AI could facilitate “knowledge transfer” and empower service users to understand how their actions, feelings, and thoughts are intertwined [37]. By increasing insight into mental health, self-monitoring allows service users to respond to symptoms and take action themselves [10]. Service users’ monitoring and managing their mental health may involve connecting with other users and supporting one another, setting reminders to take medication, and responding to prompts to engage in helpful strategies [3]. Thus, service users’ ability to monitor their mental health-related data can be empowering [33]:

A lot of the thoughts you have are that you are incapable, inadequate, cannot accomplish things. So, that [data] kind of directly speaks against that, right? “I am able to accomplish something, like reaching 10,000 steps a day.” [Participant in Ng et al [33]]

This self-management could increase awareness of early warning signs, reduce the risk of relapse, and therefore decrease demand for services, for example, by reducing hospital admissions [36].

Subtheme 3.2: Risk to Well-Being
In contrast, concerns have been raised about the accuracy of sensing technology, as making health care decisions based on unreliable sensors could potentially be harmful [3]. In addition, if service users have access to their health data, this could cause some to become hyperfocused on their data, catastrophize, or become disheartened by lack of progress or negative trends [33]. This was thought to be particularly pertinent to those who experience health anxiety [3] or paranoia, as “tracking” behavior could exacerbate symptoms [32]:

When you give this to a paranoid patient, they will think you are monitoring them. It will be so difficult
Subtheme 3.3: Data Privacy and Protection Issues

Participants reported that in practice, they will often recommend apps to service users without reviewing privacy policies, citing a lack of time as the reason for not investigating this further [23]. However, in most studies, concerns have been raised regarding privacy in relation to passive sensing data. It has been suggested that the collection of personal data through digital devices that allow passive sensing could increase the risk of loss of confidentiality and misuse of data [34,36], which could negatively impact therapeutic relationships [10]. In line with this, it was felt that service users would have less control over what they chose to share, which may feel uncomfortable for service users and lead clinicians to feel as though they are invading their privacy [10,23]:

...it does feel like it is personal information and to share all the details about their sleep and their activity levels—that could be quite tricky for some of them to share openly and knowing that we kind of can access it without them knowing or without them being there. [Participant in Byrne et al [10]]

Data management was therefore seen as an important consideration, and it was highlighted that service users should be given choice over what they share and be made aware of who could access the data, what will happen if their data are leaked [3], how their data will be kept private and secure, and what their data will be used for [10]:

I would see a risk if the app claimed: “if you go through these ten steps...then you are a different person” [laughs]...A good app would be characterized by the fact that the user does not internalize a problem centred perspective, but that he...gets the feeling: “I am okay.” [Participant in Gotz et al [37]]

Another concern highlighted by participants working in mental health wards was that service users could harm themselves using a digital device. For example, if an armband that stretches it could be used as a ligature, the design of devices is an important consideration [32]:

...how far does it stretch, can you put it round your neck? Well, that might be an issue, you know, ligatures. [Participant in Greer et al [32]]

...should be something that they cannot use as a weapon, like, there shouldn’t be any metal or something that they can use to self-harm. [Participant in Greer et al [32]]

I think it’s about having a conversation with the client at the beginning about boundaries really, and once it’s clearer, how the information will be shared and can be shared, then you can...Normally put in those boundaries and then you can understand those concerns. [Participant in de Angel et al [3]]

It is important to ensure that service users have capacity when making these decisions, as mental states can change and influence decision-making [32].

um, making sure they understand completely, cos some people are more paranoid on days...than other days so it could be they’re fine for 5 days then the sixth day they’re really paranoid. [Participant in Greer et al [32]]

Discussion

Principal Findings

Across the papers reviewed, multiple ways in which passive sensing technology and applied AI methods could augment mental health care were identified, such as supporting service users in managing their mental health, improving diagnostic accuracy, monitoring treatment trajectories, and increasing access to timely support, thereby reducing the risk of relapse of mental health difficulties. Indeed, research has shown that passive data and AI methods have the potential to provide insight into service user behavior outside the clinic environment and provide real-time detection of behavioral anomalies, which could allow early identification and intervention before a deterioration in mental health [39]. However, despite the potential benefits, concerns have been raised that clinicians could become overreliant on digital technology in practice [40]. This could have negative consequences, as participants discussed that they may not be able to fully trust the data they receive because of service user influence and inaccurate sensors. Therefore, overreliance on inaccurate data can lead to misdiagnoses or missed diagnoses. Thus, decision-making should not be delegated to technology alone [41], and it is important for clinicians to acknowledge the limitations of objective data collection and applied AI methods to avoid tension between service users and clinicians [42]. This is particularly important, as research has shown that discrepancies between experience and tracking data can lead to upset, confusion, and disengagement [43], which may negatively impact the therapeutic relationship.

The influence that the use of passive sensing and applied AI methods could have on the therapeutic relationship was further discussed across papers. Although service users should feel empowered to make choices and manage their own mental health, access to human in-person support is deemed necessary. This reflects concerns that the use of AI in health care could lead to neglect of the therapeutic aspects of in-person consultation, such as consideration of motivation and self-advocacy, attendance to nonverbal cues, and social connection that can be provided by in-person clinical contact [44]. Fears were further raised that the absence of a therapeutic relationship may lead service users to disengage or refuse mental health care altogether. Research suggests that a therapeutic...
alliance can exist between a person seeking change and a change agent, which does not necessarily have to be a human health care professional, with digital tools and apps themselves having the potential to act as change agents [45].

Concerns have been raised across studies that service users may notice a decline in their mental health if they were to monitor aspects of their behavior and interpret subsequent passively collected data in such a way that increases anxiety or results in demotivation. Research has shown that tracking behavior can reduce enjoyment in walking-based activities [46] and increase eating disorder symptomatology [47]. However, research has also found that the use of digital devices that allow passive sensing, such as wearables, can be a positive experience, with multiple psychological benefits identified by users, including increased sense of motivation and accountability [48]. Individual differences are therefore important for clinicians to consider, as certain characteristics may impact a service user’s ability to interpret their health data in a helpful way. For example, research has shown that high health literacy supports the understanding of passively collected health data and how to use it to work toward goals [49].

Across studies, clinicians discussed the impact that use of passive sensing and AI could have on their workload. Concerns appeared to be around reviewing significant amounts of data to identify clinically relevant information and risk monitoring. However, previous research has suggested that AI may in fact reduce clinicians’ workloads, as less time will be required to read through notes to understand a service user’s history, particularly because certain AI methods, such as natural language processing, could be applied to patient notes to summarize important information [50]. Furthermore, machine learning methods can facilitate work by highlighting previously inaccessible or less understood symptoms and patterns [6]. It has also been suggested that data received by clinicians regarding a service user’s behavior may allow them to identify those most in need of support and prioritize their workload, thus using their time more effectively [51]. To reduce concerns about increased workload, it would be useful for clinicians to receive data in a user-friendly format, allowing seamless access to relevant information. If devices and associated systems are not considered user-friendly and there are multiple technical issues, this will likely result in frustration and reluctance to engage [52]. Along with ease of use, training was discussed as a means to encourage clinicians to engage with devices that allow passive sensing and applied AI methods in their practice. Ways to make training useful for clinicians included ensuring that clinicians have access to clear guidance around incorporating data flows into their practice, managing risk issues, and data privacy and protection procedures. The latter is especially pertinent, as concerns about data security were a reoccurring theme throughout studies. Transparent guidelines will need to be developed, and codes of practice enforced around storage, ownership, and sharing of data [52]. However, it has been suggested that concerns about confidentiality of data may always remain; therefore, to facilitate engagement, the perceived value to clinicians and service users will need to outweigh these concerns [53]. As discussed in the reviewed studies, training should involve increasing awareness of the evidence base so that clinicians can understand the cost-benefits of engaging in passive sensing and AI in practice.

The final key issue is access. As highlighted in this review, access to technology could pose a barrier to engagement at both the service user and clinician level. For example, studies conducted in India have highlighted that not all hospitals offering mental health care have access to the internet. Therefore, considering the digital context within low- and middle-income countries, it is important to create digital-based mental health interventions intended for a global rollout. Indeed, Lee et al [54] highlighted that methods such as machine learning have the potential to advance health equity by supporting opportunities for equality in patient outcomes, performance, and resource allocation.

**Strengths and Limitations**

Meta-synthesis allows for greater scope and generalizability than individual primary studies [55]. However, as data are transposed into third-order constructs, there is potential for the findings to move away from the empirical, conceptual, and theoretical contexts of primary qualitative studies [56]. Of the papers included, 2 used content analysis, which is a more descriptive approach to coding and data interpretation (Vaismoradi et al [57]). Thus, the findings may have been more heavily influenced by studies that used more robust qualitative methods, such as thematic analysis, which can provide a more detailed and nuanced account of the data (Braun and Clarke [58]). Furthermore, the process and methods of meta-synthesis are heavily influenced by the focus and expertise of the authors, meaning that some concepts and theories may not have been considered. This limitation was managed through discussion with the research team on coding and themes as well as remaining attuned to personal perspectives that could introduce bias [59].

Efforts were made to include all eligible studies in this review and to avoid neglecting potentially important findings, such as checking the reference lists of all papers and searching the databases again at a later date to identify further studies that might have been published. However, it is possible that some studies were overlooked, particularly as the terminology in this research area can be diverse and studies were only included if they were published in the English language in peer-reviewed journals, meaning that important contributions to the literature may have been missed because of language and publication bias. The included studies were conducted across different mental health settings, such as primary care and inpatient settings, and across different countries. It is important to note that the health care systems and services within them differ globally, so the generalizability of the results may be limited. However, meta-synthesis of qualitative studies can transform findings into highly abstracted and generalizable theoretical frameworks [21].

**Future Directions**

Considering the findings from this review and wider research in this area, a key barrier to implementing digital technology innovations is end user perceptions rather than technology innovation itself [3]. Therefore, it will be important for future...
research to gain a deeper understanding of service user views as well as other stakeholders, such as policy makers. Further research into the efficacy of passive sensing and AI in mental health care is necessary to build an evidence base that would support the scaling up of these approaches to routine service delivery. Real-world studies implementing passive sensing and AI in practice are needed to understand the contextual factors that impact uptake, which will be useful to gain knowledge that can support the development of implementation frameworks [60].

**Clinical Implications**

These findings suggest that although clinicians are open-minded about the use of passive sensing and applied AI methods in mental health care, factors such as service user well-being, clinicians’ workloads, and the therapeutic relationship need to be considered. It is important to involve both service users and clinicians in the development of digital technologies and systems to ensure their ease of use. The development of policies, training, and clear guidelines on the use of passive sensing and AI in mental health care, including risk management and data security procedures, will also be key to facilitating clinician engagement and wide-scale adoption. Means for clinicians and service users to provide feedback on how the use of passive sensing and AI in practice is being received should also be considered, allowing reflection on any impact there might be on the therapeutic relationship, service user well-being, and clinicians’ workloads.

**Acknowledgments**

JF is supported by a UK Research and Innovation Future Leaders Fellowship (MR/T021780/1). SB is supported by the National Institute for Health and Care Research Professorship (NIHR300794) and acknowledges the support from the Manchester Biomedical Research Centre (NIHR203308).

**Conflicts of Interest**

SB is a Director and shareholder of CareLoop Health Ltd, which develops and markets digital therapeutics for schizophrenia and a digital screening app for postnatal depression. SB also reports research funding from the National Institute for Health and Care Research and the Wellcome Trust.

**References**


26. CASP Checklists. URL: https://casp-uk.net/casp-tools-checklists/ [accessed 2024-01-19]


Abbreviations
- **AI**: artificial intelligence
- **CASP**: Critical Appraisal Skills Programme
- **PRISMA**: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
Effectiveness of Online and Remote Interventions for Mental Health in Children, Adolescents, and Young Adults After the Onset of the COVID-19 Pandemic: Systematic Review and Meta-Analysis

Linda Fischer-Grote¹,², BSc, MSc; Vera Fössing³, BSc, MSc; Martin Aigner¹,⁴, Dr med, PhD; Elisabeth Fehrmann¹, MSc, PhD; Markus Boeckle³, MSc, PhD

¹Department of Psychology and Psychodynamics, Karl Landsteiner University of Health Sciences, Krems, Austria
²Department of Clinical Psychology and Psychotherapy, University Hospital Krems, Krems, Austria
³Research Centre Transitional Psychiatry, Karl Landsteiner University of Health Sciences, Krems, Austria
⁴Department of Psychiatry for Adults, University Hospital Tulln, Tulln, Austria

Corresponding Author:
Elisabeth Fehrmann, MSc, PhD
Department of Psychology and Psychodynamics
Karl Landsteiner University of Health Sciences
Dr.-Karl-Dorrek-Straße 30
Krems, 3500
Austria
Phone: 43 6503032923
Email: elisabeth.fehrmann@kl.ac.at

Abstract

Background: The prevalence of mental illness increased in children, adolescents, and young adults during the COVID-19 pandemic, while at the same time, access to treatment facilities has been restricted, resulting in a need for the quick implementation of remote or online interventions.

Objective: This study aimed to give an overview of randomized controlled studies examining remote or online interventions for mental health in children, adolescents, and young adults and to explore the overall effectiveness of these interventions regarding different symptoms.

Methods: A systematic literature search was conducted according to PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analysis) guidelines using PubMed, PsycInfo, Psynex, Embase, and Google Scholar. A meta-analysis was conducted using a random effects model to calculate overall effect sizes for interventions using standardized mean differences (SMDs) for postintervention scores.

Results: We identified 17 articles with 8732 participants in the final sample, and 13 were included in the quantitative analysis. The studies examined different digital interventions for several outcomes, showing better outcomes than the control in some studies. Meta-analyses revealed significant medium overall effects for anxiety (SMD=0.44, 95% CI 0.20 to 0.67) and social functioning (SMD=0.42, 95% CI −0.68 to −0.17) and a large significant effect for depression (SMD=1.31, 95% CI 0.34 to 2.95). In contrast, no significant overall treatment effects for well-being, psychological distress, disordered eating, and COVID-19–related symptoms were found.

Conclusions: The qualitative and quantitative analyses of the included studies show promising results regarding the effectiveness of online interventions, especially for symptoms of anxiety and depression and for training of social functioning. However, the effectiveness needs to be further investigated for other groups of symptoms in the future. All in all, more research with high-quality studies is required.

(JMIR Ment Health 2024;11:e46637) doi:10.2196/46637

KEYWORDS
COVID-19 pandemic; online/digital mental health intervention; e-mental health; anxiety; social functioning; depression; well-being; psychological distress; eating disorder; COVID-19 symptoms
**Introduction**

The high prevalence of psychological disorders in children and adolescents is well known, has been reported for a long time [1-4], and was estimated in 2015 to be 13.4% worldwide [4]. Psychological disorders in these age groups often show long-term impacts on adult life as well [2,5]. Childhood and adolescence are relevant periods for learning and brain maturing, possibly resulting in either a positive or negative impact [6]. Due to these developmental aspects, adolescents have, for example, been found to be especially vulnerable to addiction and addictive behavior [7].

Because of the COVID-19 pandemic and all the accompanying characteristics, prevalence rates of mental health issues have increased in the general population [8], adolescents [9], and young adults, who are among the groups most at risk of suffering from a COVID-19–related decrease in mental health [10-15]. A systematic review reported a lockdown-associated increase in anxiety and depressive symptoms in children and adolescents and an increase in sleep disorders; as risk factors, a priori mental illness and high media exposure were identified [16]. Increased stress levels are associated with respective containment measures [15].

Earlier research spanning from 1946 until 2020 showed an increased risk of depression and anxiety in children and adolescents due to loneliness and isolation [17]. This is an important aspect the current pandemic brought about in many countries [9,17] due to lockdowns and homeschooling, possibly impacting adolescents especially, as emotional support by peers is highly relevant at this age [9]. School closures resulted in a change in daily routines, which is particularly important for young people with mental health problems. Additionally, social isolation poses a risk factor for domestic violence, and an increase in worries related to the future, like school success, university access, and employment chances, has been noted [18].

However, the negative impact of the pandemic consists not only of an increase in mental health issues but also a significant impediment to the accessibility of treatment options, among other aspects, due to the need for social distancing [8,19]. Even before the pandemic, some groups of patients, like migrants [20], different groups of minorities [21], and people in remote areas [22], were difficult to reach through mental health programs. Prior to the pandemic, fewer than 50% of adolescents with depression used adequate services [23,24].

The sudden onset and accompanying restrictions of the pandemic made it even more necessary to increase the offers of online therapy to maintain the treatment of patients with mental health issues. These offers led to a sudden switch in therapists using online interventions [8,25-27], thereby seemingly decreasing perceived barriers by psychotherapists to use online or remote treatment options [8,26]. Nevertheless, the sudden switch also resulted in insecurities and the need for further guidance for therapists [27].

New media and online interventions have been developed and studied for years now [8], including in the context of children and adolescents with psychosomatic illnesses [28], with some studies even finding advantages of virtual therapy compared with face-to-face treatments [29] or at least similar outcomes [30]. Generally, reasonable user satisfaction and feasibility of interventions have been found [30], and studies show that therapeutic alliances can also be reached during videoconferencing, with clients rating bond and presence as equal to face-to-face settings [22]. Online help-seeking seems related to increased anonymity, accessibility, and inclusivity [31], and social media shows benefits for offering mental health care [32]. Applications developed to enhance mental health in children and adolescents show good acceptability [33]. Co-designed eHealth for adolescents appears to be associated with a more engaging and satisfying user experience [34-37].

Still, more research on effectiveness is needed [33], especially considering the sudden switch to online therapies due to COVID-19. Some reviews have been conducted regarding the effectiveness of online interventions for mental health related to the COVID-19 pandemic [38-41]; a review by Bonardi et al [38] focused on randomized controlled trials (RCTs) explicitly designed to address mental health issues related to COVID-19 and found some with promising effects but none for children or adolescents that met the inclusion criteria. Regarding web-based exercise interventions for depressive symptoms and anxiety, a review found no clear recommendations [39], while Valentine et al [40] found telehealth services for neurodevelopmental disorders to be primarily equal to control groups and focused on studies conducted before the onset of the COVID-19 pandemic. Yunus et al [41] found efficacy of digitalized interventions for depression in pregnant women and included studies from before the pandemic.

Nevertheless, it seems of high relevance to identify studies of interventions for mental health conducted after the onset of the pandemic with children, adolescents, and young adults, as persons of these age groups are at a higher risk of being negatively impacted by the pandemic in the long term. Whereas children, adolescents, and young adults can be considered “digital natives” [42], which might make them especially receptive to online interventions, younger individuals also seem to be especially vulnerable to negative aspects of digital media usage (eg, problematic smartphone use) [43]. Not only is it necessary to identify RCTs studying these aspects but one should also take into consideration the specific type of control condition that is used since different types of controls lead to different strengths of studies and especially in mobile health interventions, the combination of a variety of features might account for the resulting effects [44].

Thus, this systematic review and meta-analysis aimed to give a concise overview of studies examining the effectiveness of online or remotely delivered interventions or interventions delivered through digital media since the onset of the COVID-19 pandemic for specific mental health issues in children, adolescents, and young adults.
Methods

Search Strategy
To identify papers published since early 2020 (after the initial onset of the COVID-19 pandemic) until June 2023, a literature search based on the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analysis) framework [45] was conducted in PubMed, PsycINFO, Psynex, Embase, and Google Scholar. The detailed search parameters are depicted in Textbox 1. The reference search strategy was applied to locate additional relevant studies, and Google Scholar alerts were enabled. Multimedia Appendix 1 shows the PRISMA checklist, while Multimedia Appendix 2 shows the search strategy in more detail.

Textbox 1. Search parameters used in the literature search.

<table>
<thead>
<tr>
<th>Databases</th>
<th>PubMed</th>
<th>PsycINFO</th>
<th>Psyndex</th>
<th>Embase</th>
</tr>
</thead>
<tbody>
<tr>
<td>Search parameters</td>
<td>((depression) OR (anxiety) OR (mental health) OR (eating disorder) OR (stress) OR (sleeping disorder) OR (quality of life)) AND ((post covid) OR (long covid) OR (Covid) OR (Sars-cov-2)) AND ((adolescent) OR (child) OR (Juvenile) OR (teenager) OR (youth) OR (young adults) OR (emerging adult)) AND ((Psychology) OR (Psychotherapy) OR (psychiatry)) AND ((online) OR (digital) OR (video-based) OR (tele*)) AND ((effectiveness) OR (efficacy)) AND ((RCT) OR (Randomized controlled trial) OR (Case control) OR (observational cohort))</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Study Selection Process
Before examining full texts, 2 authors (LFG, VF) independently screened the titles and abstracts for inclusion and exclusion criteria. In case of a mismatch between the 2 authors, all authors conferred and made a joint decision. See Figure 1 for the detailed exclusion process at each stage. Studies were included if they were (1) original, interventional studies; (2) published not earlier than 2020 (after the onset of the COVID-19 pandemic); (3) in peer-reviewed journals; (4) written in English or German; (5) focused on psychological or psychotherapy interventions that were delivered remotely (eg, online, via mobile app, or via telephone); (6) targeted at mental health issues like distress, depression or anxiety, psychological well-being, or quality of life (QoL); (7) conducted with children, adolescents, or young adults (from the age of 6 years to the age of 30 years, as emerging adulthood is defined as ages up to 30 years [46]). As outcome measures, we included standardized, validated, and reliable instruments designed for children, adolescents, and young adults to assess the listed mental health issues.
Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analysis) flow diagram.

Statistical Analysis

Meta-analyses were conducted to examine the interventions' effectiveness using standardized mean differences (SMDs) as the outcome measure. The SMD compares postintervention scores between treatment and control groups. Only RCTs or pilot RCTs were included in the meta-analyses. A positive SMD indicates lower outcome scores in the treatment group than in the control group. Eligible studies were grouped by outcome type (anxiety, depression, well-being, disordered eating, psychological stress, social functioning, and COVID-19–related outcomes), and separate analyses were carried out for each group. Score polarity had to be reversed in 1 study [47]. Effect sizes were pooled using the “metafor” package [48] in the R environment. A random effects model was fitted to the data to account for variations in sample size, measures, and methodologies between the different studies. Heterogeneity was assessed using Higgins $I^2$ [49]. Interpretation of the effects sizes is based on Cohen $d$ [50,51].

Additionally, a risk of bias assessment for studies included in the meta-analysis was conducted based on the Joanna Briggs Institute Critical Appraisal Checklist for Randomized Controlled Trials and on the Joanna Briggs Institute Critical Appraisal Tool for Quasi-Experimental Studies [52]. All statistical analyses were conducted in the R environment for statistical computing [53].

Results

Sample of Included Studies

A total of 155 articles were found in the initial database search process, and 2 additional studies were identified through the reference search strategy. Of the total number of articles, 9 duplicates had to be removed. We examined 55 articles at a full-text level. Of these, 20 articles were excluded since they were not original, were case studies, were not intervention studies, did not target the right outcome variables or groups, were not available, or were carried out before 2020. Additionally, 18 studies had to be excluded at the end of the search process as results still needed to be published for these trials. See Figure 1 for a detailed description of the inclusion and exclusion process.

The final sample of articles in June 2023 comprised 17 articles for the qualitative analysis, with an overall 8732 participants. Of these studies, 13 articles were included in the quantitative analysis. RCTs were reported in 16 articles, whereas 1 study [54] had only a quasiiexperimental design with no control group. Only 1 study [55] explicitly compared the online intervention with an intervention conducted in a face-to-face setting. In
addition, 1 study [56] was adapted to an online format during data collection due to the beginning of the COVID-19 pandemic. Of the 17 studies included in the qualitative analysis, 5 (29%) [47,57-60] were conducted in the United States, 4 (24%) [54,61-63] were conducted in China, 2 (12%) each were conducted in Australia [56,64] and the United Kingdom [65,66], and 1 (6%) each was carried out in Canada [67], Italy [55], Iran [19], and Tunisia [68]. Of the included studies, 4 had an approximately (SD 15%) equal distribution of female and male participants [19,54,66,67]. In contrast, 9 had more female participants [47,57-59,61,62,64,65,68], 2 had more male participants [55,63], and 1 study was conducted with female participants only [56]. In addition, 1 additional article reported on 5 studies, of which 4 had more female participants, and 1 had an approximately equal distribution of female and male participants [60]. Of the included articles, 8 focused on children and adolescents [54,55,57-59,62,65,67], while another 8 included young adults [19,47,56,61,63,64,66,68]. The article reporting on 5 studies had samples with only adolescents and samples including young adults [60].

Characteristics of the included studies can be viewed in Tables 1 and 2, and the risk of bias assessment is depicted in Tables 3 and 4.
Table 1. Sample characteristics of the included studies.

<table>
<thead>
<tr>
<th>Study</th>
<th>Sample size</th>
<th>Sample recruitment</th>
<th>Gender</th>
<th>Age</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chen et al (2023) [62]</td>
<td>N=76</td>
<td>Research flyers through social media or school referral</td>
<td>Intervention: female=78.9%; male=21.1%; control: female=76.3%; male=23.7%</td>
<td>Intervention: 11-18 years, mean 16.45 (SD 1.52) years; control: 13-18 years, mean 16.37 (SD 1.24) years</td>
<td>China</td>
</tr>
<tr>
<td>Duan et al (2022) [54]</td>
<td>N=76</td>
<td>Online broadcasting platform</td>
<td>Female=56.58%; male=43.42%</td>
<td>10-12 years, mean 10.72 (SD 0.48) years</td>
<td>China</td>
</tr>
<tr>
<td>He et al (2022) [63]</td>
<td>N=148</td>
<td>Social media, online platforms, university communities, referred by counselors</td>
<td>Female=37.2%</td>
<td>17-21 years, mean 18.78 (SD 0.89) years</td>
<td>China</td>
</tr>
<tr>
<td>Krifa et al (2022) [68]</td>
<td>N=366</td>
<td>Health care students: class visits, posters in university, on website</td>
<td>Female=94%</td>
<td>Mean 20.74 (SD 1.64) years</td>
<td>Tunisia</td>
</tr>
<tr>
<td>Kutok et al (2021) [57]</td>
<td>N=80</td>
<td>Targeted Instagram advertisement</td>
<td>Female=59%</td>
<td>13-17 years, mean 15.3 (SD 1.35) years</td>
<td>US</td>
</tr>
<tr>
<td>Malboeuf-Hurtubise et al (2021) [67]</td>
<td>N=37</td>
<td>In 2 elementary schools</td>
<td>Female=43%; male=57%</td>
<td>Mean 8.18 years</td>
<td>Canada</td>
</tr>
<tr>
<td>Nicol et al (2022) [59]</td>
<td>N=18</td>
<td>Adolescents with depressive symptoms treated in practice-based research networks</td>
<td>Female=88%</td>
<td>13-17 years</td>
<td>US</td>
</tr>
<tr>
<td>Pavarini et al (2022) [65]</td>
<td>N=100</td>
<td>Advertisement on social media</td>
<td>Female=84%; male=14%; male (transgender)=1%; nonbinary=1%</td>
<td>16-18 years</td>
<td>UK</td>
</tr>
<tr>
<td>Prato et al (2022) [55]</td>
<td>N=40</td>
<td>Patients diagnosed with Tourette syndrome at a child and adolescent neuropsychiatry unit</td>
<td>Female=10%; male=90%</td>
<td>9-16 years, mean 13.5 (SD 2.0) years</td>
<td>Italy</td>
</tr>
<tr>
<td>Schleider et al (2022) [58]</td>
<td>N=2452</td>
<td>Instagram advertisements</td>
<td>Female=88.09% (biological sex)</td>
<td>13-16 years</td>
<td>US</td>
</tr>
<tr>
<td>Shabahang et al (2021) [19]</td>
<td>N=150</td>
<td>Convenient sample from Guilan University, Iran; online advertisement in college student social network</td>
<td>Female=51.33%; male=48.67%</td>
<td>Mean 24.7 (SD 5.4) years</td>
<td>Iran</td>
</tr>
<tr>
<td>Simonsson et al (2021) [66]</td>
<td>N=177</td>
<td>Students from the University of Oxford, UK</td>
<td>Female=64.4%</td>
<td>18-24 years (71.8%)</td>
<td>UK</td>
</tr>
<tr>
<td>Suffoletto et al (2021) [47]</td>
<td>N=52 (intervention: n=34; usual care group: n=18)</td>
<td>Young adults with a current mental health diagnosis recruited from primary care or a mental health clinic</td>
<td>Female=85%</td>
<td>Intervention: mean 18.7 (SD 0.42) years; usual care group: mean 18.7 (SD 0.48) years</td>
<td>US</td>
</tr>
<tr>
<td>Sun et al (2022) [61]</td>
<td>N=114</td>
<td>University students, online via WeChat-based flyers and websites targeting college students</td>
<td>Female=73.7%</td>
<td>Mean 22.21 (SD 2.67) years</td>
<td>China</td>
</tr>
<tr>
<td>Torok et al (2022) [64]</td>
<td>N=455</td>
<td>Social media: persons with suicidal thoughts in the past 12 months</td>
<td>Female=84.4%</td>
<td>18-25 years, mean 21.5 (SD 2.18) years</td>
<td>Australia</td>
</tr>
<tr>
<td>Yeager et al (2022) [60]</td>
<td>Study 1: n=2534; study 2: n=790; study 3: n=160; study 4: n=200; study 5: n=119; study 6: n=351</td>
<td>Character Lab Research Network, school students, university students (in specific courses)</td>
<td>Study 1: female=49%, male=49%, nonbinary=2%; study 2: female=64%, male=36%; study 3: female=72.3%, male=27.7%; study 4: female=81.5%, male=18.5%; study 6: similar to study 2</td>
<td>Study 1: 13-18 years; study 2: 17-21 years; study 3: 18-26 years; study 4: 18-32 years; study 5: 14-16 years; study 6: similar to study 2</td>
<td>US</td>
</tr>
<tr>
<td>Zhou and Wade (2021) [56]</td>
<td>N=100 (pre-COVID-19: n=41; during COVID-19: n=59)</td>
<td>University students at risk of developing an eating disorder</td>
<td>Female=100%</td>
<td>17-26 years, mean 19.85 (SD 2.01) years</td>
<td>Australia</td>
</tr>
<tr>
<td>Study</td>
<td>Study type</td>
<td>Intervention</td>
<td>Control conditiona</td>
<td>eHealth technology</td>
<td>Target outcomes</td>
</tr>
<tr>
<td>-----------------------</td>
<td>------------</td>
<td>-----------------------------------------------------------------------------</td>
<td>--------------------</td>
<td>---------------------------</td>
<td>-----------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Chen et al (2023)</td>
<td>RCTb</td>
<td>Online solution-focused brief therapy (SFBT), active intervention group: 38/76, 50%</td>
<td>No treatment, wait list control, 38/76, 50%</td>
<td>Teleconference</td>
<td>Primary outcome: anxiety; secondary outcomes: depressive symptoms and coping styles</td>
</tr>
<tr>
<td>Duan et al (2022)</td>
<td>Quasiexperimental, no control group</td>
<td>Online Strength-informed Acceptance and Commitment Therapy (SACT), active intervention group: 76/76, 100%</td>
<td>No control group</td>
<td>Video conferencing system</td>
<td>Quality of life (QoL) and anxiety</td>
</tr>
<tr>
<td>He et al (2022)</td>
<td>RCT</td>
<td>CBTc-based mental health chatbot (XiaoE), active intervention group: 49/148, 33.1%</td>
<td>2 mHealthd minimal active controls: e-book: 49/148, 33.1%; general chatbot: 50/148, 33.8%</td>
<td>Main intervention: XiaoE, unguided CBT-based chatbot, 1 module a day for 1 week</td>
<td>Primary outcome: depressive symptoms; secondary outcomes: working alliance, usability, acceptability</td>
</tr>
<tr>
<td>Krifa et al (2022)</td>
<td>RCT</td>
<td>CAREe program: internet-based positive psychology intervention, active intervention group: 183/366, 50%</td>
<td>No treatment, wait list control: 183/366, 50%</td>
<td>8-week online self-program: lecture, videos, psychoeducation, practices</td>
<td>Stress, anxiety, depression, emotional regulation, optimism, hope, study engagement, well-being</td>
</tr>
<tr>
<td>Kutok et al (2021)</td>
<td>RCT</td>
<td>IMPACTf, active intervention group: 36/80, 45%</td>
<td>Placebo minimal: enhanced web-based resources: 44/80, 55%</td>
<td>Video intervention plus app-based automated messaging; control: enhanced web-based resources</td>
<td>Cyberbullying: to reduce consequences of cyber victimization, to increase bystander intervention</td>
</tr>
<tr>
<td>Malboeuf-Hurtubise et al (2021)</td>
<td>RCT</td>
<td>MBIg (16/37, 43.2%); P4Ch (21/37, 56.8%); both group-based, delivered online</td>
<td>Comparison of 2 active intervention groups (comparative efficacy)</td>
<td>Teleconferencing platform</td>
<td>Anxiety, inattention symptoms, basic psychological need satisfaction (BPN) in the context of COVID-19</td>
</tr>
<tr>
<td>Nicol et al (2022)</td>
<td>Pilot RCT</td>
<td>CBT, active intervention group: 10/18, 55.6%</td>
<td>No treatment, wait list control, 1:1: 8/18, 44.4%</td>
<td>mHealth app with embedded conversational agent</td>
<td>Primary outcomes: depression severity, anxiety; secondary outcomes: feasibility, acceptability, usability</td>
</tr>
<tr>
<td>Study</td>
<td>Study type</td>
<td>Intervention</td>
<td>Control conditiona</td>
<td>eHealth technology</td>
<td>Target outcomes</td>
</tr>
<tr>
<td>-----------------------</td>
<td>------------</td>
<td>-----------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>--------------------</td>
<td>-----------------------------------------------------------------------------------</td>
<td>-------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Pavarini et al (2022)</td>
<td>RCT</td>
<td>Online peer support training course “Uplift Peer Support Training,” active intervention group: 50/100, 50%</td>
<td>No treatment, wait list control: 50/100, 50%</td>
<td>Zoom, smaller groups in breakout rooms or via WhatsApp</td>
<td>Primary outcomes: motivation to provide support, perceived support-giving skills, frequency of support provided, compassion toward others, connectedness with peers; secondary outcomes: mental well-being, emotional symptoms, self-efficacy, civic engagement</td>
</tr>
<tr>
<td>Prato et al (2022)</td>
<td>RCT</td>
<td>Behavior therapy for youths with Tourette syndrome during COVID-19</td>
<td>Non-mHealth evidence-based active control: comparison of online vs face-to-face intervention: 20/40, 50% each</td>
<td>Video conference vs face-to-face</td>
<td>Tic symptoms, obsessive compulsive symptoms, ADHD symptoms, anxiety, depressive symptoms</td>
</tr>
<tr>
<td>Schleider et al (2022)</td>
<td>RCT</td>
<td>Online single-session intervention (SSI) for depressive symptoms (behavioral activation SSI: 821/2452, 33.5% vs growth mindset SSI: 813/2452, 33.2% vs supportive therapy SSI as the control</td>
<td>Placebo active control: supportive control, (structurally similar [eg, matched in length]): 818/2452, 33.4%</td>
<td>Self-administered online intervention</td>
<td>Depressive symptoms, hopelessness, agency, generalized anxiety, COVID-19–related trauma, restrictive eating</td>
</tr>
<tr>
<td>Shabahang et al (2021)</td>
<td>RCT</td>
<td>Video-based CBT intervention for COVID-19 anxiety, active intervention group: 75/150, 50%</td>
<td>No treatment, wait list control: 75/150, 50%</td>
<td>Self-administered video-based strategies, online booklet</td>
<td>COVID-19 anxiety, health anxiety, anxiety sensitivity, somatosensory amplification</td>
</tr>
<tr>
<td>Simonsson et al (2021)</td>
<td>RCT</td>
<td>Online, guided, 8-week mindfulness program, active intervention group: 88/177, 50%</td>
<td>No treatment, wait list control: 89/177, 50%</td>
<td>Online classes via Zoom,</td>
<td>Anxiety, depression</td>
</tr>
<tr>
<td>Suffoletto et al (2021)</td>
<td>Pilot RCT</td>
<td>Mobile Support Tool for Mental Health (MoST-MH), active intervention group: 34/52, 65.4%</td>
<td>mHealth minimally active control: enhanced usual care (eUC; weblink to psychoeducational videos): 18/52, 34.6%</td>
<td>Text messaging, web-based check-ins, video feedback (psychoeducation)</td>
<td>Mental health symptoms, mental health self-efficacy, mental health care use</td>
</tr>
</tbody>
</table>

a. Conditions are grouped by intervention method (eHealth technology)
<table>
<thead>
<tr>
<th>Study</th>
<th>Study type</th>
<th>Intervention</th>
<th>Control condition</th>
<th>eHealth technology</th>
<th>Target outcomes</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sun et al (2022)</td>
<td>RCT</td>
<td>Mindfulness-based mHealth intervention, active intervention group: 57/114, 50%</td>
<td>mHealth minimally active control (matched social support mHealth control): 57/114, 50%</td>
<td>Videoconferencing via Zoom, WeChat-based mini-program</td>
<td>Primary outcomes: anxiety, depression; secondary outcomes: mindfulness, social support, emotional suppression</td>
<td>Reduction in anxiety and depression and increase in mindfulness and social support in both groups; greater effect on anxiety through mindfulness intervention; greater engagement with and higher acceptability of mindfulness mHealth</td>
</tr>
<tr>
<td>Torok et al (2022)</td>
<td>RCT</td>
<td>Self-guided smartphone app based on DBT, active intervention group: 228/455, 51.1%</td>
<td>Placebo active control, smartphone app with general information: 227/455, 49.9%</td>
<td>Smartphone app (LifeBuoy)</td>
<td>Primary outcome: suicidal ideation symptom severity; secondary outcomes: depression, generalized anxiety, distress, well-being</td>
<td>Significantly higher effects of intervention regarding suicidal ideation; no superior effects regarding secondary outcomes</td>
</tr>
<tr>
<td>Yeager et al (2022)</td>
<td>RCT</td>
<td>Synergistic mindset intervention, active intervention group: study 1: 1208/2534, 47.7%; study 2: 387/790, 49%; study 3: 74/160, 46%; study 4: growth only; 52/200, 26%; stress only; 65/200, 32.5%; synergistic, 39/200, 19.5%; study 5: 61/119, 51.3%; study 6: 179/351, 51%</td>
<td>Placebo active control, study 1: 1326/2534, 52.3%; study 2: 403/790, 51.0%; study 3: 86/160, 54%; study 4: 44/200, 22%; study 5: 58/119, 48.7%; study 6: 172/351, 49%</td>
<td>Self-administered online training module</td>
<td>Studies 1 and 2: stress-related cognition; studies 3 and 4: cardiovascular reactivity; studies 4 and 5: psychological well-being; study 5: daily cortisol levels, academic success; study 6: anxiety levels during COVID-19 lockdowns</td>
<td>Improvements in outcomes greater in treatment group than in control group in all experiments</td>
</tr>
<tr>
<td>Zhou and Wade (2021)</td>
<td>RCT</td>
<td>Online intervention to reduce disordered eating, active intervention group: 77/100, 77%</td>
<td>No treatment, assessment only control: 23/100, 23%</td>
<td>Online format introduced in April 2021</td>
<td>Disordered eating, body image flexibility, self-compassion, fear of self-compassion, negative affect</td>
<td>Significantly higher symptomology during COVID-19 than pre-COVID-19, active intervention significantly increased self-compassion compared with control, no other significant time x condition effects</td>
</tr>
</tbody>
</table>

aTypology of control groups based on Goldberg et al [44].
bRCT: randomized controlled trial.
cCBT: cognitive behavioral therapy.
dmHealth: mobile health.
eCARE: Coherence, Attention, Relationship, and Engagement.
fIMPACT: Intervention Media to Prevent Adolescent Cyber-Conflict Through Technology.
gMBI: mindfulness-based intervention.
hP4C: philosophy for children.
iADHD: attention-deficit/hyperactivity disorder.
jDBT: dialectical behavior therapy.
## Table 3. Risk of bias assessment for randomized controlled trials.

<table>
<thead>
<tr>
<th>First author (year)</th>
<th>True randomization</th>
<th>Concealed allocation</th>
<th>Similar groups at baseline</th>
<th>Participants, personnel, or outcome assessors blinded to assignment</th>
<th>Identical treatment of groups</th>
<th>Follow-up: complete or full description</th>
<th>Analysis of participants in their groups</th>
<th>Outcome measurement: equal and reliable</th>
<th>Appropriate statistical analysis</th>
<th>Appropriate trial design and deviations accounted for</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chen et al (2023) [62]</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Participants and personnel: no; outcome assessors: yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>He et al (2022) [63]</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Krifa et al (2022) [68]</td>
<td>Yes</td>
<td>Unclear</td>
<td>Yes</td>
<td>Unclear</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Kutok et al (2021) [57]</td>
<td>Yes</td>
<td>Randomized but stratified by age and gender</td>
<td>Yes</td>
<td>Participants and personnel: no; outcome assessors: yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Malboeuf-Hurtubise et al (2021) [67]</td>
<td>Unclear</td>
<td>Unclear</td>
<td>No</td>
<td>Unclear</td>
<td>Yes</td>
<td>No follow-up</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No real control group, 2 interventions</td>
</tr>
<tr>
<td>Nicol et al (2022) [59]</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Unclear</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Pavarini et al (2022) [65]</td>
<td>Yes</td>
<td>Unclear</td>
<td>Yes</td>
<td>Unclear</td>
<td>More assessments in treatment group</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Prato et al (2022) [55]</td>
<td>Yes</td>
<td>Unclear</td>
<td>Yes</td>
<td>Participants and personnel: no; outcome assessors: unclear</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Schleider et al (2022) [58]</td>
<td>Unclear</td>
<td>Unclear</td>
<td>Yes</td>
<td>Unclear</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Shabahang et al (2021) [19]</td>
<td>Yes</td>
<td>Unclear</td>
<td>Yes</td>
<td>Unclear</td>
<td>Yes</td>
<td>No follow-up</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Simonsson et al. (2021) [66]</td>
<td>Yes</td>
<td>Unclear</td>
<td>Yes</td>
<td>Unclear</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Suffoletto et al (2021) [47]</td>
<td>Unclear</td>
<td>Yes</td>
<td>Partially</td>
<td>Outcome assessors: yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Sun et al (2022) [61]</td>
<td>Yes</td>
<td>Unclear</td>
<td>Unclear</td>
<td>Participants and research assistant: yes (at first)</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>
Characteristics of Online Interventions Used in the Included Studies

In most (9/17, 53%) of the included studies [54-57,61,62,65-67], different versions of videoconferencing systems were used to deliver the interventions remotely (see Table 2). In an intervention targeted at cyberbullying (Intervention Media to Prevent Adolescent Cyber-Conflict Through Technology [IMPACT]), Kutok et al [57] added app-based automated messaging to their video-delivered intervention. Pavarini et al [65] added the possibility for smaller group discussions by using breakout rooms and WhatsApp for their online peer support training. The mindfulness-based mobile health intervention by Sun et al [61] was supplemented by a WeChat-based mini-program. Other interventions delivered remotely via videoconferencing were the online Strength-informed Acceptance and Commitment Therapy (SACT) [54], mindfulness-based interventions [66,67], philosophy for children (P4C) [67], behavior therapy for Tourette syndrome [55], an intervention to reduce disordered eating [56], and the online solution-focused brief therapy (SFBT), primarily to reduce symptoms of anxiety [62].

Next to these online interventions with teleconferencing systems, 4 studies used self-administered online interventions. Schleider et al [58] examined online single interventions for depressive symptoms, and Shabahang et al [19] targeted COVID-19–related anxiety with self-administered video-based strategies and online booklets. An 8-week self-program with lectures and videos was delivered as an intervention by Krifa et al [68], and Yeager et al [60] used self-administered online training to reduce stress-related symptoms. Text messaging, web-based check-ins, and video feedback with psychoeducation were applied in a study by Suffoletto et al [47] in their Mobile Support Tool for Mental Health (MoST-MH). A cognitive behavioral therapy (CBT)–based mental health chatbot (XiaoE) was used to reduce depressive symptoms by He et al [63]. One study [64] used a smartphone app (LifeBuoy) based on dialectical behavior therapy (DBT) to target suicidal ideation. In contrast, a second study [59] used an app with an embedded conversational agent based on CBT to primarily reduce depressive symptoms.

Not all studies reported on the feasibility and acceptability of their interventions. Those that did, however, found the intervention to be feasible [57,59] and acceptable [57-59,63], met with high satisfaction [19,62], and more accepted and engaging in the treatment group than in the control group [61].

Effectiveness of Online Interventions Regarding Mental Health Outcomes

Mental health–related outcomes varied in the included studies (see Table 2). They included anxiety, depression, mental well-being, social functioning, COVID-19–related symptoms, cyberbullying, Tourette syndrome, disordered eating, suicidal ideation, and psychological stress, among others.

Anxiety

Several studies reported reduced anxiety [19,47,54,55,59-62,66-68]. The impact on anxiety was more prominent in some studies for the treatment group than for the control group [19,59-61,66,68]. In contrast, others found only partial differences [58], equal effects, or no differences between groups [55,64]. P4C had a more significant impact on anxiety than a mindfulness-based intervention in one study [67].
A meta-analysis of 10 studies, with 9 targeting generalized anxiety disorder and 1 targeting health anxiety, showed an overall significant positive effect of interventions in the form of decreased symptoms (SMD=0.44, 95% CI 0.20 to 0.67; \( I^2=82.9\% \)). Figure 2A shows a forest plot of the observed outcomes.

**Figure 2.** Meta-analysis of treatment effect regarding (A) anxiety and (B) depression, shown using the overall and individual study standardized mean difference (SMD) and 95% CIs (those that include 0 show nonsignificant effects), where a positive effect size indicates a decrease in symptoms.

### Depression

Reduced symptoms of depression were found in several studies \([47,55,58,59,61-63,68]\), with superiority of the intervention group found in some \([47,55,58,59,62,63,68]\). The studies conducted by Sun et al \([61]\), Simonsson et al \([66]\), and Torok et al \([64]\) found no superior effects of the treatment on depressive symptoms.

Nevertheless, a meta-analysis of 9 studies found a strong treatment effect (SMD=1.31, 95% CI 0.34 to 2.95; \( I^2=99.63\% \)).

The observed outcomes are depicted in a forest plot in Figure 2B.

### Mental Well-Being, Quality of Life, Agency or Self-Efficacy

Several studies found increased well-being \([47,57,60,65,67,68]\). QoL was increased at the 3-month follow-up in 1 study \([54]\). Self-efficacy or agency was increased in some studies \([58,65]\), while 1 study found no effect \([47]\). Only 3 studies \([64,65,68]\) were eligible for a meta-analysis analyzing the treatment effect on well-being. However, no significant effect was shown in the meta-analysis (see Figure 3A).
Figure 3. Meta-analysis of treatment effect regarding (A) well-being and (B) social functioning, shown using the overall and individual study standardized mean difference (SMD) and 95% CIs (those that include 0 show nonsignificant effects), where a negative effect size indicates an increase of well-being and social functioning.

Other Main Outcomes

A study targeting cyberbullying increased bystander intervention in the treatment group [57], while another showed promising results regarding increased social support-giving skills, compassion toward others, and civic engagement, among other outcomes [65]. Tic and obsessive-compulsive symptoms in children and adolescents with Tourette syndrome were equally reduced via videoconference and face-to-face interventions [55]. Using a smartphone app, 1 study was able to significantly reduce suicidal ideation [64], and 1 study targeting disordered eating increased self-compassion through treatment. At the same time, no other effect was found [56].

Meta-analyses were conducted for several of these outcomes. Regarding disordered eating (see Figure 4A), psychological stress (see Figure 4B), and COVID-19–related trauma or anxiety (see Figure 5), no significant treatment effects were found across studies, while a significant medium effect (SMD=−0.42, 95% CI −0.68 to −0.17; $I^2=0.0\%$) was found for interventions targeting social functioning in 3 studies [47,61,65] (see Figure 3B). For the outcomes of attention and emotional functioning, more data were needed for the meta-analyses.
Figure 4. Meta-analysis of treatment effect regarding (A) disordered eating and (B) psychological stress, shown using the overall and individual study standardized mean difference (SMD) and 95% CIs (those that include 0 show nonsignificant effects), where a positive effect size indicates a decrease in symptoms.
Figure 5. Meta-analysis of treatment effect regarding COVID-19–related symptoms, shown using the overall and individual study standardized mean difference (SMD) and 95% CIs (those that include 0 show nonsignificant effects), where a positive effect size indicates a decrease in symptoms.

Discussion

Principal Findings

This systematic review and meta-analysis is the first of its kind investigating the effectiveness of online or remote interventions for psychological symptoms and disorders in children, adolescents, and young adults after the onset of the COVID-19 pandemic. We examined 17 studies conducted between the pandemic’s start and June 2023 for the impacts of their online interventions. Despite the necessary fast development due to the increased need for remote interventions during the COVID-19 pandemic, the results are promising. All the studies observed positive effects on some of the outcomes they targeted through their remote interventions.

Of the 17 included articles, 16 [19,47,55-68] were RCTs. However, control conditions differed across the RCTs. Only 1 study directly compared an online intervention with a face-to-face intervention [55], which is a control condition that, based on a typology proposed by Goldberg et al [44], provides high comparison strength. The mentioned study found almost equal effects of the online and in-person interventions [55]. Nevertheless, it is expected, due to the great need for fast solutions to deliver interventions amid the ongoing pandemic and obstacles like quarantine, lockdowns, and increased safety measures, that comparisons with face-to-face-interventions were not possible in most cases. Some of the included studies had wait list or no-treatment control groups [19,56,59,62,65,66,68], which can be considered as a control condition with low comparison strength [44]. Others used different content in the control groups [47,57,58,60,61,63,64], mostly providing medium comparison strength [44], or compared different kinds of interventions [58,67] (high comparison strength [44]). These different kinds of control conditions have to be considered when comparing effect sizes of the included studies, and future research should try to use control conditions that provide high comparison strength. Most of the included studies used a videoconferencing system, although several applied interventions that were developed or adapted especially for online delivery.

Meta-analyses on treatment effectiveness yielded significant effects regarding depressive symptoms and medium effects regarding anxiety and social functioning. The results indicate that online or remote interventions show promising results regarding the aforementioned variables. This is a slightly more favorable result than earlier reviews on online interventions or prevention programs with young people conducted before the pandemic. Earlier results were not entirely conclusive, but there were some promising findings for depressive symptoms [69,70] and anxiety [69]. For adults, a review found digitalized CBT interventions to reduce depressive symptoms in pregnant women [41]. However, specifically in the context of COVID-19, a need for more high-quality research has been identified. A systematic review that included only studies with adults published after the onset of the pandemic found some encouraging results for online interventions targeting anxiety and depression [38]. In another review on web-based exercise interventions for adults, the superiority of the interventions over the control conditions was present in only 1 of 3 studies for depressive symptoms and in none for anxiety symptoms [39].

Concerning other variables, the picture is even more unclear: Interventions to improve well-being and reduce psychological stress, disordered eating, and COVID-19–related psychological symptoms did not show significant effects across studies in this meta-analysis. One must consider, however, that only a few
studies for each outcome were eligible for these calculations. Only 3 studies could be included in the analyses regarding well-being. One used a Zoom-based intervention focusing on peer support [65], one used an app to reduce suicidal ideation [64], and the third used a self-administered online positive psychology intervention for different mental health outcomes [68]. Thus, in addition to various main interventions using different kinds of online or remote applications, it can be assumed that the baseline state of well-being or differently expressed amount of suffering was quite different between the 3 studies, which might explain the lack of a significant overall treatment effect.

Psychological stress was analyzed in 3 additional studies [47,64,68]; 2 were included in the meta-analysis of interventions for COVID-19–related outcomes [19,58]. There were also 2 different outcomes used in the latter 2 studies: 1 study [58] examined COVID-19–related trauma, while the other [19] focused on COVID-19–related anxiety. However, the systematic review by Bonardi et al [38] found 3 high-quality studies that were able to decrease different COVID-19–related symptoms like anxiety and depressive symptoms in adults postintervention [71] or at least 6 weeks after the intervention [72,73]; this shows that there seems to be some online or remote interventions available for adult participants.

Regarding disordered eating, the lack of overall treatment effect across the studies could indicate that more than remote therapy is needed for eating disorders and symptoms. Eating disorders might require approaches that treat the somatic aspects in a clinical setting to regularly control for treatment compliance [74]. A previous meta-analysis found similar results, with the lowest effectiveness for online interventions for eating disorders [75]. The study by Zhou and Wade [56] compared symptoms before the onset of COVID-19 and during COVID-19 and found more symptoms during the pandemic, underlining the increased need for interventions due to the pandemic. Although disordered eating and body image flexibility decreased in patients entering the study both before and during the pandemic, the impact of the intervention on self-compassion decreased during the pandemic. All in all, for all the variables showing no overall treatment effects, the few studies available suggest that more research is needed before a clear conclusion regarding the effectiveness of remote or online interventions can be drawn for these symptoms.

Most of the included studies based their online interventions primarily on well-studied therapy forms like CBT [19,47,55,57,59,63] and extensions of CBT or therapy forms related to it like acceptance commitment therapy [54], mindfulness-based interventions [61,66,67], and DBT [47,64]. Interventions based on positive psychology [68] and SFBT [62] were also included in the sample. Thus, interventions were developed from evidence-based forms of therapy. No clear superiority of any form of therapy can be found in this sample of studies. Interventions differed according to length, from 1 session [58] to 3 months [47,59]. Nevertheless, even the single-session intervention was effective [58].

Although some studies show promising results regarding interventions for adolescents [55,58-60,62,65] or young adults [19,47,60,61,63,64,66,68], it is more unclear how effective such interventions are for younger children, as only 2 studies focusing on elementary school children [67] or children up to the age of 12 years [54] could be included. In 1 of these studies [67], a philosophy-based intervention was more effective than the mindfulness-based intervention, possibly hinting at a higher effectiveness of more creative approaches when working with younger children.

Most studies recruited through social media, primary care centers, or at a university. However, in 2 articles, schools were involved: 1 study [62] used school referrals, while the other conducted the intervention in elementary school classes [67], thus showing that, especially with younger children, interventions can also be set within the school context, even if online.

It must be noted that most studies were conducted in North America or China. Although it can be assumed that technical opportunities might be equal in most of Europe, it needs to be clarified how the results can be adapted to lower-income countries, where financial aspects might impede technical opportunities.

The digital transition to online or remotely delivered interventions seems essential, not only considering challenging circumstances like the COVID-19 pandemic, which made face-to-face treatment in many cases impossible, but also in light of the ever-increasing numbers of children, adolescents, and young adults experiencing mental health issues or who have psychological disorders. Thus, it is relevant to develop low-threshold interventions [8,9,14,16]. Nevertheless, several factors must be considered: Legal frameworks might need to be adapted for different countries [8], and therapists might need support when transitioning to online interventions [8,27]. Regarding the development of such interventions, studies have shown positive effects by including peer groups in the development process [34,36] and using peers as advisors [35,37].

However, using digital media and smartphones in and of themselves might pose risk factors for children and adolescents: An increase in cyber victimization through media use has been found [57], and young people are more at risk for addictive behavior in general [7]. Problematic behavior has also been discussed for problematic smartphone use [43], which has been found to impede mental well-being and QoL in children and adolescents [76].

The clinical implications of this meta-analysis are both immediate and far-reaching; the results underscore the versatility and applicability of online therapeutic interventions across diverse settings. As the COVID-19 pandemic amplified the demand for remote interventions, the emergence of promising outcomes, despite rapid development, demonstrates the adaptability and resilience of the mental health sector. These findings suggest that online and hybrid therapeutic modalities not only provide a viable alternative to traditional face-to-face sessions but also bridge the accessibility gap. They offer crucial mental health support to those confronted with problems of accessibility rooted in personal, communicative, geographical, or logistical barriers, as well as challenges stemming from limited mobility due to mental or physical disorders. Such
restrictions often make traditional therapeutic settings challenging, underscoring the importance of versatile, remote solutions.

The utility of remote interventions and the promising outcomes of these methodologies have transformative potential for various contexts. In educational environments, for instance, schools can leverage online interventions to address the mental health needs of students who may be reluctant or unable to access traditional counseling services [35-37,77]. Primary care settings can also integrate telehealth solutions into their care regimes, ensuring patients have consistent and comprehensive mental health support alongside their physical health needs. Additionally, psychiatric rehabilitation provides supportive care, but the therapeutic effects often decrease after discharge [78]. Implementing online or hybrid care modalities post-inpatient treatment could potentially bolster and prolong the beneficial outcomes of therapy, offering a more sustained therapeutic impact for patients in the long run [79].

Telehealth and hybrid systems can be transformative in delivering mental health services. A hybrid approach, which blends traditional face-to-face therapy with online sessions, can cater to diverse patient needs and preferences, enhancing treatment adherence, accessibility, and comfort. For example, patients might initiate their therapeutic journey via face-to-face consultations and later transition to online sessions for convenience, or vice versa. Schools can adopt similar hybrid models, offering in-person counseling sessions and providing digital platforms for students to access support during out-of-school hours or remote learning periods. Likewise, primary care facilities can offer a combination of in-person consultations with remote follow-ups, ensuring continuity of care. The potential of these strategies will need detailed scientific investigation.

This adaptability was also evident during crises like the Syrian and Ukrainian wars, where online interventions were effectively used to support individuals suffering from trauma and distress in their home countries or while migrating (eg, [80,81]). Therapists and mental health professionals from other countries could remotely provide much-needed assistance, showcasing the potential of such platforms in transcending international borders [82]. The success of these interventions in various crises—health pandemics or geopolitical conflicts—signals the need to reevaluate conventional therapeutic models. A shift toward hybrid care models that combine digital and in-person strategies could be pivotal, especially in future catastrophic events. The results underscore that the digital landscape allows for more straightforward, accessible engagement with young populations, demonstrating equal effectiveness as traditional therapeutic methods. Still, research on this population is limited so far. Notably, online and app-based interventions provide a compelling alternative to face-to-face therapy, showcasing notable efficacy, particularly in addressing symptoms of anxiety and depression and improving social functioning. The commitment from young individuals toward these interventions seems robust and encouraging. Although there is a pressing need for further high-quality research on various interventions with heightened comparability, it is evident that there are tangible, effective alternatives to in-person therapeutic interventions. Considering these findings, incorporating online intervention techniques should be paramount in the future training of clinical

Limitations
Although this meta-analysis and literature review is the first to report the effects of online interventions for children, adolescents, and young adults during the pandemic, some limitations of this systematic review and meta-analysis must be noted. We would like to clarify that only studies with experimental or quasi-experimental designs were included in the systematic review. One of the included studies was not an RCT [54]. However, the results of this study are only reported descriptively, as they were not included in the meta-analysis. As for the type of comparison groups, no restrictions were made due to the novelty of the field, aiming to capture a comprehensive range of experimental approaches. We believe this approach provides a more inclusive representation of the current state of research in this domain. Control groups in most studies do not consist of face-to-face interventions due to the pandemic’s nature, potentially affecting the validity of conclusions about the true effectiveness of online interventions. A risk assessment was carried out for all included studies, revealing, in some cases, a need for more details regarding allocation concealment and blinding of participants, personnel, and outcome assessors. Proper randomization was only evident in certain cases (see Tables 3 and 4), possibly affecting the overall quality of the studies included. Above all, the various examined outcomes and the different media or renditions of the online or remote interventions mean that no 2 studies in the sample looked at the exact same intervention. Although understandable in a rapidly evolving field like online interventions, more consistent research on specific interventions is necessary for in-depth meta-analyses and subanalyses. The sample is relatively small, leading to even smaller sample sizes of the outcome groups that were analyzed quantitatively. When the literature search was conducted, several other studies matching the search criteria were registered, but results were unavailable.

Conclusion
All in all, the included studies exhibit promising results regarding the implementation of online or app-based interventions for mental health issues for children, adolescents, and young adults. This is relevant not only in times of crises such as the COVID-19 pandemic or catastrophic events but also given the increasing prevalence rates for psychological disorders in these demographics. The results underscore that the digital landscape allows for more straightforward, accessible engagement with young populations, demonstrating equal effectiveness as traditional therapeutic methods. Still, research on this population is limited so far. Notably, online and app-based interventions provide a compelling alternative to face-to-face therapy, showcasing notable efficacy, particularly in addressing symptoms of anxiety and depression and improving social functioning. The commitment from young individuals toward these interventions seems robust and encouraging. Although there is a pressing need for further high-quality research on various interventions with heightened comparability, it is evident that there are tangible, effective alternatives to in-person therapeutic interventions. Considering these findings, incorporating online intervention techniques should be paramount in the future training of clinical
psychologists and psychotherapists to ensure they remain adaptive, effective, and relevant in our ever-evolving digital age.

Acknowledgments
The authors appreciate the contribution of NÖ Landesgesundheitsagentur, legal entity of University Hospitals in Lower Austria, for providing the organizational framework to conduct this research. We also want to acknowledge support from the Open Access Publishing Fund of Karl Landsteiner University of Health Sciences, Krems, Austria.

Conflicts of Interest
None declared.

Multimedia Appendix 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analysis) checklist.
[DOCX File, 34 KB - mental_v11i1e46637_app1.docx ]

Multimedia Appendix 2
Database search strategy in more detail.
[DOCX File, 17 KB - mental_v11i1e46637_app2.docx ]

References


**Abbreviations**

- CBT: cognitive behavioral therapy
- DBT: dialectical behavior therapy
- IMPACT: Intervention Media to Prevent Adolescent Cyber-Conflict Through Technology
- MoST-MH: Mobile Support Tool for Mental Health
- P4C: philosophy for children
- PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analysis
- QoL: quality of life
- RCT: randomized controlled trial
- SACT: Strength-informed Acceptance and Commitment Therapy
- SFBT: solution-focused brief therapy
- SMD: standardized mean difference

©Linda Fischer-Grote, Vera Fössing, Martin Aigner, Elisabeth Fehrmann, Markus Boeckle. Originally published in JMIR Mental Health (https://mental.jmir.org), 05.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Mental Health, is properly cited. The complete bibliographic information, a link to the original publication on https://mental.jmir.org/, as well as this copyright and license information must be included.
Effectiveness of Virtual Reality–Based Well-Being Interventions for Stress Reduction in Young Adults: Systematic Review

Joy Xu¹; Areej Khanotia²; Shmuel Juni³; Josephine Ku⁴; Hana Sami⁴; Vallen Lin⁴; Roberta Walterson²; Evelyn Payne³; Helen Jo³; Parmin Rahimpoor-Marnani⁵

¹David Geffen School of Medicine at UCLA, Los Angeles, CA, United States
²Faculty of Science, University of Ottawa, Ottawa, ON, Canada
³Faculty of Arts and Science, University of Toronto, Toronto, ON, Canada
⁴Faculty of Health Sciences, McMaster University, Hamilton, ON, Canada
⁵Faculty of Science, York University, Toronto, ON, Canada

Corresponding Author:
Joy Xu
David Geffen School of Medicine at UCLA
417 Charles E Young Drive West
Los Angeles, CA, 90024
United States
Phone: 1 3109481476
Email: joyjxu@gmail.com

Abstract

Background: Adolescents can be especially vulnerable to various stressors as they are still in their formative years and transitioning into adulthood. Hence, it is important for them to have effective stress management strategies.

Objective: This systematic review investigates current well-being interventions that are aimed at reducing stress among young adults. In particular, interventions using the medium of virtual reality (VR) are explored.

Methods: This mixed methods systematic review follows the PRISMA-P (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Protocols) guidelines, and papers were gathered from databases such as PsycINFO, PubMed, Science Direct, Web of Science, OpenGrey, and Edutopia. Predetermined criteria and specific keywords were used to search for the papers. Search results were screened and extracted with all article screening or extraction delegated among all authors. Any disagreements after reconciliation were settled by a third author. The quality and risk of bias of included studies were assessed using the GRADE (Grading of Recommendations Assessment, Development, and Evaluation) Tool for Quantitative Studies. Studies were analyzed qualitatively.

Results: In total, 20 studies were included, and qualitative analysis was performed to evaluate the effectiveness of VR-based interventions in 3 domains: nature, stress, and academics.

Conclusions: Studies using VR interventions, overall, promoted a reduction in stress and an increase in well-being. The findings suggest that VR may serve as an accessible and affordable medium of stress reduction for students and young adults. Larger sample sizes, and a greater number of included studies, may be required in future directions.

(JMIR Ment Health 2024;11:e52186) doi:10.2196/52186

KEYWORDS
well-being; well-being; virtual reality; VR; stress, nature; academic; student; intervention; young adults; teens; adolescent; stressors; stress management; systematic review; accessible; accessibility; students; affordable

Introduction

The COVID-19 pandemic has impacted millions of lives across the world and provided a spotlight on systemic disparities including present limitations to well-being and accessibility for support. The reverberations of this global phenomenon have reconciled greater focus on adolescents and young adults in particular, as their formative years were inevitably impacted by social distancing. The crucial years of their education were faced with obstacles in course delivery, academic opportunities, and social spheres. Furthermore, stresses were compounded with other hardships such as economic setbacks, limited socialization,
and more issues that cumulatively burden one’s well-being, especially in the transition into adulthood. Thus, there has been significant effort to identify potential targets for interventions through research. Broadly, interventions aim to help study some change in individual experiences through strategies and processes after a systematic modification [1]. The main objective is to measure the effect of a process or program on certain situations [1]. In this review, virtual reality (VR) interventions refer to programs or treatments that target one or more determinants of health using a VR headset, which displays a visual environment. This review refers to the Canadian Index’s definition of Well-being stated as “The presence of the highest possible quality of life in its full breadth of expression focused on but not necessarily exclusive to good living standards, robust health, a sustainable environment, vital communities, an educated populace, balanced time use, …” [2]. Thus, some aspects of an individual’s well-being can be measured through the extent to which an individual is stressed. Stress can be understood as a “psychophysiological response” to some form of danger and involves biological components including nervous and hormonal responses to stimuli [3].

This systematic review questions the effectiveness of VR interventions in reducing stress and promoting well-being in students and young adults. Currently, mindfulness-based interventions (MBIs) through digital or computerized means are common, such as mindfulness-based stress reduction, group mindfulness-based intervention, and self-direct mindfulness-based intervention through digital delivery. There is evidence that these interventions have significant improvement in regulating emotion and mindfulness [4]. Similarly, there has been progressive growth in present research demonstrating the efficacy of VR interventions for well-being. VR can be defined as an artificial, 3D digital environment that a user experiences through a computer headset [5]. Although many reviews in the literature summarize the efficacy or need for MBIs delivered on a computer or digital program, there remains a need for a comprehensive review to specifically assess the impact of VR-based interventions on stress and mental well-being among students and young adults.

Currently, the use of nature-based settings alone was found to improve mood to being “good” and “calm” in older adults, without the need for an MBI curriculum [6]. Thus, the papers included in this systematic review use auditory, visual, and even olfactory aids to simulate an MBI to reduce stress and increase mindfulness in young adults, such as integrating MBI techniques within VR or similar tools or resources.

This systematic review aims to understand the effectiveness of current well-being VR interventions in reducing stress among young adults. Due to the nature of learning in the 21st century relying heavily on digital resources, implementing digital tactics is imperative to combating these new sources of stress. In an era where Generation Z (the generation born between 1997 and 2012) is more stressed than previous generations and most are experiencing burnout [7], having digital resources allows for immediate and low-maintenance aid.

Methods

A diverse range of literature was assessed according to the inclusion and exclusion criteria, which were determined prior to the search. These criteria qualified the contextual and scientific relevance of the data by developing a standardized expectation for the literature’s content and experimental purpose. The PRISMA-P (Preferred Reporting Items for Systematic Review and Meta-Analysis Protocols; Multimedia Appendix 1 [8]) guideline was used for this qualitative review. There was an inclusion of gray literature in accordance with Assessing the Methodological Quality of Systematic Reviews-2 (AMSTAR-2), as this body of research encompasses material produced outside of traditional commercial or academic publications to reduce publication bias (eg, government databases and preprints). Papers published between 1980 and 2022, which were randomized controlled trials, were gathered from PsycINFO, PubMed, Science Direct, Web of Science, Open Grey, and EduTopia through an initial screening. A specific keyword search was used to gather these papers including the following keywords from each of the search platforms (“wellbeing” OR “well-being” AND “student” AND “virtual” AND “reality” AND “quantitative” AND “intervention”). Duplicates and papers not matching the screening questionnaire were eliminated.

This review considers young adults to be within the age range of 15-40 years. All studies included are published in English. Included studies focused on the analysis of diverse student demographics in the scope of university programs, socioeconomic status, and culture for a robust analysis of student well-being. The interventions under study in this paper include any VR-based environments that can be accessed by a participant through a VR headset, where concepts can be incorporated into a mobile app. Furthermore, stress reduction interventions that use either physiological or psychological questionnaires were included. Exclusion criteria included any research that generated qualitative data and studies published by organizations with conflicts of interest. Studies with interventions geared toward individuals with psychological illnesses and preexisting psychological illnesses were also excluded to reduce confounding factors and external influences or experiences that may uniquely affect the study outcomes in these demographics.

Data for this study were collected and screened using the Covidence software (Veritas Health Innovation). Multiple authors (JX, AK, SJ, JK, HS, VL, RW, EP, HJ, and PRM) screened a collection of research for relevant studies in 2 steps: abstract screening and full-text screening. From a dropdown menu, authors selected a reason for excluding an article (ie, wrong outcomes: Where the paper reports on findings not in line with the research question of systematic reviews, such as pilot studies, no quantitative data, and biological markers of stress). Two authors were assigned to each study. The authors worked independently and reconciled any disagreements and discrepancies after each step. Reconciliation proceeded by a third author who reevaluated the inclusion and exclusion criteria of the study based on the screening or extraction tools independently of the 2 other authors. Following screening, the authors extracted relevant information to be used within this
review, which included methods of study design, interventions, subject information, and outcomes. EndNote (McMaster University) was used to combine all studies found and to remove any duplicates. GRADE (Grading of Recommendations Assessment, Development, and Evaluation) was used to assess the quality of the studies. GRADE consists of 4 levels of quality: very low, low, moderate, and high. The 2 authors independently evaluated the included studies using the GRADE methodology. When discrepancies arose regarding the assessed quality levels, the said authors collaborated to resolve them, ultimately reaching a consensus on the final quality rating for those studies.

In this case, all studies that did not appropriately follow the inclusion criteria were marked as wrong outcomes, study interventions, and so forth.

Results

The study selection and screening process is outlined in Figure 1. In total, 401 papers were identified from the database search, out of which 20 were included studies. Among these included studies, 11 were randomized controlled trials and 4 were experimental designs. Both the published and gray literature papers went through abstract screening and full-text screening. After the abstract screening, 60 papers remained in total. After full-text screening, 20 papers were left. Through full-text screening, 20 papers remained with categories of nature, stress, and academic contexts within the scope of well-being within the study designs. Studies primarily focused on undergraduate students with included studies encompassing Canada, Ireland, the United States, China, Italy, Europe, Spain, and Russia. Table 1 shows an overview of diverse studies examining the impact of VR environments, with a focus on the nature setting. Key details include authors, publication year, methods, participant populations, interventions, and main findings related to positive and negative affect schedule scores. The studies demonstrate the varied effects of nature-based VR on well-being across different populations. Table 2 presents a synthesis of various studies investigating the effectiveness of VR interventions on stress management and well-being. The studies, conducted using randomized controlled trials and between-subject designs, explore the impact of VR interventions on stress reduction, emotional regulation, relaxation, and positive affect across diverse participant populations. Table 3 provides a comprehensive overview of studies interventions for stress management and well-being among student populations. It outlines various research methodologies, participant demographics, intervention types, and significant findings, offering insights into the use of VR within educational contexts to promote mental health.
Figure 1. Structural diagram presenting screening and included studies.
Table 1. Summary of VR\textsuperscript{a} environment studies (nature).

<table>
<thead>
<tr>
<th>Authors and year of publication</th>
<th>Methods</th>
<th>Population</th>
<th>Interventions</th>
<th>Main findings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Valtchanov and Ellard (2010) [9]</td>
<td>Randomized controlled trial</td>
<td>• Undergraduate students aged 18-26 years from the University of Waterloo, Canada • Participants who can read or write English fluently, do not experience “seizures, vertigo, or motion sickness” or severe visual impairment</td>
<td>3 VR settings: • Nature (tropical scenery) • Geometric (assorted 3-dimensional shapes) • Urban (to-scale model of Shibuya station, Tokyo, Japan)</td>
<td>Mean positive affect scores in terms of ZIPERS scores were only significant in the Nature group post immersion in VR, with a mean of 2.21 (SD 0.71) prior to VR compared to 3.03 (SD 0.98) postintervention ($P&lt;.001$). Both geometric and urban VR settings had no significant impact on improving positive effects</td>
</tr>
<tr>
<td>O’Meara et al (2020) [10]</td>
<td>Randomized controlled trial</td>
<td>• 18 years old or older • Experience high test anxiety and no treatment • Control: low or healthy level of test anxiety • From University College Cork, Ireland</td>
<td>2 VR settings: • Urban environment • Nature environment</td>
<td>Only the high anxiety group of students significantly benefited from the nature VR intervention within the 4-minute session ($P=.02$). The authors conclude that simulated exposure to nature can therefore reduce negative affect, thereby reducing test anxiety. Conversely, virtual exposure to nature settings did not significantly improve test scores</td>
</tr>
<tr>
<td>Browning et al (2020) [11]</td>
<td>Randomized controlled trial</td>
<td>• 18-27 years old, mean age 20 years • University students, from the University of Illinois at Urbana-Champaign, the United States • Excluded if: Diagnosed mood disorder, treated for mental illnesses, hearing impairments, use of alcohol or prescription drugs not normally taken in 24 hours. Intense physical activity in 24 hours • Control: sit in front of a blank wall • Outdoor group: Forest • VR group: same setting as forest group</td>
<td>Individual VR environment interventions: • Gray • Blue • Open green • Partly open green • Partly closed green • Closed green</td>
<td>Both VR and outdoor interventions have positive and significant results ($P&lt;.001$), indicating high positive effect scores. All 3 conditions reveal a significant decrease in negative affect values (outdoor group $P=.034$; control group $P&lt;.001$, VR $P=.03$). Only the outdoor group reported a statistically significant increase in positive effect for mood effects ($P=.044$)</td>
</tr>
<tr>
<td>Gao et al (2019) [12]</td>
<td>Randomized controlled trial</td>
<td>• 120 Chinese college students with mean age of 20.7 years • Myopia degrees &lt;800</td>
<td>6 different VR environment interventions: • Gray • Blue • Open green • Partly open green • Partly closed green • Closed green</td>
<td>Partly open green had the most impact on reducing negative mood, while closed green had the least. Additionally, student preference for environments revealed that blue was most preferred, with gray and closed green as least preferred. A strong positive correlation is revealed between preference for a given environment and positive mood improvement</td>
</tr>
</tbody>
</table>

\textsuperscript{a}VR: virtual reality.
Table 2. Overview of studies on VR\(^a\) interventions for stress management and well-being.

<table>
<thead>
<tr>
<th>Authors and year of publication</th>
<th>Methods</th>
<th>Population</th>
<th>Interventions</th>
<th>Main findings</th>
</tr>
</thead>
</table>
| Villani and Riva (2012) [13]    | Between-subject design | 36 Italian participants were seen to be on or past the higher quartile on stress | • 6-session stress management intervention  
• VR experimental: used “ESCAPE” VR  
• Video experimental group  
• Audio experimental group | Participants were able to reduce their heart rate across conditions and with time. This was specifically seen in the VR condition, where they were more able to reduce their heart rate which in turn helped their emotional state |
| Wayment et al (2015) [14]       | Randomized controlled trial | 32 female first-year undergraduate students | • Audio recording describing the 4 characteristics of the quiet ego  
• Audio recording with VR of a park scene  
• Control group: read nature magazines | The VR with the audio Quiet Ego Contemplation reduced the degree to which participants felt “in the moment” as compared to the audio-only group (\(P<.05\)) |
| Villani and Riva (2008) [15]    | Randomized controlled trial | 36 individuals who were students or office workers from Milan | • Nature VR (park, waterfall, river, garden, etc)  
• Nature video (park, waterfall, river, garden, etc)  
• Relaxing audio | VR, video, and audio of nature settings can help induce relaxation and help in stress management, increase positive emotions, enhance self-awareness, and contribute to emotional regulation. Of the 3 interventions, the VR intervention was seen to have the greatest psychological and physiological effects |
| Villani et al (2007) [16]       | Randomized controlled trial | 34 female students and 30 male students between the ages of 21 and 28 years from the Catholic University of Milan | • VR with video of tropical islands paired with audio of therapeutic sounds  
• DVD of relaxing tropical videos  
• Audio of therapeutic narrative | VR of a relaxing nature environment can enhance the position |
| Taneja et al (2017) [17]        |         |             | • Participants who scored 14 or less on the DASS-21\(^b\) test were deemed normal and eligible for the control group  
• Participants who scored 14 or more were eligible for the stress group | • SCWT\(^c\) task  
• VR-based stress therapy intervention | PANAS\(^d\) questionnaire: a mean score of 32 for positive affect and a mean score of 15 for negative affect were generated |

\(^a\)VR: virtual reality.  
\(^b\)DASS-21: Depression Anxiety Stress Scales-21 items.  
\(^c\)SCWT: Stroop Color and Word Test.  
\(^d\)PANAS: positive and negative affect schedule.
<table>
<thead>
<tr>
<th>Authors and year of publication</th>
<th>Methods</th>
<th>Population</th>
<th>Interventions</th>
<th>Main findings</th>
</tr>
</thead>
</table>
| Kaplan-Rakowski et al (2021) [18] | Between-subjects randomized design | European-based, university, business students, enrolled in an introductory computer science course, willing to participate in meditation activities | • VR intervention group which used a VR headset for meditation  
• Video control group which used a monitor for meditation | The mean differences in test scores between examinations before and after intervention were 0.03 for the VR group (SD 0.27; P=.01) and -0.19 for the control group (SD 0.42). VR was more effective |
| Hernández-Ortega et al (2021) [19] | An experimental, analytical, longitudinal, and prospective study | Participants were included if they were enrolled in Practicum I in the 2017-2018 academic year. Additionally, if they were being treated for anxiety with medication then they were excluded | • Cognitive behavioral therapy (digital)  
• Progressive muscle relaxation (digital) | Participants in IG1 showed lower overall scores than those in the control group, which were also statistically significant for the KEZKAK questionnaire (P=.019) and state-trait anxiety (P<.004). Additionally, no statistical significance was found between IG1 and IG2 regarding stress and anxiety |
| Chen et al (2012) [20] | A randomized controlled trial | Nursing students from a Spanish public university conducted clinical practices between 2017 and 2019 | Digitally guided meditation sessions | SAS scores showed that postintervention participants reported lower rates of anxiety compared to preintervention and the control group. On the other hand, SDS scores did not show a great difference between preintervention, postintervention, and the control group |
| Küchler et al (2020) [21] | Three-armed randomized controlled trial | College students with moderate to low mindfulness levels from universities in Germany, Austria, and Switzerland were recruited. Eligibility criteria include age 18 years or above, enrollment in university or college, proficiency in German, internet access, and moderate to low mindfulness (Freiburg Mindfulness Inventory FMI ≤ 37), with exclusion for ongoing psychotherapy or mindfulness interventions. | Studicare meditation—guided and un-guided | There are no reported results at this time |
| Modrego-Álarcón et al (2021) [22] | Three-armed randomized controlled trial | • 18 years of age or older  
• Enrollment in undergraduate or master’s degree course  
• Enrolled in social sciences or health sciences at the Zaragoza campus or the Calatayud campus  
• Able to speak and write in Spanish  
• Mindfulness-based program that consisted of 90-minute group sessions of mindfulness training  
• Mindfulness-based program paired with short virtual reality sessions  
• Relaxation control group | The results showed higher mean stress levels in the control group (17.73, SD 4.42) when compared to both the mindfulness-based program (15.33, SD 4.50), and the mindfulness-based program along with virtual reality (15.75, SD 4.51). VR on its own was unable to prove to be an effective intervention against stress |
Measures of Well-Being Through Exposure to Nature

From the included studies that underwent data extraction, a considerable number explored the role of natural environments in a VR setting on well-being. It is well known that exposure to nature can positively improve well-being, such as daily walks in urban parks, hikes, or gardening [24]. In research led by White et al [25], it was revealed that dedicating 2 hours weekly to nature activities correlated with significantly better health and well-being outcomes compared to individuals with no nature exposure. Specifically, those who spent this time in nature had a 20% higher chance of reporting good health and a 33% higher chance of experiencing high well-being. However, access to stimulating biodiversity may not be possible for every individual, such as students in highly urbanized campuses, disabled individuals, and students who live in climates with long winter seasons. Therefore, VR may serve as an accommodation. Browning et al [11] reveal that 6 minutes of VR in an outdoor nature setting results in high positive affect levels postintervention (P<.001). The study emphasizes the importance of 360-degree nature videos that are immersive and interactive for adequate similarity to the natural world. O’Meara et al [10] revealed that students with high anxiety can benefit from VR nature exposure by significantly reduced negative affect (P=.02). The authors argue that these negative effects can be associated with examination anxiety; thus, VR serves as a well-being tool to reduce student stress.

The significance of the setting in the VR experience, namely nature, over other landscapes is explored within Valtchanov and Ellard’s [9] study. A neutral geometric environment, an urban environment, and a natural environment were randomly allocated to students with induced stress. The nature setting increased positive effects significantly (P<.001) from a mean of 2.21 pre-VR to 3.03 post-VR compared to no effect of the other settings. Self-reported stress significantly decreased in the natural environment (P<.005), unlike the geometric or urban settings [9]. Additionally, differences in the biodiversity of natural environments in VR were also compared within the included studies. Gao et al [12] conclude that despite no significant differences between the VR settings regarding impact on physiological stress, positive mood, or attention, the authors reveal that restorations in negative mood were significantly different (P=.03). Partly open green spaces, defined in the study as having a 10%-30% composition of trees and shrubbery in the VR environment, had the highest significant effect on reducing negative mood (at P<.01) [12]. To summarize, natural environments presented in the medium of VR can promote positive mood, lowered stress, and overall well-being that can supplement a lack of access that students may face. The included studies feature VR through headsets, such as the Oculus Rift. Although such devices may be cheaper than a cell phone, the average student may not want to invest in a ~US $260 headset for the purpose of accessing well-being interventions [26]. Thus, future directions of implementing VR through cell phones that can be mounted onto cheap headsets, such as the Google Cardboard paired with audio, maybe a cost-efficient but impactful VR intervention with nature for students.

Measures of Well-Being via Stress Levels

Five of the included studies discussed interventions influencing well-being as seen through measures of stress levels. Among many therapy approaches to the reduction of stress and anxiety, relaxation techniques are seen to be very effective [17]. Specifically, the effectiveness of VR therapy was demonstrated through the use of VR environments and relaxing audio. Through the use of the PANSAS questionnaire, a mean score of 32 for positive affect was found as compared to a mean score of 15 for negative affect. Similarly, a study by Plante et al [27] looked at the impacts of VR (paired with aerobic exercise in a laboratory setting) on well-being and stress. Stress and energy were measured under 4 conditions: exercise outdoors, exercise with VR in laboratory conditions, exercise without VR in laboratory conditions, and VR without exercise. Outcomes were measured using Activation-Deactivation Adjective Checklists. It was found that exercising outdoors resulted in the most significant decreases in stress and increases in energy in men.
(P<.10) and especially in women (P<.50). Women also felt calmer (P<.50) following the VR intervention, although the findings were not significant for the men. While there is room for more research, this finding strengthens the credibility of VR as an intervention to reduce stress.

The rapid development of technology in recent years has impacted how one perceives, interprets, and organizes their day-to-day lives, which may influence one’s health and well-being [28]. When assessing well-being, an aspect that may be considered includes an individual’s sense of presence, which can be described as how one experiences and engages in events and situations at the moment [28]. Specifically, in the case of digital environments and simulations, presence can be referred to as an experience where the user is immersed in the VR environment [29]. A study conducted by Villani et al [16] found that VR intervention involving an environment with nature can help induce psychological and physiological effects such as relaxation (P<.005) and lowered anxiety (P<.01). A similar study conducted by Villani and Riva [15] reported that VR can significantly reduce anxiety levels (P<0.05), manage a state of anxiousness (P<.005), and increase relaxation (P<.01). Both studies used the State-Trait Anxiety Inventory to gauge anxiety levels, in addition to employing the Coping Orientation to Problems Experienced Questionnaire to evaluate participant stress levels [15,16]. In the past, researchers have suggested that a greater sense of presence can increase the level of engagement with a digital simulation [29,30]. One study conducted by Mostajeran et al [31] assessed the cognition and stress levels of office workers before and after a VR intervention involving digital plants. It was found that a digital environment with plants (SD 0.82) significantly increased the sense of presence (P<.05) compared to a digital environment where there were no plants (SD 0.87) [31]. Additionally, it was observed that participants performed significantly better on memory tests such as the digit span backward test (P<.05) after the VR intervention [31]. The findings by Mostajeran et al [31] suggest that natural VR environments can help promote a sense of presence, which can have positive effects on productivity levels, as well as mediating stress, which aligns with the findings in this study.

However, not all of the studies showed a benefit from VR. In a study from 2015, a total of 32 female undergraduate students dealing with stress and anxiety in their transition to university completed a quiet ego contemplation intervention, which reminded them of 4 characteristics of the quiet ego: detached awareness, inclusive identity, perspective taking, and growth [14]. They were split into 3 groups: an audio recording describing the 4 characteristics, the same audio recording paired with VR of a park scene, and a control group in which participants perused nature magazines. The participants completed three 15-minute sessions of their assigned condition over 6 weeks. The study found that adding the VR component to quiet ego contemplation reduced the degree to which participants felt “in the moment” (t(29)=2.66; P<.05). The authors hypothesized that this could be explained by the low-quality headset that was used, which was uncomfortable for participants and by the self-directed aspect of the VR, which may have been more distracting than a guided VR experience would have been.

Measures of Well-Being via Academic Contexts

The final theme extracted from the studies included covered measures of well-being as seen in academic contexts. These studies included participants in academic institutions who are facing many stressors as a result of their academic careers. Two of those studies recruited nursing students; 1 study used cognitive behavioral therapy and progressive muscle relaxation, and the other study administered a mindfulness meditation intervention [19,20]. Ortega et al [19] used the KEZKAK questionnaire and the state-trait anxiety test as a measuring tool for stress levels. They found overall lower scores for both tests among participants who underwent the intervention with P values of .019 for the KEZKAK questionnaire and .004 for the state-trait anxiety test [19]. Similarly, Chen et al [20] used the self-ratings anxiety scale as a measurement tool, where lower rates of anxiety were reported for participants in the intervention group. In considering these findings, the strength of digital-based interventions to decrease stress is further reinforced. The authors of both studies argue that these types of intervention are necessary for students to be able to function healthily and excel in their fields [19,20].

A common intervention method among these studies was the use of VR headsets. Kaplan-Rakowski et al [18] used VR headsets in their study for a meditation intervention. This was used for the experimental group, while the control group went through the intervention by watching it on a monitor. Participants went through a pretest and posttest, which entailed basic computer science tasks. It was found that the VR group (0.03) showed a higher mean difference in test scores as compared to the control group (~0.19). Berezina et al’s [23] intervention also used a VR headset. Participants in experimental group A watched a stimulating scene, followed by a relaxing scene, while participants in experimental group B watched a relaxing scene followed by a stimulating scene. Finally, there was a control group, which did not watch anything. They found that while both experimental groups experienced a decrease in fatigue, experimental group B showed a less pronounced decrease as compared to experimental group A. The final study on this theme used a VR intervention aimed at reducing stress. There were 3 groups: one which underwent a group mindfulness session, another which went through a VR session, and a control group. Modrego-Alarcón et al [22] found that participants in the mindfulness-based program showed the lowest levels of stress as compared to any other group. Additionally, the VR group scored lower stress levels compared to the control group. Overall, the studies show that VR, well-being interventions are effective in reducing stress levels and promoting relaxation among students, especially those in higher institutions.

Strengths and Limitations

Included studies analyzed within defined categories that supported qualitative analyses of excerpts within niches of productivity that are rarely explored in demographics of young adults. This systematic review provided a diverse understanding of various techniques and research approaches to well-being in psychology in the present literature. However, given the limited research conducted on the demographic of young adults within well-being from a digital standpoint, included studies
demonstrated heterogeneity, which precluded meta and statistical analyses; thus, a qualitative analysis of studies was provided. Other limitations included exclusions based on languages that were not English, which do not extend the research boundaries to non–English-speaking demographics. Additionally, given the nuance of recent developments in technology, a lack of longitudinal studies may limit the application of this systematic review. Thus, elements that may be beneficial or consequential to study participants may not be adequately observed as well-being can be a culminated outcome of long-term behaviors.

Conclusions
In conclusion, this systematic review aimed at evaluating the efficacy of VR interventions to promote the well-being of students and young adults. To achieve this, studies were divided into themes of nature, stress, and academic contexts as focuses of interventions. Overall, the included studies reveal that VR interventions pose a promising medium to reduce the stress experienced by young adults and students, which can ultimately improve well-being. These findings reveal that VR may serve as an accessible, affordable tool for students and young adults to promote well-being or lower stress levels. However, there are some limitations to the review. The included studies tend to have smaller sample sizes, which may not be representative of students as a whole. A total of 20 studies were included in the final phase of extraction. Future directions may include expanding the search criteria to include more studies that may have higher sample sizes.

Acknowledgments
We would like to extend our sincere appreciation to all researchers and authors whose valuable contributions have enriched the body of literature in this field. We express our gratitude to Liza Iaralova, Valerie Lo, Cherry Tagra, and Ammal Riaz for their dedicated efforts in conducting literature searches, data extraction, and analysis. Their meticulous work has been instrumental in synthesizing the findings presented in this review. We also acknowledge the academic and institutional support that has facilitated our research endeavors. Lastly, we extend our thanks to the broader scientific community whose collective efforts continue to propel the exploration of novel technologies for enhancing well-being.

Conflicts of Interest
None declared.

Multimedia Appendix 1
PRISMA 2020 checklist.
[PDF File (Adobe PDF File), 1080 KB - mental_v11i1e52186_app1.pdf ]

References


Abbreviations

AMSTAR-2: Assessing the Methodological Quality of Systematic Reviews
GRADE: Grading of Recommendations Assessment, Development, and Evaluation
MBI: mindfulness-based intervention
PRISMA-P: Preferred Reporting Items for Systematic Review and Meta-Analysis Protocol
VR: virtual reality
Implementation of an Electronic Mental Health Platform for Youth and Young Adults in a School Context Across Alberta, Canada: Thematic Analysis of the Perspectives of Stakeholders

Gina Dimitropoulos¹,²,³,⁴, MSW, PhD, RSW, RMFT; Emilie M Bassi¹, BA, MSc; Katherine S Bright⁵,⁶, BSc, BN, MN, PhD; Jason Gondziola⁷, BA, MBA; Jessica Bradley⁷, RN, MN; Melanie Fersovitch⁷, BN; Leanne Stamp⁷, BEd; Haley M LaMonica⁸, PhD, ABPP-CN; Frank Iorfino⁸, BSc, MBMSc, PhD; Tanya Gaskell⁷, BA, MBA; Sara Tomlinson⁷, MSc, CHE; David Wyatt Johnson⁹,¹⁰, MD

¹Faculty of Social Work, University of Calgary, Calgary, AB, Canada
²Calgary Eating Disorders Program, Alberta Health Services, Calgary, AB, Canada
³Mathison Centre for Mental Health Research and Education, University of Calgary, Calgary, AB, Canada
⁴Alberta Children's Hospital Research Institute, University of Calgary, Calgary, AB, Canada
⁵School of Nursing and Midwifery, Faculty of Health, Community, and Education, Mount Royal University, Calgary, AB, Canada
⁶Heroes in Mind, Advocacy, and Research Consortium (HiMARC), Faculty of Rehabilitation Medicine, College of Health Sciences, University of Alberta, Edmonton, AB, Canada
⁷Provincial Addiction and Mental Health, Alberta Health Services, Calgary, AB, Canada
⁸Brain and Mind Centre, University of Sydney, Sydney, Australia
⁹Maternal Newborn Child and Youth Strategic Clinical Network, Alberta Health Services, Calgary, AB, Canada

Corresponding Author:
Gina Dimitropoulos, MSW, PhD, RSW, RMFT
Faculty of Social Work
University of Calgary
2500 University Drive NW
Calgary, AB, T2N 1N4
Canada
Phone: 1 403 220 7332
Email: gdimit@ucalgary.ca

Abstract

Background: Youth, aged 15 to 24 years, are more likely to experience mental health (MH) or substance use issues than other age groups. This is a critical period for intervention because MH disorders, if left unattended, may become chronic and serious and negatively affect many aspects of a young person’s life. Even among those who are treated, poor outcomes will still occur for a percentage of youth. Electronic MH (eMH) tools have been implemented in traditional MH settings to reach youth requiring assistance with MH and substance use issues. However, the utility of eMH tools in school settings has yet to be investigated.

Objective: The objective of this study was to gain an understanding of the perspectives of key school staff stakeholders regarding barriers and facilitators to the implementation of the Innowell eMH platform in secondary schools across the province of Alberta, Canada.

Methods: Guided by a qualitative descriptive approach, focus groups were conducted to elicit stakeholder perspectives on the perceived implementation challenges and opportunities of embedding the Innowell eMH platform in secondary school MH services. In total, 8 focus groups were conducted with 52 key school staff stakeholders.

Results: Themes related to barriers and facilitators to youth and school MH care professional (MHCP) capacity in implementing and using eMH tools were identified. With respect to youth capacity barriers, the following themes were inductively generated: (1) concerns about some students not being suitable for eMH services, (2) minors requiring consent from parents or caregivers to use eMH services as well as confidentiality and privacy concerns, and (3) limited access to technology and internet service among youth. A second theme related to school MHCP barriers to implementation, which included (1) feeling stretched with high caseloads and change fatigue, (2) concerns with risk and liability, and (3) unmasking MH issues in the face of limited
resources. In contrast to the barriers to youth and MHCP capacity, many facilitators to implementation were discussed. Youth capacity facilitators included (1) the potential for youth to be empowered using eMH tools, (2) the platform fostering therapeutic relationships with school personnel, and (3) enhancing access to needed services and resources. MHCP capacity facilitators to implementation were (1) system transformation through flexibility and problem-solving, (2) opportunities for collaboration with youth and MHCPs and across different systems, and (3) an opportunity for the continuity of services.

Conclusions: Our findings highlight nuanced school MHCP perspectives that demonstrate critical youth and MHCP capacity concerns, with consideration for organizational factors that may impede or enhance the implementation processes for embedding eMH in a school context. The barriers and facilitators to implementation provide future researchers and decision makers with challenges and opportunities that could be addressed in the preimplementation phase.

(JMIR Ment Health 2024;11:e49099) doi:10.2196/49099

KEYWORDS
electronic mental health; eMH; digital mental health; youth and young adult mental health; secondary schools; implementation science; qualitative descriptive methods; mental health platform; mental health; mobile phone

Introduction

Background

Globally, mental health (MH) issues are on the rise among young people, particularly since the onset of the COVID-19 pandemic [1,2]. A recent meta-analysis showed that of all people with MH disorders, 62.5% had onset before the age of 25 years [3]. Furthermore, 1 in 10 youths is estimated to experience an MH disorder in their lifetime, including depression, anxiety, and eating disorders [4]. Approximately 1 in 3 youths is estimated to experience an MH disorder in their lifetime [5]. Since the onset of the COVID-19 pandemic, the rates of depression and anxiety are estimated to have doubled among youth compared with the prepandemic period [6]. A recent systematic review found evidence of decreases in access to preventive MH supports coupled with increases in attempted suicide, self-harm, and suicidal ideation among adolescents during the COVID-19 pandemic [2]. The increase is linked to a disruption to daily routines and family connections [4], a reduction in the availability of social support [1], and a lack of necessary coping skills to navigate challenges and stressors experienced by youth [4]. Furthermore, public health restrictions and the increasing rates of social isolation have resulted in higher rates of internet and gaming problems among youth [7,8].

Despite the concerns about the rise in MH problems among youth, many barriers to help-seeking behaviors have been noted, including stigma [9], low MH literacy, and a reluctance to seek MH services [10]. There are many systemic barriers that further interfere with youth accessing needed services when they seek support, including long waitlists and limited availability of evidence-based treatments that are youth friendly and developmentally appropriate [11]. In a recent systematic review, long wait times for MH services were linked to a deterioration in MH outcomes [12]. As a result of the pandemic, web-based care options, including the application of electronic MH (eMH) tools, were widely and rapidly implemented, demonstrating their utility in reaching youth affected by MH and substance misuse and their ability to transcend geographic distance during public health restrictions [13].

The global spike in poor youth MH calls for early detection and intervention of MH disorders using traditional and nontraditional MH services, including eMH services [9,14]. eMH is defined as the “provision of guided mental health care where consumers navigate a rapid and more effective system experience of service entry, skilled assessment, and multidisciplinary and coordinated care, as well as ongoing outcome-based monitoring” [15]. A proliferation of eMH services has been observed using smartphone apps and electronic tools (e-tools) [16,17]. eMH services may supplement various aspects of MH service delivery, including disseminating resources and information, performing ongoing assessments, tracking changes to symptoms and behaviors, delivering psychosocial therapies, and providing peer support [16]. Randomized controlled trials have consistently demonstrated a reduction in some MH symptoms, including mild to moderate depression and anxiety [18-20]. These randomized controlled trials have especially focused on delivering treatment via smartphone interventions (those that provided in-app feedback, and those used to enhance or support face-to-face interventions) [18]. However, research on the implementation of eMH tools outside of routine clinical settings is lacking.

Research suggests the importance of the school setting as a place of first contact for youth who seek MH support [21-23], positioning teachers as the first point of contact to see MH “red flags” [24]. The use of eMH technology provides an opportunity to transform services, including improving the early detection of MH issues, better care coordination and care plans, and referrals to necessary services for youth [25]. However, there is limited research on the implementation and use of eMH tools in school systems.

Existing studies have highlighted general barriers and facilitators to the implementation of digital MH tools among both youth and MH care professionals (MHCPs). Stigma remains a daunting barrier to using eMH services, deterring individuals from seeking help owing to societal and internal biases [26,27]. The quality of the therapeutic relationship, fostered by well-designed interfaces and personalized content, serves as a facilitator, particularly in remote settings [26,27]. Systemic influences (eg, access and infrastructure) can both hinder and facilitate engagement [26]. In addition, one’s level of MH literacy significantly impacts one’s decision to engage, with greater literacy leading to the increased acceptance and use of eMH resources [26]. In summary, these intertwined factors create a
complex landscape for eMH and digital MH interventions. The available literature exploring the barriers and facilitators to implementing eMH interventions in the school setting is notably scarce, underscoring the need for a deeper exploration of this important topic. To our knowledge, this study is the first to explore the barriers and facilitators to implementing eMH interventions in the school setting.

The Innowell eMH platform is a configurable web-based tool that seeks to provide a more personalized approach using measurement-based care to complement existing MH services [25]. The platform empowers youth and young adults to actively participate in their care plan by using measurement-based MH care, including a comprehensive questionnaire consisting of psychometric tools exploring 20 biopsychosocial domains (eg, depression, anxiety, and social connectedness). Examples of health priorities and health cards on the platform’s dashboard are provided in Figure 1 [28] and Figure 2 [29]. After completing baseline measures, platform users are immediately provided access to vetted web-based resources, including apps, e-tools, and crisis line options to use between sessions, and cues to discuss care options with their MHCPs. The platform can assist “with the assessment, feedback, management, and monitoring of their mental ill health and maintenance of well-being by collecting personal and health information from a young person, their clinician(s), and supportive others” [25]. The platform was designed using participatory design in Australia. The participatory design processes included participants with lived experience, health professionals, and service staff from diverse service populations to optimize the usability of the platform design [25,30]. It is important to note that the “platform does not provide stand-alone medical or health advice, risk assessment, clinical diagnosis, or treatment. Instead, it guides and supports (but does not direct) young people and their providers to decide what may be suitable care options” [25].

Figure 1. Participant and clinician electronic mental health Innowell platform dashboard health priorities [28].
**Objectives**

Hence, it is critical to evaluate the implementation of an eMH tool such as the Innowell eMH platform in secondary schools to assess its acceptability, usability, and efficacy in school contexts. Furthermore, there is a need to examine how eMH tools can be customized for school-based organizations where youth often spend most of their time interacting with educators, guidance counselors, and school MHCPs. This study seeks to elicit the perspectives of key stakeholders within the school divisions regarding their perceptions of barriers and facilitators to the potential implementation of an eMH platform in schools across the province of Alberta, Canada.

**Methods**

This study used a qualitative thematic analysis of school stakeholder focus groups (FGs) to better understand their perceived barriers and facilitators to the implementation of an eMH platform for youth and young adults in Alberta.

**Ethical Considerations**

This research study received human participant research ethics approval from the University of Calgary Research Ethics Board (project name “Pre-Implementation and Implementation Phase of E-Mental Health for Youth and Young Adults in Alberta”; REB20-1137). Informed consent was completed on the web before the commencement of the FGs. The study participants’ data were deidentified to ensure privacy and confidentiality. Participants were not compensated for their participation.

**Qualitative Descriptive Methodology**

A qualitative descriptive methodology was used to describe barriers and facilitators to the implementation of the Innowell eMH platform through the exploration of participants’ perceptions and experiences [31]. This methodology is well suited to health research and is descriptive in nature [31,32], capturing events or conditions from the perspective of individuals and in their everyday language [33]. We sought to obtain a rich description of key stakeholder perspectives rather than researcher-provided interpretations of their preconceived views about the value of an eMH platform. We used an FG approach to obtain data from a group of key stakeholders in schools who may hold diverse concerns about the utility of the platform. In contrast to individual interviews, FGs offer participants the opportunity to share a range of thoughts and to iteratively build on ideas and considerations for integrating an eMH platform in school settings across different geographic regions and diverse student populations [34]. The sample consisted of 52 school district key stakeholders representing 8 communities across Alberta. The web-based FGs were conducted from February 1, 2021 to May 31, 2022.

A semistructured interview guide was created with input from the research and implementation teams on the project. The questions included a focus on thoughts and feelings about the platform and the implementation processes. The interview guide was structured into 3 separate but related sections and was constructed to suit the study objectives, research questions, and the theoretical domains framework (TDF). The TDF is an integrative framework that includes 14 theoretical domains derived from 33 validated health and social psychology theories.
and 128 constructs that have the potential to drive and explain health-related behavior change [35]. The 14 theoretical domains are knowledge; skills; social or professional role and identity; beliefs about capabilities; optimism; beliefs about consequences; reinforcement; intentions; goals; memory, attention, and decision processes; environmental context and resources; social influences; emotions; and behavioral regulation [36]. The interview guide was designed to explore which TDF domains were relevant for the implementation of the Innowell eMH platform.

FGs were conducted to evaluate the barriers and facilitators that may affect the implementation of the Innowell eMH platform in existing school services. The interview guide covered unique factors about the service setting and organization, unique community contexts that can influence implementation, questions about the unique youth population that the school serves, and the barriers and facilitators that might influence implementation.

The study setting for recruitment included kindergarten to grade 12 schools situated in rural and urban communities across Alberta. This study used a targeted email and web-based meeting recruitment, whereby practice leads from the implementation team working with each community shared recruitment materials with senior decision makers and clinical supervisors at schools taking part in the eMH study. Key stakeholders in this area were recruited, including teachers, administrators, psychologists, school counselors, community connectors, and social workers. The site clinical lead for each community identified potential participants who would be involved in using the platform from a range of different roles, including managers, leaders, and frontline clinicians. Three members of the research team moderated all FGs.

There were 3 inclusion criteria for participants of this study. Each participant had to be employed in the school district as an MHCP, decision maker, educator, or administrator; be proficient in written and spoken English; and use web-based technologies on a laptop computer, desktop computer, tablet, or smartphone.

The FGs were conducted by the research team and included an implementation practice lead, 2 facilitators, and a notetaker who was often a youth research partner. The FGs lasted approximately 90 minutes. Of the 52 participants, 3 (6%) to 10 (19%) participated in each FG. All interviews were audio recorded and transcribed verbatim by a professional transcriber. The transcribed interviews were checked for accuracy by a member of the research team.

A total of 8 FGs were conducted with the 52 key stakeholders from 11 school divisions. The FGs collectively contained representation from 21 schools, including high schools (grades 10-12; n=10, 48%), an elementary school (kindergarten to grade 6; n=1, 5%), combined elementary to high school (kindergarten to grade 12; n=3, 14%), combined junior high to high school (grades 7-12; n=3, 14%), as well as outreach and web-based schools (n=4, 19%). Participants included teachers, administrators, psychologists, school counselors, community connectors, and social workers. Community connectors in the school division are members of the community who support and inform individuals about how to access support groups, services, and information that might help improve their MH and well-being [37]. Sample size was justified at 52 key stakeholders when the research team determined that the sample provided “information power,” a concept used in qualitative methodology to denote that the sample held substantial information [38]. There was representation from public, Catholic, Francophone, and web-based and alternative outreach schools for students (also referred to as clients by some of the participants) who benefit from nontraditional learning methods. The FGs were composed of diverse school district MHCPs, from managers and administrators to teachers.

**Analytic Plan: Thematic Analysis**

Using a combined inductive-deductive approach, the research team used the 6 stages of thematic analysis outlined by Braun and Clarke [39]. Thematic analysis is a method of identifying and reporting on themes that emerge through the data [39]. The six-stage analytical process consists of (1) familiarization with the data, (2) coding, (3) generating themes, (4) reviewing themes, (5) defining and naming themes, and (6) producing the report [39].

As part of the project’s youth engagement strategy, a team of coders, including youth and young adults with varying research experience, reviewed the transcribed FGs independently to increase familiarization with the data. After familiarization, the coders conducted a preliminary descriptive coding of their assigned transcripts. The research team generated these codes independently using Microsoft Word and Taguette qualitative coding software. The coders then exchanged FG transcripts and conducted secondary coding. Each transcript was coded by a minimum of 2 coders, which ensured reliability because the codes depended on 2 different individuals achieving and deciding on the same code outcome [40]. The research team then worked collaboratively to group the codes into broader themes through discussion and consensus. Finally, illustrative quotes for each theme were identified. The core research team met regularly to share their impressions of patterns across and within the FG transcripts.

The team closely followed guidelines for publishing qualitative data [41]: for instance, the research team recorded decisions about the coding and the decision-making processes for establishing the themes. The coders worked independently to analyze the data and met regularly to discuss the codes established and assigned. Through consensus, the coding group members formulated their themes based on patterns across the FGs with different school settings. As a form of member checking, our team presented our findings to the broader research team consisting of researchers from different academic institutions to ensure that the data resonated with their experiences and understanding of the academic literature. The practice of reflexivity was used through memoing and group discussions, where members of the research team reflected on their biases and their positionality and questioned their own assumptions throughout the direction of the research process. The practice of keeping written memos was used to record the process and document the rationale for how the data were coded [42].
**Results**

**Participant Demographics**

Most of the participants (38/52, 73%) identified as women. Many of the participants were counselors or therapists (12/52, 23%) or social workers (11/52, 21%). Of the 52 participants, 19 (37%) were aged between 40 and 49 years, and 32 (62%) had been practicing their profession for ≥11 years. The majority of the participants were born in Canada (45/52, 87%) and held a graduate degree (22/52, 42%). A complete list of descriptive information about the participants is presented in Table 1.
Table 1. Descriptive quantitative survey information of school stakeholders (N=52).

<table>
<thead>
<tr>
<th>Variable</th>
<th>Values, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Gender</strong></td>
<td></td>
</tr>
<tr>
<td>Woman</td>
<td>38 (73)</td>
</tr>
<tr>
<td>Man</td>
<td>7 (14)</td>
</tr>
<tr>
<td>Not reported</td>
<td>7 (13)</td>
</tr>
<tr>
<td><strong>Profession</strong></td>
<td></td>
</tr>
<tr>
<td>Counselor or therapist</td>
<td>12 (23)</td>
</tr>
<tr>
<td>Social worker</td>
<td>11 (21)</td>
</tr>
<tr>
<td>Teacher</td>
<td>7 (13)</td>
</tr>
<tr>
<td>Administrator</td>
<td>6 (12)</td>
</tr>
<tr>
<td>Psychologist</td>
<td>6 (12)</td>
</tr>
<tr>
<td>Community Connector</td>
<td>2 (4)</td>
</tr>
<tr>
<td>Not reported</td>
<td>8 (15)</td>
</tr>
<tr>
<td><strong>Age group (y)</strong></td>
<td></td>
</tr>
<tr>
<td>20-29</td>
<td>3 (6)</td>
</tr>
<tr>
<td>30-39</td>
<td>15 (29)</td>
</tr>
<tr>
<td>40-49</td>
<td>19 (37)</td>
</tr>
<tr>
<td>50-59</td>
<td>8 (15)</td>
</tr>
<tr>
<td>60-69</td>
<td>2 (4)</td>
</tr>
<tr>
<td>Not reported</td>
<td>5 (10)</td>
</tr>
<tr>
<td><strong>Member of a visible minority group</strong></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>5 (10)</td>
</tr>
<tr>
<td>No</td>
<td>39 (75)</td>
</tr>
<tr>
<td>Not reported</td>
<td>8 (15)</td>
</tr>
<tr>
<td><strong>Length of time practicing (y)</strong></td>
<td></td>
</tr>
<tr>
<td>&lt;1-5</td>
<td>6 (12)</td>
</tr>
<tr>
<td>6-10</td>
<td>9 (17)</td>
</tr>
<tr>
<td>≥11</td>
<td>32 (62)</td>
</tr>
<tr>
<td>Not reported</td>
<td>5 (10)</td>
</tr>
<tr>
<td><strong>Length of time at organization (y)</strong></td>
<td></td>
</tr>
<tr>
<td>&lt;1</td>
<td>6 (12)</td>
</tr>
<tr>
<td>1-5</td>
<td>10 (19)</td>
</tr>
<tr>
<td>6-10</td>
<td>9 (17)</td>
</tr>
<tr>
<td>≥11</td>
<td>22 (42)</td>
</tr>
<tr>
<td>Not reported</td>
<td>5 (10)</td>
</tr>
<tr>
<td><strong>Employment status</strong></td>
<td></td>
</tr>
<tr>
<td>Full time</td>
<td>35 (67)</td>
</tr>
<tr>
<td>Part time</td>
<td>7 (13)</td>
</tr>
<tr>
<td>Other</td>
<td>5 (10)</td>
</tr>
<tr>
<td>Not reported</td>
<td>5 (10)</td>
</tr>
<tr>
<td><strong>Ethnicity (n=54)^a</strong></td>
<td></td>
</tr>
<tr>
<td>Indigenous to North America</td>
<td>5 (9)</td>
</tr>
<tr>
<td>Other North American origins</td>
<td>7 (13)</td>
</tr>
</tbody>
</table>
Barriers and Facilitators

In each of the following subsections, we highlight youth and MHCP capacity barriers and facilitators to the implementation of eMH tools in school settings. Youth capacity to engage in eMH services refers to their ability to effectively use digital tools and platforms, such as eMH, for addressing MH concerns. This involves not only their technical skills but also their understanding of the potential benefits and limitations of eMH services. Clinician capacity to engage in eMH services refers to their proficiency in using digital tools and platforms, such as assessment tools, to provide MH services to their patients. This also encompasses their ability to navigate the technical aspects of an eMH platform as well as their competence in delivering quality MH care while adhering to the regulations of their profession’s governing body as well as their health care organization’s ethics and regulatory guidelines in the eMH realm.

First, youth capacity barriers were identified, which included the suitability of measurement-based care for all youth with different types of MH disorders that may impact their engagement in eMH services; the role of caregivers, confidentiality, and perceived consent concerns among youth; and inadequate device and internet access. We then established a second theme relating to MHCP capacity barriers to implementation, which included MHCPs feeling stretched with high caseloads and change fatigue, concerns about liability and risk considerations, and the potential to unmask MH issues in the face of service and resource constraints.

By contrast, the final 2 themes consisted of many facilitators to implementation. The third theme—youth capacity facilitators—included the potential for an active and empowered role in care, the potential to foster and enhance therapeutic relationships, and the importance of improving access to services and resources. The fourth theme covered MHCP capacity facilitators to implementation, including the unique flexibility and natural problem-solving skills of school staff that can contribute to system transformation, the potential for collaboration with MHCPs across the continuum and with different systems, and an opportunity to strengthen the continuity of services. Collectively, these themes highlight the potential of the school setting to create opportunities for system transformation through the implementation of eMH tools.

Barriers to the Implementation of eMH Tools in School Settings

Youth Capacity Barriers

Individual Characteristics and Implications for the Engagement of Measurement-Based Care

Participants identified numerous factors as potential barriers to youth using the measurement-based care protocol integrated within the Innowell eMH platform. Given the large and diverse youth population served by secondary schools in Alberta, participants expressed concern that some youth with certain types of MH issues and sociodemographic characteristics may be unable to use the Innowell eMH platform. Participants perceived that measurement-based care may not be suitable for youth with the following conditions: severe MH diagnoses (eg, moderate to severe depression or anorexia nervosa), disability diagnoses (eg, attention-deficit/hyperactivity disorder), low cognitive capacities, those experiencing suicidal and self-harm behaviors, and low literacy levels and new or migrating Canadians not fluent in English. Participants were concerned that measurement-based care, which includes the 20-domain assessment features on the Innowell eMH platform, might be lengthy, difficult to understand, and overwhelming:

I do wonder about say students just with their literacy levels, like their comprehension. And especially those who maybe present a specific neurotypical, they present that way but actually maybe their comprehension is really, really weak. And so, onboarding and then going through the diagnostic

Values, n (%)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Values, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>European origins</td>
<td>28 (52)</td>
</tr>
<tr>
<td>Latin, Central, and South American origins</td>
<td>1 (2)</td>
</tr>
<tr>
<td>Not reported</td>
<td>11 (21)</td>
</tr>
<tr>
<td>Born in Canada</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>45 (87)</td>
</tr>
<tr>
<td>No</td>
<td>1 (2)</td>
</tr>
<tr>
<td>Not reported</td>
<td>6 (12)</td>
</tr>
<tr>
<td>Level of education</td>
<td></td>
</tr>
<tr>
<td>Some university, college, or trades school</td>
<td>1 (2)</td>
</tr>
<tr>
<td>Completed college or trades school</td>
<td>5 (10)</td>
</tr>
<tr>
<td>Bachelor’s degree from a university</td>
<td>18 (35)</td>
</tr>
<tr>
<td>Graduate school</td>
<td>22 (42)</td>
</tr>
<tr>
<td>Not reported</td>
<td>6 (12)</td>
</tr>
</tbody>
</table>

Numbers do not equal 52 because participants were allowed to select multiple responses.
tools and the self-assessment tools. Like I do wonder about the implications of that. [FG21-P02]

This quote demonstrates many participants’ reflections about the implications of onboarding youth who may be unable to easily navigate certain aspects of the platform on their own.

Participants also described a concern that the extensive measurement-based care assessment protocol on the Innowell eMH platform might exacerbate symptoms for some of the youth experiencing complex MH issues:

I do worry that part of this platform, although [it] gives them the tools and the apps and all of those things, and reminds them to speak to the therapist or go see their person. We’re also asking a lot of kids who are struggling...Right? So, they’re not feeling very mentally well and then we’re asking them to do other things to take care of their own wellness. And I worry—are they capable of that? I guess we’ll see. [FG17-P02]

This quote demonstrates a perception that some youth may find it difficult to navigate the Innowell eMH platform when they are struggling with MH issues and wellness.

Role of Caregivers, Confidentiality, and Caregiver Consent

Other potential barriers identified by participants are problems that may arise in obtaining consent from caregivers and protecting the confidentiality of youth who will access the platform but do not wish to disclose their MH concerns to family members:

And if a lot of the issues surrounds the parent relationship, then that becomes a problem and that has definitely been the past barrier to them getting services sometimes is they don’t want their parents having to sign consents. [FG22-P03]

Although not a pervasive issue, participants raised concerns about some youth living in unsafe family environments where it would not be acceptable to disclose an MH issue owing to stigma and fear of reprisals from family members. These factors were described as significant barriers if consent is required from caregivers for the access and use of the Innowell eMH platform. In the absence of clear confidentiality assurances, it may prevent young people from engaging in the apps and e-tools available through the platform and in MH services more broadly.

Information-sharing requirements within the school district with caregivers was also described as a potential barrier to implementation. Participants were concerned about who owns the assessment information completed by youth and having to navigate information-sharing requests or requirements from caregivers or other MHCPs:

I think it would need to be very clear what, if any, of the information would be getting shared with their parents, like abundantly clear. I get questions on this all the time when I meet the kids for the first time is, “What exactly is confidentiality?” And many, many questions surrounding that. [FG10-P03]

This participant builds on this concern, stating that many young people come to their first session reluctantly, which could spread to minimal uptake of the Innowell eMH platform if information-sharing protocols about consent and parental involvement are unclear. Participants shared that being unclear about confidentiality with young people would be a barrier not only to using the platform but also to young people accessing MH services and resources more generally.

Limited Access to Technology and Internet Service Among Youth

A significant barrier described in the school context is a lack of access to technology for youth to use the Innowell eMH platform. With the shift to web-based school and MH services during the COVID-19 pandemic, many MHCPs learned of the access-related challenges with remote and web-based schooling for some youth, especially those residing in remote or rural geographic regions. Participants learned that many youth lack access to technology or devices that would be necessary to use the platform; and many youth also have unstable, or completely lack, Wi-Fi, data, or an internet connection needed to access the platform:

One of the things that has been a challenge over—specifically with the course of the pandemic, is it really has brought to light how many students—because we are in a rural, northern remote area where Wi-Fi access and device access, cell phone access, all that stuff can be really spotty if not almost impossible for some of our families which means some of the students don’t have access to this kind of thing outside of typical school hours. [FG22-P03]

Participants were concerned that introducing an eMH platform, the use of which requires access to devices and a stable Wi-Fi connection, would be an unsuitable initiative for many of their youth population, especially those who cannot afford Wi-Fi access, thus not addressing existing barriers to care for youth considered disadvantaged and vulnerable.

MHCP Capacity Barriers

Feeling “Really Stretched” With High Caseloads and Change Fatigue

Participants described the school system as often the first point of contact for “a revolving door” (FG17-P02) of students experiencing diverse challenges, from interpersonal problems to the onset of MH disorders. This creates difficulty for them to envision the integration of an eMH platform that might worsen their already high caseloads. Participants described feeling “really stretched” (FG5-P08) and supporting students in the school district as “mentally taxing” (FG21-P02). A participant highlighted these concerns:

Like my vested interest in this point-in-time is it not just a huge, huge source of time and energy suck for them because obviously, before this project, they were busy with their full-time work anyways. Now this is like another thing on their plate. So, that would be my sort of trepidation with this is how it will add more to their plate and potentially increase their stress. [FG21-P02]
Participants were concerned about implementing a new eMH platform that would be time consuming and potentially overwhelming to navigate, given the many conflicting demands on their schedules in an ever-changing work environment. Participants also described a reluctance to undergo the necessary organizational changes needed to integrate eMH into their routine practice:

\[\text{We’re always given new tools constantly. “Try this, these are different assessments and mental health promotions has this going on.”} \text{[FG17-P02]}\]

Participants questioned how a new tool could be integrated into the existing organizational flow and day-to-day tasks for school staff. This includes the allocation of resources to integrate a new eMH tool, the limited contact time with youth, how this tool factors into waitlist management, and how to integrate this into current data management practices and software. Participants described other data management tools that they have been grappling with in the school system and a concern about how additional eMH tools may create work duplication, especially if there is a lack of interoperability.

Furthermore, the impact of the COVID-19 pandemic on the school system resulted in higher workloads for many school MH staff. Some of the participants noted experiencing a high volume of cases with students with severe MH acuity and a reduction in the availability of community-based MH services; for example, a participant shared how the COVID-19 pandemic has impacted the availability of other community-based MHCPs:

\[\text{[L]ack of resources. We lost our child psychiatrist in the community. The contract wasn’t renewed. So, that puts an extra burden on the schools for services. That affects both school jurisdictions as well as with the current COVID situation again, it’s the lack of service providers being able to access the school environment.} \text{[FG22-P04]}\]

Thus, when grappling with external factors, such as the unexpected consequences of the COVID-19 pandemic and the constant upgrading of organizational tools, participants were apprehensive about additional training requirements for another tool to integrate into their daily school routines.

**MHCP Concerns About Liability**

One of the key concerns among participants was the possibility of being liable for the suicidal thoughts and behaviors notification embedded in the Innowell eMH platform. Many of the participants shared that they were concerned that a slow response time to a notification that a student is reporting high levels of suicidal thoughts and behaviors may inadvertently prevent timely intervention and therefore contribute to a youth’s suicide. This lag time in responding to a notification of acute crisis could be seen as negligence and thus render the MHCP liable. Participants described a concern that youth might erroneously assume that school professionals are immediately made aware that they are in distress or experiencing an emotional crisis after triggering the suicidal thoughts and behaviors notification:

\[\text{[B]ut the consistent challenge that happens over and over again is when the school closes at 3:30 PM, our mental health closes at 4:30 PM, what do these parents—so now we know that there’s ideation and we’re at high risk. What do we do in [the community] with these students?} \text{[FG22-P02]}\]

Thus, participants were concerned about who would be responsible for responding to youth outside of school hours and during school closures, including weekends and summer holidays.

**Unmasking MH Issues in the Face of Limited Services**

Finally, many of the participants expressed concern that they may not have the capacity to respond to the needs of the students within their school district. Some of the participants worried that the increased assessments conducted using measurement-based care embedded within the Innowell eMH platform would unmask the degree to which MH problems impact their student population. Increased assessment would inevitably lead to increased identification of students requiring MH care within student MH services in the school environment. However, participants expressed concern that they would not be able to access needed resources, supports, and specialty services for students presenting with MH concerns owing to long wait times and the absence of evidence-based treatments:

\[\text{So, to me it’s no different than a lot of rural communities, just a lack of support services to address mental health needs and the negative stigma around mental health and its needs as well.} \text{[FG22-P04]}\]

Participants were concerned about the potential for this to create a crisis in the school district where MHCPs would be inundated with students with MH problems without proper services and resources to refer them to in the community:

\[\text{Like I don’t want to say that as “Be careful, let’s not [open the floodgates].” I think we just need to be prepared for that and how are we going to support everybody in the meantime or through that?} \text{[FG5-P07]}\]

The knowledge of the shortage of primary care services and specialty clinics within the school and across the MH community created significant reservations among participants in considering using the Innowell eMH platform.

**Facilitators to Implementation of eMH Tools in School Settings**

**Youth Capacity Facilitators**

**Creates Potential for an Empowered and Active Role in Care**

Participants argued that the most important facilitator to implementation is the belief that the Innowell eMH platform may empower youth to take an active role in their care journey. Being able to access the platform independently, completing assessments that are relevant to them, and accessing apps and e-tools that reflect their needs are key facilitators to implementation. Relatedly, participants acknowledged that young people are far more comfortable using technology, and therefore their readiness to adopt the Innowell eMH platform may be high. Many young people are already using technology...
to complete and upload homework assignments, schedule appointments, and, in some cases, attend appointments and classes. Thus, participants in our study asserted that there was strong potential for young people to use the platform as an extension of these existing practices.

Participants further suggested that integrating an eMH platform into the school system from primary to secondary schools would allow students to incrementally develop self-management skills and scaffold the information and resources they need for wellness and MH care. Using the Innowell eMH platform within the school context would also allow young people to monitor MH changes over time:

[A] grade 9 student could carry through this platform for the next 4 years and actually manage their own care. [FG5-P04]

Collectively, participants viewed the many advantages of eMH technology for their students, including youth empowerment and the ability to acquire self-management skills, resources, and psychoeducation information:

But also, empowering youth. I love the idea that they can access this and it’s at their fingertips and they can start to really see their growth and that’s exciting to me. [FG5-P01]

Participants acknowledged youth being able to see their growth over time and access apps and e-tools while in the school system as a salient facilitator to them agreeing to adopt and implement eMH technology in their school. Thus, capitalizing on the young people’s receptivity to technology, embracing their desire to self-manage, and acknowledging this tool as an opportunity to empower youth are all potential implementation enablers.

**Fosters Therapeutic Relationships**

Many of the participants highlighted the important ways in which the Innowell eMH platform can foster a stronger therapeutic relationship between themselves and youth. Participants described several ways in which this can occur, including the potential for youth to develop greater MH literacy, which would lead to awareness and the use of terminology about MH and well-being, leading to a shared language with their MHCP and increased collaboration to explore what care is needed:

[It can be a great tool by the looks of it for youth to identify areas. We might know how they’re feeling but not quite able to categorize. “Oh, maybe I’m struggling due to grief,” and being able to recognize, name, and have the language for that, I think, can also be an empowering tool. [FG5-P02]

Many of the participants further highlighted that the repeated use of the measures to assess MH issues might be an effective way for youth to alert their MHCP on how they are coping between sessions or how their MH symptoms are changing over time. Participants described the ease of reviewing assessment information that could be updated between sessions and identifying trends that students may not be verbally sharing with their MHCP:

I can see the benefit of it in terms of that—the narrative in between sessions. I often have encouraged the youth that I work with to use the accompanying email to sort of do exactly this piece...there’s a narrative in between things that are happening between sessions, I can’t answer all the time, but you can just send me little notes to say, “This is for next session.” And just sort of in that hopes of creating some sort of form of accountability. [FG8-P01]

Participants also anticipated the measurement-based care of the Innowell eMH platform providing information on students’ progress at different time points during the academic year. In a school setting, the platform can facilitate the sharing of information and the strengthening of the therapeutic relationship over many years. The accessibility of the baseline assessments and records of MH changes over time helps inform the students and their MHCP while also increasing the continuity of care.

**Enhances Access to Services and Resources**

The final youth capacity facilitator identified by participants is the opportunity to enhance access to information, resources, wellness apps and e-tools, and web-based support by using the Innowell eMH platform. Closely interconnected to the previous facilitator, participants suggested that youth can learn about, explore, and access resources, services, apps, and e-tools more easily by using the platform:

I feel like this could be beneficial to them because they can kind of access areas that is of interest or is of concern to them and they can get resources right there and then. That they can click on those apps, those different phone numbers or websites are things that I saw when I was clicking on some of those. And I think that that will be great for them. [FG10-P01]

Participants recognized that long wait times for specialty services, limited resources, and barriers to accessing MH services, especially in rural areas, might be temporarily addressed by students accessing resources via the Innowell eMH platform.

**MHCP Capacity Facilitators**

**System Transformation Through Flexibility and Problem-Solving**

Participants view the school environment as creating the conditions for them to be early adopters of the Innowell eMH platform because educators are flexible, resilient, and solution focused when faced with challenges:

Yeah. I think our program is unique just in the flexibility that we have...We have the flexibility that, you know, you meet a student, the next day they’re in tears, they can stop by. Or you have a high-risk who could check in in a couple days as opposed to having to wait for 2 weeks. [FG14-P03]

Participants suggested that by using the Innowell eMH platform, youth would be monitored more carefully, which is different from what might typically occur at the MH clinics in the community. Many of the participants also brainstormed solutions...
to the barriers listed in the preceding section; for example, a participant brainstormed solutions for youth who may not have access to devices that allow them to use the Innowell eMH platform and apps:

> Even if they don’t have the media devices, there’s nothing wrong with using the Chromebook [in the school] and sitting in the next room and working on it. [FG5-P05]

Despite the many barriers delineated by the school personnel regarding the implementation of the Innowell eMH platform, participants frequently brainstormed solutions to some of their unique challenges while participating in the FGs. This speaks to the flexibility and natural problem-solving ability of school personnel to contribute to creating system transformation.

### Collaboration With MHCPs Across the Continuum and With Different Systems

By developing a comprehensive baseline assessment of students using measurement-based care on the Innowell eMH platform, participants believed that they would develop a better understanding of the care required to meet their students’ unique needs. The information gleaned from the measurement-based care assessment protocol was further perceived as supporting a stepped and staged care approach. This was highlighted by a participant who described how the information could be used:

> For like, a clinician’s dashboard—to be able to see those alerts right away would help us really prioritize where we need to target. [FG14-P01]

Participants further highlighted that sharing the results of the assessments and ongoing monitoring of MH symptoms with other MHCPs would allow them to share the mutual understanding and language that would enhance collaborations with other MHCPs:

> I agree, collaboration and communication is so important. I think being able to get out of our silos and work as a team of professionals will only serve to benefit our young people in need of supports. [FG9-P02]

Participants were eager to highlight the potential to remove silos from the MH care system by embracing an eMH platform that could be used across MH care systems. Collectively, many of the MHCPs recognized the potential for these factors to foster a therapeutic relationship with youth, especially if there is increased collaboration among MH teams.

### Opportunity for the Continuity of Services

Finally, participants shared that the Innowell eMH platform may be used to facilitate the continuity of care across services in the community. Participants highlighted the potential for all members of the care team to share, communicate, and access similar information on the platform about a shared youth client and thus promote the continuity of care. This can help in streamlining the service process by enhancing the continuity of care among services:

> ‘Cause we want a wraparound service for the kids and I know there’s been lots of privacy issues and that’s kind of been better actually for some clinicians—the more open sharing. [FG9-P04]

Some of the participants suggested that MHCPs may embrace the implementation of the Innowell eMH platform as an opportunity to create change within the health care system:

> We have a beautiful opportunity here at the middle to drive what happens at the top and at the bottom from a client perspective, our students and our families that’ll be impacted and our community and then our leadership. Our superintendents, our mayors, all of those people. And we right now are the driving middle force that can actually change what’s happening. And we need to utilize this opportunity well. [FG5-P04]

Participants of this study strongly advocated that school staff hold a unique role within the community: they establish relationships with a majority of the youth and young adults in every community for many years, while also maintaining relationships with leaders of the school districts and with municipal and provincial levels of government as well, thus providing them with the opportunity to build stronger relationships and drive MH care change by embracing this implementation project.

### Discussion

#### Principal Findings

This study aimed to explore perceived barriers and facilitators relating to the implementation of the Innowell eMH platform in secondary schools in Alberta. Using a descriptive qualitative methodology, we held FGs with key stakeholders in school divisions, including administrators, teachers, management staff, school counselors, psychologists, and community connectors (N=52). Our research shows interconnected barriers and facilitators to implementation as it relates to youth and school MHCP capacities, with system-level considerations. We conclude the discussion with a summary of recommendations for addressing implementation in school settings (Textbox 1).
Textbox 1. Recommendations from the qualitative focus groups with school stakeholders.

<table>
<thead>
<tr>
<th>Recommendation and description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Clear policies and processes for consent</strong> (with regard to accessing mental health [MH] services)</td>
</tr>
<tr>
<td>• School-based leaders and decision makers establish policies and processes regarding consenting mature minors and obtaining and navigating parental consent</td>
</tr>
<tr>
<td>• Development or tailoring of existing policies and processes in the local context and culture around the consenting process</td>
</tr>
<tr>
<td><strong>Web-based training environment</strong></td>
</tr>
<tr>
<td>• Create interactive learning activities to enhance educators’ knowledge regarding the application of electronic MH (eMH) tools with diverse students</td>
</tr>
<tr>
<td><strong>Level of education, knowledge, and skills</strong></td>
</tr>
<tr>
<td>• Train school MH care professional (MHCPs) on how to support youth to identify when and how to share MH issues with caregivers, if appropriate</td>
</tr>
<tr>
<td>• Establish communities of practice as an approach to provide post training education and supervisory support to ensure that school personnel can apply their knowledge and skills of measurement-based care and eMH tools</td>
</tr>
<tr>
<td><strong>Level of support and supervision</strong></td>
</tr>
<tr>
<td>• Decision makers within school settings identify how to support staff to receive adequate training and supervision to learn to use, and embrace the implementation of, eMH tools and apps</td>
</tr>
<tr>
<td>• Ongoing mentorship, supervision, and support is needed to integrate eMH tools into the school settings</td>
</tr>
<tr>
<td><strong>Structure of the school system and contexts of practice</strong></td>
</tr>
<tr>
<td>• Integrated eMH tools fit into established workflows and processes, and work duplication is removed where possible to maximize implementation efforts</td>
</tr>
<tr>
<td>• The process of adaption and adoption requires attention to the cultural and contextual components of assessment, formulation, and intervention, including the ways school personnel recognize, explain, and manage distress</td>
</tr>
<tr>
<td><strong>Existing socioeconomic barriers to access</strong></td>
</tr>
<tr>
<td>• Considering socioeconomic status and access in the communities of implementation is a key pillar of equity that should be addressed in the implementation of eMH tools and measurement-based care</td>
</tr>
<tr>
<td>• Emphasize the inclusion and integration of local culture beliefs, practices, language, social norms, family, community, and social network for better understanding of help-seeking behaviors</td>
</tr>
<tr>
<td><strong>Address liability concerns and ensure crisis response protocol</strong></td>
</tr>
<tr>
<td>• Liability concerns among stakeholders should be heard, integrated, and rapidly addressed through training and clinical supervision to increase willingness to use eMH tools in school settings</td>
</tr>
<tr>
<td>• Ensure that MHCPs have the competencies to effectively respond to a student’s disclosure of suicidal thoughts and behaviors via the Innowell eMH platform</td>
</tr>
<tr>
<td>• School administrators and decision makers must establish risk mitigation protocols and procedures to assure school MHCPs that clear pathways are determined and easily implemented to rapidly respond to students experiencing suicidal thoughts and behaviors</td>
</tr>
<tr>
<td>• Clinical supervision and administrative support must be made available to school MHCPs requiring assistance with students’ disclosures of suicidal thoughts and behaviors and need of acute care</td>
</tr>
<tr>
<td><strong>Youth focus</strong></td>
</tr>
<tr>
<td>• Keeping youth at the center of eMH implementation strategies could inspire and enliven uptake among MHCPs</td>
</tr>
<tr>
<td><strong>Youth engagement</strong></td>
</tr>
<tr>
<td>• Use eMH tools to enhance and improve the way that youth and MHCPs interact with each other and the way that MH teams from different systems communicate</td>
</tr>
<tr>
<td>• Youth should be included in discussions about how to implement eMH in schools</td>
</tr>
</tbody>
</table>

The first theme highlights concern about the extensive assessments embedded within the Innowell eMH platform and youth capacity (eg, attention span and literacy skills) to complete assessment measures. In general, there is concern about whether the literacy level of the Innowell eMH platform matches that of its intended users in the community. In line with our findings,
MH status and demographic variables are among the primary user capacity factors described in the literature that can affect eMH use [26,43]. A systematic review by Borghouts et al [43] found 59 studies that reported that the MH status of the user can play a significant role in the engagement and uptake of eMH tools. Although the severity of MH symptoms can be a barrier to some patients using eMH tools [44,45], this may also depend on the type of eMH tool; for example, individuals with more severe MH symptoms may be more likely to use eMH tools inclusive of assessment measures [43,46]. To achieve this, researchers have co-designed a game that trains clinicians to identify how and when to adopt eMH tools with their clients [47]. A similar type of learning activity can be used to enhance educators’ knowledge about the application of eMH tools with diverse students. MHCPs would also benefit from evaluating digital tools through the adoption of adapted rating scales and digital navigators [48].

The results of our study also suggest that school personnel share a concern about how to obtain parental consent and maintain student confidentiality when considering using eMH tools with youth. In keeping with existing research, this finding shows that one of the most significant barriers to young people accessing school MH services is a lack of clarity about confidentiality, especially from their caregivers [10,49]. Young people may avoid accessing school-based MH services and resources if they anticipate negative consequences from their family members [50,51] or if they are required to obtain parental consent [52,53]. This concern also extends to digital MH tools: Cavazo-Rehg et al [51] found that two-thirds of underaged youths displaying symptoms of eating disorders were unwilling to obtain parental consent to access a mobile MH intervention. Researchers and clinicians alike strongly advocate that MHCPs use parental consent waivers, reassure young people of their privacy and autonomy, and address adolescent stigma concerns [51] to increase the use of eMH tools. Given the pervasive concerns about confidentiality and the requirement of parental consent to use the Innowell eMH platform in schools, we strongly recommend that school-based leaders and decision makers establish clear policies and processes about consenting mature minors and navigating parental consent. School staff would benefit from informed consent policies, inclusive of teachers, social workers, and administrators that have been approved across school districts in the province or country. We also recommend training school MH personnel on how to support young people to identify when and how to share MH issues with family, if appropriate.

Related to youth capacity barriers, participants were concerned that introducing an eMH platform, the use of which requires access to devices and a stable internet connection, would be unsuitable for many of their youth population, especially for those residing in remote and rural communities. This finding reaffirms the concern about how to integrate eMH technology, given the existing barriers to service access noted in multiple studies [43,54]. Strudwick et al [54] reviewed a total of 31 mobile apps and 114 web-based applications and resources that had the potential to support the MH needs of the broader Canadian population during the COVID-19 pandemic. Key barriers of concern tended to be access, cost, and poor connectivity [54]. Socioeconomic status and access in the communities of implementation are considered key pillars of equity that should be addressed to support the success of future implementation strategies and ensure equitable access of this opportunity [54].

This study points to specific issues and concerns about the lack of available time to build capacity and integrate the Innowell eMH platform into practice. Time constraints, burnout, and change fatigue were also identified as significant barriers to implementation. In alignment with our findings, although high-quality person-centered care is a priority for MH services [55], there are issues pertinent to school MHCPs, such as limited time, competing demands, high caseloads [56], high degrees of burnout [57], and insufficient training and administrative support [58], all of which can create barriers to providing optimal care and adopting a new eMH platform in a system. Furthermore, LaMonica et al [28] argue that if digital solutions are to be successfully used by MH professionals, decision makers must reduce the administrative burden and responsibilities placed on individuals to adopt eMH technology. Importantly, if eMH tools are introduced in an MH service setting, implementation strategies must consider what could be removed or combined to avoid increasing workloads of school MHCPs. We recommend that decision makers within school settings identify how to support staff to receive adequate training and supervision to learn to use, and embrace their curiosity about the implementation of eMH tools and apps.

From an MHCP capacity perspective, participants expressed concern about how a new eMH platform could be integrated into the existing organizational flow and day-to-day tasks for school MHCPs. If an eMH platform is operationalized into current vision, mission, priorities, and work plans, this is reported to enable implementation and delivery [59]. Operationalization factors that increase implementation are reported to include workflow processes; leadership, including workplace culture and management; and systems, including the organization of people and resources to meet the needs of the community [59]. Greenhalgh et al [60] draw attention to the importance of integrating technological advances in MH care into the work processes and existing tools and resources used by MHCPs. When eMH tools do not fit into traditional workflows and processes, there is a risk of low engagement and poorly sustained implementation once trials end [17,60,61]. Thus, eMH initiatives must fit into the standard workflows of the health system setting [61,62] to improve youth MH outcomes [63]. Work duplication could be removed by reducing administrative burden on professionals by ensuring the interoperability of MH tools with existing management systems (removing the need to enter the same data across multiple systems) [28]. This finding is particularly meaningful for school district professionals who already use existing data management processes, as well as MH tools and resources, and are concerned about the integration of eMH tools into the established organizational workflow. Our findings support the recommendation that eMH tools must fit into established workflows and processes and work duplication removed where possible to maximize implementation efforts.
A major implementation barrier concern among participants is the potential to be held liable for a youth’s suicidal thoughts and behaviors notification alerted through the platform. Although liability concerns have not been sufficiently discussed in the literature, Scott et al [64] examined telehealth policy implications and suggested that some risk should be anticipated and expected in the implementation of eMH tools. Thus, liability concerns among stakeholders should be heard and considered and rapidly addressed through training and clinical supervision to increase willingness to use eMH tools in school settings. Our research sheds light on key liability concerns among school staff that should be considered and urgently addressed during the preimplementation phase to ensure that school MHCPs have the competencies to effectively respond to a student’s disclosure of suicidal thoughts and behaviors via the Innowell eMH platform. This points to a key improvement in quality care by providing young people the opportunity to assess and detect suicidality and, furthermore, empowering the young person to seek support and care pathways for suicidality [65]. Equally important, school administrators and decision makers must establish risk mitigation protocols and procedures to assure school MHCPs that clear pathways are determined and easily implemented to rapidly respond to students experiencing suicidal thoughts and behaviors. A systematic review of MH training for secondary teachers shows that most training interventions have been carried out through facilitated course trainings and workshops, such as MH first aid training, peer support, and suicide prevention and postvention, to name a few [66]. This review showed an improvement in MH knowledge and attitudes among teachers, and the interventions reviewed should be considered in training and preparing schools to implement eMH tools, especially when suicide assessments and alert systems are included in the eMH options. Finally, clinical supervision and administrative support must be made available to school MHCPs requiring assistance with students’ disclosures of suicidal thoughts and behaviors and need of acute care. An area of future research is to review the legal and ethical considerations of telehealth services [64] across different systems to facilitate the successful implementation of eMH tools.

Many of the participants also expressed concern that they may not have the capacity to respond to the needs of the students within their school district owing to limited resources in their school and the broader community. These concerns regarding the availability and accessibility of resources are reported to be challenges with accessing MH services across the globe [67]. Research has shown that school MHCPs tend to have variable or insufficient training coupled with a lack of support from administration, affecting their ability to respond to the diverse needs of students in the school setting [58]. Without the resources available to meet the needs of students who are identified as requiring more MH support, there is great concern that MH symptoms and risk of suicide will worsen [68]. Our findings suggest that early assessment and intervention, including the capacity to respond to the needs of young people and potentially refer them on to additional resources, are critical to the improvement of long-term health and social outcomes [69]. Introducing an eMH platform in a school setting may inevitably lead to the identification of youth who urgently need MH care, validating this concern. Alternatively, the Innowell eMH platform may also identify youth who are able to self-manage through the platform’s apps and electronic resources, preserving current MH services for those individuals in greatest need. Furthermore, gaining an understanding of the number of youth requiring MH services provides a starting point for advocating for increased publicly funded MH services. This advocacy can be done by capitalizing on the surge of interest in digital MH tools through building awareness regarding ways to modernize access to MH services, highlighting evidence of the effectiveness of digital MH tools, and advocating for financial investment [14].

When discussing facilitators to the implementation of eMH tools, participants described the capacity among young people to use technology to access services. We learned that participants view young people as having a strong desire to direct and manage their own care. The ability of young people to take an active role in their own care journey and access resources and information at times that work for them, whether independently or with their MHCP, all highlight the importance of youth capacity as a facilitator to the implementation of the Innowell eMH platform. In line with this finding, Iorfino et al [70] suggest that health systems will see an increased push toward youth owning and managing their own health data to the benefit of both youth and MHCPs. Many studies have addressed the importance of eMH tools as empowering their users by affording more control, choice [71], and shared decision-making [72]. By contrast, MHCPs who are early adopters of digital tools tend to be those who believe that the initiatives will be beneficial to their clients [73]. The scoping review by Hawke et al [74] demonstrated that youth prefer to provide feedback on the care they receive because they have a strong desire to be involved in decision-making. Playing an active role in their care enables young people to cope better, increases their sense of empowerment, and strengthens their connections with health professionals [75]. As MHCPs expressed enthusiasm around the implementation of the Innowell eMH platform, particularly when considering its potential benefits for their students, keeping youth at the center of eMH implementation strategies could inspire and enliven uptake among school MHCPs.

The findings of our study suggest that school MHCPs may use measurement-based care through an eMH platform to identify and monitor how students’ MH concerns are progressing and where new MH problems may be emerging. In fact, our study demonstrated that MHCPs view the measurement-based care protocol embedded in the Innowell eMH platform as providing critical information about a young person’s MH status, a common language for talking about MH concerns, and when changes in MH symptoms are a signal for adaptations in the intensity of services required. This common language as well as the ability to communicate, understand, and interpret their problems and strengths, subjective symptoms, and preferences for care with their MHCP could be established in their collaboration with other MHCPs as well [76]. The changes or lack of changes in the MH symptoms could open collaboration through thoughtful conversations between youth and their MHCP about the direction for care moving forward [77]. To inspire uptake among MHCPs, we support stressing the
importance of eMH tools enhancing and improving the way that youth and MHCPs interact with each other and the way that MH teams from different systems communicate, where possible [78].

This study further highlighted organizational implications to the implementation of eMH tools, recognizing the potential for increased access to community-based services and resources and available apps and e-tools. In a scoping review of MH service-level factors for access and engagement for young people, Anderson et al [79] note that choices around resources, increased information, flexible treatment delivery, and person-centered care contributed to young people being engaged in MH services. These benefits also extend to MHCPs and service delivery by helping to better connect young people to the services that they need and providing options to explore when their MHCP is unavailable. Study participants highlighted the potential for all members of the school MH team to share and access similar information about a student to ensure the continuity of care. Furthermore, participants also perceived school MHCPs being involved in the implementation of eMH tools as a turning point for MH services in the community. The unique relationship of school districts with municipal and provincial levels of government was seen as an implementation facilitator that could support significant transformation of youth MH services. Davenport et al [15] demonstrate how primary MH services can be “flipped” using digital health tools. Their work highlights the potential of digital MH services when assessment, triage, and care pathways are created to ensure that young people are matched to the care they need [15]. More recently, the Innowell eMH platform has been used to compare the needs of clients among services and across geographic regions, which can be used to understand the needs of a heterogeneous population and plan services accordingly [80]. This highlights the potential of eMH tools to ensure that youth are accessing the right care at the right time and promoting the continuity of care to avoid the fragmentation of services [81].

Our findings point to the potential of the school setting to support “flipping” youth MH service experiences and outcomes. The school setting is of particular importance, given the unique community-based relationships formed with young people and the potential to participate in deciding the direction of youth MH strategies with local and provincial governments, especially through the implementation of new eMH tools.

**Limitations and Strengths**

This study has a few limitations and strengths. Dynamics varied and differed among the FGs, with some groups having more representation from diverse stakeholders than others despite our best efforts to encourage diversity. An open discussion was embraced with each group; therefore, some groups spoke more organically about systemic and organizational implementation challenges than others. A potential limitation is that the clinical lead was often present in the FGs, which may have prevented some people from being forthright about the barriers and enablers specific to their community context as well as their views of the implementation of eMH tools and organizational culture. FGs may also result in certain types of socially acceptable opinions emerging and certain participants dominating. Therefore, this may not be the collective voice we expected. Some strengths of this study that we would like to highlight are the intentional involvement of youth research partners in the analysis of the FG data. This study also recruited from diverse school settings with respect to size, approach, and location (both urban and rural areas in the province). An opportunity for future research would be to target more diverse individuals and ask more targeted questions about organizational and systemic barriers and facilitators experienced by MHCPs and youth, respectively. Future studies could also use multiple methods to increase validity, such as observations and interviews. In addition, this gives way to include young people with lived experiences as participants, which was done as part of this study, and these results will be published in a subsequent manuscript.

**Conclusions**

This study sought to explore school MHCPs’ perspectives relating to the implementation of the Innowell eMH platform. Schools are a critical setting to implement eMH tools for youth. Our findings highlight the nuanced perspectives among MHCPs with regard to implementation. Their insights demonstrate critical youth and MHCP concerns, with considerations for organizational-level factors that may impede or enhance the implementation processes for embedding eMH tools in the school context. The identified barriers and facilitators to implementation in a school setting provide future researchers and decision makers with expected (and unexpected) challenges that could be addressed in the preimplementation phase.

**Acknowledgments**

The authors are grateful for the contributions from their research assistants Brea McLaughlin, Simron Sidhu, Katelyn Greer, Karina Pintson, and Sarah Daniel. The authors acknowledge the contributions of the team implementing the electronic mental health tools in coordinating and delivering the focus group presentations on the Innowell electronic mental health platform and providing manuscript feedback. The authors could not have conducted this research without the support of the various stakeholders and participants of this study. Finally, the authors would like to acknowledge the Alberta Children’s Hospital Foundation and the Partnership for Research and Innovation in the Health System for funding and supporting the project.

**Data Availability**

The data generated and analyzed during this study are restricted because the transcripts in entirety could potentially identify participants and are confidential.
Authors' Contributions

GD, EMB, and KSB wrote the initial draft of the manuscript. All authors participated in the final refinement of the draft, critically reviewed it, and provided feedback on the final version submitted for publication in accordance with the International Committee of Medical Journal Editors criteria.

Conflicts of Interest

HML is a Section Editor for JMIR Aging (at the time of this publication). All other authors declare no other conflicts of interest.

References


34. Kaplowitz MD, Hoehn JP. Do focus groups and individual interviews reveal the same information for natural resource valuation? Ecol Econ 2001 Feb;36(2):237-247 [FREE Full text] [doi: 10.1016/s0921-8009(00)00226-3]


47. Bierbooms J, Feijt MA, Jsselstijn WA, Bongers IM. Design of a game-based training environment to enhance mental health professionals' skills in using e-mental health: multiple methods user requirements analysis. JIMIR Serious Games 2022 Jul 27;10(3):e34700 [FREE Full text] [doi: 10.2196/34700] [Medline: 3589632]


64. Scott RE, Jennett PA, Hebert M, Rush B. Telehealth and e-health policy considerations for Alberta. Faculty of Medicine, University of Calgary Research and Training Program. 2004 Aug. URL: https://prism.ucalgary.ca/server/api/core/bitstreams/6906529d-4483-4284-9214-6512b5c657e4/content [accessed 2024-01-04]


**Abbreviations**

- **eMH**: electronic mental health
- **e-tools**: electronic tools
- **FG**: focus group
- **MH**: mental health
- **MHCP**: mental health care professional
- **TDF**: theoretical domains framework

---

Please cite as:

Dimitropoulos G, Bassi EM, Bright KS, Gondziola J, Bradley J, Fersovitch M, Stamp L, LaMonica HM, Iorfino F, Gaskell T, Tomlinson S, Johnson DW. Implementation of an Electronic Mental Health Platform for Youth and Young Adults in a School Context Across Alberta, Canada: Thematic Analysis of the Perspectives of Stakeholders. JMIR Ment Health 2024;11:e49099

URL: https://mental.jmir.org/2024/1/e49099
doi:10.2196/49099
PMID:38231558

©Gina Dimitropoulos, Emilie M Bassi, Katherine S Bright, Jason Gondziola, Jessica Bradley, Melanie Fersovitch, Leanne Stamp, Haley M LaMonica, Frank Iorfino, Tanya Gaskell, Sara Tomlinson, David Wyatt Johnson. Originally published in JMIR Mental Health (https://mental.jmir.org), 17.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Mental Health, is properly cited. The complete bibliographic information, a link to the original publication on https://mental.jmir.org/, as well as this copyright and license information must be included.
Original Paper

Reasons for Acceptance or Rejection of Online Record Access Among Patients Affected by a Severe Mental Illness: Mixed Methods Study

Julian Schwarz1,2,3, MD; Eva Meier-Diedrich1,3, MSc; Katharina Neumann1, MSc; Martin Heinze1,2,3, MD; Yvonne Eisenmann2,3, PhD; Samuel Thoma1,3, MD, PhD

1Department of Psychiatry and Psychotherapy, Center for Mental Health, Immanuel Hospital Rüdersdorf, Brandenburg Medical School Theodor Fontane, Rüdersdorf, Germany
2Center for Health Service Research Brandenburg, Brandenburg Medical School Theodor Fontane, Rüdersdorf, Germany
3Faculty of Health Sciences Brandenburg, Brandenburg Medical School Theodor Fontane, Neuruppin, Germany

Corresponding Author:
Julian Schwarz, MD
Department of Psychiatry and Psychotherapy
Center for Mental Health, Immanuel Hospital Rüdersdorf
Brandenburg Medical School Theodor Fontane
Seebad 82/83
Rüdersdorf, 15562
Germany
Phone: 49 33638 83 5
Email: julian.schwarz@mhb-fontane.de

Abstract

Background: Over the past few years, online record access (ORA) has been established through secure patient portals in various countries, allowing patients to access their health data, including clinical notes (“open notes”). Previous research indicates that ORA in mental health, particularly among patients with severe mental illness (SMI), has been rarely offered. Little is known about the expectations and motivations of patients with SMI when reading what their clinicians share via ORA.

Objective: The aim of this study is to explore the reasons why patients with SMI consider or reject ORA and whether sociodemographic characteristics may influence patient decisions.

Methods: ORA was offered to randomly selected patients at 3 university outpatient clinics in Brandenburg, Germany, which exclusively treat patients with SMI. Within the framework of a mixed methods evaluation, qualitative interviews were conducted with patients who chose to participate in ORA and those who declined, aiming to explore the underlying reasons for their decisions. The interviews were transcribed and analyzed using thematic analysis. Sociodemographic characteristics of patients were examined using descriptive statistics to identify predictors of acceptance or rejection of ORA.

Results: Out of 103 included patients, 58% (n=60) wished to read their clinical notes. The reasons varied, ranging from a desire to engage more actively in their treatment to critically monitoring it and using the accessible data for third-party purposes. Conversely, 42% (n=43) chose not to use ORA, voicing concerns about possibly harming the trustful relationship with their clinicians as well as potential personal distress or uncertainty arising from reading the notes. Practical barriers such as a lack of digital literacy or suspected difficult-to-understand medical language were also named as contributing factors. Correlation analysis revealed that the majority of patients with depressive disorder desired to read the clinical notes (P<.001), while individuals with psychotic disorders showed a higher tendency to decline ORA (P<.05). No significant group differences were observed for other patient groups or characteristics.

Conclusions: The adoption of ORA is influenced by a wide range of motivational factors, while patients also present a similar variety of reasons for declining its use. The results emphasize the urgent need for knowledge and patient education regarding factors that may hinder the decision to use ORA, including its practical usage, its application possibilities, and concerns related to data privacy. Further research is needed to explore approaches for adequately preparing individuals with SMI to transition from their inherent interest to active engagement with ORA.

Trial Registration: German Clinical Trial Register DRKS00030188; https://drks.de/search/en/trial/DRKS00030188
KEYWORDS
open notes; patient-clinician relations; electronic health record; clinical notes; visit notes; patient participation; online record access; mental illness; patient portal; mental health; qualitative interview; patient education

Introduction

In recent years, several countries have established secure patient portals to enable online record access (ORA), allowing patients to view their health data, including clinical notes (“open notes”) from their health care providers [1]. The United States, Canada, and Scandinavian countries, particularly Estonia, Sweden, and Norway, have been at the forefront, providing access to a significant number of patients across multiple regions [2,3]. More recently, the United Kingdom introduced the NHS app, offering access to primary care provider medical records since just last year [4]. In Germany, offering ORA was made mandatory for statutory health insurance providers by 2021, although the inclusion of open notes remains uncertain [5,6].

Research conducted in general health settings indicates clear benefits of patient access to clinical notes, including improved treatment satisfaction, transparency, patient engagement, patient-clinician communication, and health literacy [7-9]. Additionally, ORA enhances medication adherence and security in patients, helps patients to identify and correct treatment errors [10], increases a sense of control over the treatment, and reduces anxieties regarding the treatment process [11]. Health care providers generally view ORA as a valuable tool to promote patient engagement, even though it may be connected to an additional workload related to documentation and communication [12].

Studies conducted in mental health generally yield similar results to those in general health settings, but they also highlight distinct ethical and practical challenges [13,14]. For example, these challenges encompass navigating disagreements between patients and health care professionals (HCPs) regarding clinical notes, as well as discussing exceptions to or limitations of access for patient groups with specific diagnoses or acute conditions, such as severe mental illnesses (SMI) [15,16]. SMI is commonly characterized by conditions including (1) nonorganic psychoses, bipolar disorders, personality disorders, or severe chronic depression; (2) a prolonged psychiatric history involving multiple hospitalizations or outpatient treatments; or (3) moderate impairment in work and leisure activities alongside mild impairment in basic needs [17]. Clinicians often hold reservations about offering ORA to patients affected by SMI. Their concerns primarily revolve around the apprehension that ORA might contribute to self-harming or violent behaviors, especially in patients with whom establishing a trusting relationship is challenging [14]. Patient surveys, however, have demonstrated that individuals with SMI using ORA also exhibit improved understanding of medication prescriptions, higher medication adherence, and a greater sense of control over their medication [18]. Nevertheless, the adoption of ORA among patients with SMI remains lower compared to those receiving treatment for somatic conditions [18,19].

Apart from assumed positive effects on the patient-clinician relationship and concerns regarding data security, little is known about the motivations and expectations of patients with SMI toward ORA [14,20]. Examining these aspects is crucial for addressing the barriers that hinder the widespread acceptance of ORA among patients with SMI. Therefore, this study aims to thoroughly explore the reasons why patients with SMI consider or reject ORA and whether sociodemographic characteristics may influence patient decisions. More specifically, the following research questions will be addressed (1) which factors influence the decision of patients with SMI, who are offered access to their clinical notes, to either embrace or reject this option? (2) Does the decision for or against ORA in the context of SMI relate to any patient characteristics?

Methods

Design

This study is part of the PEPPPSY project (Piloting and Evaluation of a Participatory Patient-Accessible Electronic Health Record in Psychiatry and Somatics; 2021-2023) that focuses on piloting and evaluating a participatory patient record in psychiatry and somatic medicine [21,22]. It aims to examine the development, implementation, processes, and outcomes of the corresponding patient portal, also known as PEPPPSY, from the perspectives of patients and HCPs. Based on a concurrent mixed methods design, a qualitative methodology was used to comprehensively explore the reasons for and against ORA as well as the potential benefits and challenges [23]. This was complemented by a quantitative analysis to study the possible association between the decision for or against ORA and patient characteristics. Qualitative and quantitative data were analyzed separately. The concurrent design was chosen in order to obtain a comprehensive view of the research question [23], as it simultaneously provides in-depth, qualitative insights into the reasons for acceptance or rejection of ORA and broad, quantitatively measurable data on the patient characteristics in relation to this approval or rejection.

Ethical Considerations

The ethics approval from the Ethics Committee of the Brandenburg Medical School (MHB) was obtained (E-01-20210727), and the study was registered with the German Clinical Trial Register (DRKS00030188).

PEPPPSY App

The patient portal PEPPPSY was developed as part of a research collaboration between the Norwegian University of Science and Technology (NTNU) and the MHB. It emerged from an iterative process of participatory design, development, application, and evaluation [21,22]. PEPPPSY’s primary function is to provide patients with secure, 2-factor authenticated access to their physician’s notes, but it also includes other information such

https://mental.jmir.org/2024/1/e51126

JMIR Ment Health 2024 | vol. 11 | e51126 | p.298
as laboratory results and a list of prescribed medications. In the current second phase of the pilot, PEPPPSY is being expanded to serve a broader patient population, offering additional features such as a messaging function to allow communication between patients and clinicians concerning the open notes.

**Study Setting**

The study was conducted at 3 psychiatric outpatient university clinics of psychiatry and psychotherapy at MHB, Immanuel Klinik Rüdersdorf. Located in the metropolitan region Berlin/Brandenburg, the clinics are responsible for providing mental health care services to approximately 255,000 inhabitants in the catchment area. These outpatient clinics offer specialized care for patients who require a comprehensive, multidisciplinary approach due to the nature, severity, or duration of their conditions. The eligibility for receiving treatment in these psychiatric outpatient clinics is based on specific diagnoses, including SMI and other diagnoses, as determined by the insurance providers and the hospital association.

**Recruitment and Sampling**

From January to June 2023, eligible patients were randomly selected from the 3 study centers by the PEPPPSY research team consecutively. Participants had to meet the following inclusion criteria: age 18 years or older, diagnosed with SMI and confirmed by an external report, and currently receiving treatment in an outpatient psychiatric clinic. Exclusion criteria were previous ORA use related to mental health issues; acute psychiatric conditions or symptoms such as disorientation, severe delusions, hallucinations, katatonia, or agitation that may be associated with significant impairment of cognitive and social functioning; or acute self-harm or harm to others. Eligible patients were informed about the study and written informed consent was obtained. The former included detailed information about ORA, such as how it works, the health information it provides, and offers to participate in (1) this qualitative study and (2) the intervention part of the study, that is, to try out ORA for a period of several months in the study setting described.

**Data Collection**

The interviews that form the basis of the data in this study were conducted immediately after informed consent was obtained, which included a decision about whether or not participants wished to receive ORA. Sociodemographic data were collected, followed by face-to-face interviews. Information on the diagnoses of the study participants according to the International Classification of Diseases (ICD) was taken from the patient’s medical records. The interviews were performed at the aforesaid outpatient clinics by the authors and psychiatrists, ST and JS, who were not the outpatient treatment providers of the participating patients. The interviewers conducted the interviews out of genuine interest in understanding why ORA is accepted or rejected. For the interviews, a semistructured approach based on an interview guide was chosen in order to ensure the comparability of the interviews. This interview guide (see *Multimedia Appendix 1*) was developed deductively with the participation of all researchers on the basis of a desktop study (or literary research) on the topic of acceptance of ORA among patients with psychiatric disorders using Google Scholar and PubMed [24]. The interview guide was then tested in 2 pilot sessions with patients within the authors’ institution. Since no changes to the guide were necessary, the sample interviews could be included in the analysis.

The interviews explored each patient’s reasons for acceptance or rejection of ORA usage and had a mean duration of 11.3 (SD 4.5) minutes. In addition to the interviews, the researchers also took field notes, which were later included in the data analysis. Data collection continued until thematic saturation of categories was reached, which occurred when no new themes emerged from the transcripts. The saturation of categories was defined as the point at which no new codes appeared and the meaning of the category and subcategories were established [25].

**Data Analysis**

The interviews were transcribed verbatim and pseudonymized (JS and ST) and analyzed with thematic analysis (JS, ST, and KN using the MAXQDA Software (Verbi Software Ltd). Thematic analysis is a flexible approach that was used to inductively (“bottom up”) analyze data gathered from semistructured interviews [23]. Data analysis was initially conducted by 2 researchers on each interview individually and verified for consensus; a third person joined in when coders could not reach consensus. The analysis proceeded in six steps (1) familiarizing oneself or becoming familiar with the data, (2) generating initial codes, (3) generating initial themes, (4) reviewing themes, (5) defining and adequately naming themes related to the research questions, and (6) formulating key concepts. After all the themes were generated for each of the interviews, they were divided into 2 groups, reasons for acceptance or rejection of ORA. Subsequently, these themes were clustered within these groups and overarching categories and subcategories were formed. This was a recursive process where different categories were repeatedly tested for coherence and differentiability from the other categories and subcategories. In the final step, 2 researchers jointly selected the most relevant and succinct quotes from the subjects for each of the categories and subcategories. For the group of acceptance of ORA, 4 categories and 13 subcategories were formed. For the group of rejection of ORA, 5 categories and 13 subcategories were formed. For quality assurance purposes, the Consolidated Criteria for Reporting Qualitative research (COREQ) checklist was used (see *Multimedia Appendix 2*) [26].

The sociodemographic data of participants were analyzed according to their group affiliation (acceptance vs rejection ORA). Descriptive statistics were used to examine possible differences in sociodemographic characteristics between groups using R software (R Core Team) [27], which is available license-free. These between-group differences and their significance were assessed using chi-square test and t test [28]. No data were excluded from the data analysis.

**Results**

**Sociodemographic Data**

Out of the 124 eligible patients, 83.1% (n=103) agreed to participate in the study about reasons to use or not to use ORA. Sociodemographics are summarized in *Table 1*. The respondents...
had an average age of 46.1 (ranging from 19 to 86) years. The majority of the participants were women (n=64, 62%) with an average age of 47.2 years. Respondents who were men had an average age of 44.2 years. Among the approached patients, 58% (n=60) expressed a desire to use ORA, while 42% (n=43) declined ORA. When differentiating for gender, 56% (n=36) of the respondents who were women agreed to participate, while 44% (n=28) declined. Among respondents who were men, 62% (n=24) agreed to participate, while 38% (n=15) declined. The willingness to participate was highest among younger respondents (aged 18 to 39 years) and among patients aged 50 to 59 years. In terms of diagnosis, a high willingness to participate was observed among individuals with affective disorders (ICD 10, F3 [mood (affective) disorders]) at 91% (n=48; P<.001). The lowest agreement was found among individuals with schizophrenia, schizotypal, and delusional disorders (ICD 10, F2 [schizophrenia, schizotypal, and delusional disorders]) at 35% (n=6; P=.01).

Table 1. Sociodemographic characteristics of the study sample (N=103).

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>All (N=103)</th>
<th>Do you want online record access?</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender, n (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Women</td>
<td>64 (62.1)</td>
<td>Yes (n=60): 36 (56.2)</td>
<td>.68</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No (n=43): 28 (43.8)</td>
<td></td>
</tr>
<tr>
<td>Men</td>
<td>39 (37.9)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age (years), mean (SD)</td>
<td>46.06 (16.9)</td>
<td>Yes (n=60): 45.1 (16.44)</td>
<td>.51 b</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No (n=43): 47.4 (17.66)</td>
<td></td>
</tr>
<tr>
<td>Age (years), n (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18-29</td>
<td>20 (19.4)</td>
<td>Yes (n=60): 12 (60.0)</td>
<td>&gt;.99</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No (n=43): 8 (40.0)</td>
<td></td>
</tr>
<tr>
<td>30-39</td>
<td>23 (22.3)</td>
<td>Yes (n=60): 15 (65.2)</td>
<td>.34</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No (n=43): 8 (34.8)</td>
<td></td>
</tr>
<tr>
<td>40-49</td>
<td>15 (14.6)</td>
<td>Yes (n=60): 7 (46.7)</td>
<td>.25</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No (n=43): 8 (53.3)</td>
<td></td>
</tr>
<tr>
<td>50-59</td>
<td>20 (19.4)</td>
<td>Yes (n=60): 13 (65.0)</td>
<td>.61</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No (n=43): 7 (35.0)</td>
<td></td>
</tr>
<tr>
<td>≥60</td>
<td>25 (24.3)</td>
<td>Yes (n=60): 13 (52.0)</td>
<td>.35</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No (n=43): 12 (48.0)</td>
<td></td>
</tr>
<tr>
<td>Diagnosis, n (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>173 (100.0)</td>
<td>Yes (n=60): 116 (67.1)</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No (n=43): 57 (32.9)</td>
<td></td>
</tr>
<tr>
<td>F1c</td>
<td>20 (11.6)</td>
<td>Yes (n=60): 11 (55.0)</td>
<td>.32</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No (n=43): 9 (45.0)</td>
<td></td>
</tr>
<tr>
<td>F2d</td>
<td>17 (9.8)</td>
<td>Yes (n=60): 6 (35.3)</td>
<td>.01</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No (n=43): 11 (64.7)</td>
<td></td>
</tr>
<tr>
<td>F3e</td>
<td>53 (30.6)</td>
<td>Yes (n=60): 48 (90.6)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No (n=43): 5 (9.4)</td>
<td></td>
</tr>
<tr>
<td>F4f</td>
<td>36 (20.8)</td>
<td>Yes (n=60): 21 (58.3)</td>
<td>.23</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No (n=43): 15 (41.7)</td>
<td></td>
</tr>
<tr>
<td>F6g</td>
<td>15 (8.7)</td>
<td>Yes (n=60): 8 (53.3)</td>
<td>.25</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No (n=43): 7 (46.7)</td>
<td></td>
</tr>
<tr>
<td>Othersb</td>
<td>32 (18.5)</td>
<td>Yes (n=60): 22 (68.8)</td>
<td>.68</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No (n=43): 10 (31.2)</td>
<td></td>
</tr>
<tr>
<td>Number of diagnosis, mean (SD)</td>
<td>1.88 (0.87)</td>
<td>Yes (n=60): 1.95 (0.81)</td>
<td>.37 b</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No (n=43): 1.79 (0.94)</td>
<td></td>
</tr>
</tbody>
</table>

aN/A: not applicable.
bP values were calculated using t test, while all other values were calculated based on chi-square test.
cF1: Mental and behavioral disorders due to psychoactive substance use.
dF2: Schizophrenia, schizotypal, and delusional disorders.
eF3: Mood (affective) disorders.
fF4: Neurotic, stress-related, and somatoform disorders.
gF6: Disorders of adult personality and behavior.
hOthers: All mental disorders in the International Classification of Diseases-F chapter beyond those previously listed were subsumed under this category.

Reasons for Acceptance and Rejection of Participation in ORA

The categories and subcategories for the respective reasons provided by the respondents are summarized in Table 2.
Table 2. Categories and subcategories of stated reasons for acceptance and declination of ORA (N=103).

<table>
<thead>
<tr>
<th>Categories and subcategories</th>
<th>Values&lt;sup&gt;a&lt;/sup&gt;, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Reasons for acceptance of ORA&lt;sup&gt;b&lt;/sup&gt;</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Wish to engage in treatment</strong></td>
<td></td>
</tr>
<tr>
<td>Improved self-understanding and self-knowledge</td>
<td>14 (13.6)</td>
</tr>
<tr>
<td>Interest in the external perspective provided by clinicians</td>
<td>12 (11.7)</td>
</tr>
<tr>
<td>Continual contact and exchange</td>
<td>10 (9.7)</td>
</tr>
<tr>
<td>Incentive for increased engagement in treatment</td>
<td>6 (5.8)</td>
</tr>
<tr>
<td><strong>Understanding the treatment process</strong></td>
<td></td>
</tr>
<tr>
<td>Reminder of content discussed in therapy sessions</td>
<td>27 (26.2)</td>
</tr>
<tr>
<td>Ability to track the progress of treatment over time</td>
<td>5 (4.9)</td>
</tr>
<tr>
<td>Interest in medical translation of own symptoms</td>
<td>3 (2.9)</td>
</tr>
<tr>
<td><strong>Critically assessing clinicians</strong></td>
<td></td>
</tr>
<tr>
<td>Gaining more transparency into the perspective of health care providers on patients</td>
<td>6 (5.8)</td>
</tr>
<tr>
<td>Needing to verify the correctness of the notes in order to be able to trust the clinician</td>
<td>10 (9.7)</td>
</tr>
<tr>
<td>Avoiding and correcting misunderstandings</td>
<td>13 (12.6)</td>
</tr>
<tr>
<td><strong>Sharing personal health data with third parties</strong></td>
<td></td>
</tr>
<tr>
<td>Improving communication about the illness with significant others</td>
<td>7 (6.8)</td>
</tr>
<tr>
<td>Ability to share own health data with public institutions</td>
<td>3 (2.9)</td>
</tr>
<tr>
<td>Having access to own health data</td>
<td>2 (1.9)</td>
</tr>
<tr>
<td><strong>Reasons for rejection of ORA</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Feeling well supported in face-to-face interactions</strong></td>
<td></td>
</tr>
<tr>
<td>Sufficient oral “transmission” of notes</td>
<td>15 (14.6)</td>
</tr>
<tr>
<td>Desire to address problems and inquiries more effectively in direct conversation</td>
<td>10 (9.7)</td>
</tr>
<tr>
<td>Adequate satisfaction with in-person appointments</td>
<td>7 (6.8)</td>
</tr>
<tr>
<td><strong>Self-uncertainty</strong></td>
<td></td>
</tr>
<tr>
<td>Feeling emotionally burdened by reading the notes</td>
<td>16 (15.5)</td>
</tr>
<tr>
<td>Fear of excessive confrontation with one’s own condition</td>
<td>3 (2.9)</td>
</tr>
<tr>
<td>Adequate satisfaction with one’s own perspective</td>
<td>7 (6.8)</td>
</tr>
<tr>
<td><strong>Uncertainty in the relationship with the clinician</strong></td>
<td></td>
</tr>
<tr>
<td>Control weakens trust</td>
<td>2 (1.9)</td>
</tr>
<tr>
<td>Trust does not require control</td>
<td>19 (18.4)</td>
</tr>
<tr>
<td>Concern about technical demands for clinicians</td>
<td>1 (1.0)</td>
</tr>
<tr>
<td><strong>Concerns about the misuse of health data by third parties</strong></td>
<td></td>
</tr>
<tr>
<td>Worries about data security</td>
<td>5 (4.9)</td>
</tr>
<tr>
<td>Concerns of unwanted control by family members</td>
<td>2 (1.9)</td>
</tr>
<tr>
<td><strong>Practical barriers</strong></td>
<td></td>
</tr>
<tr>
<td>Difficulties in dealing with technology</td>
<td>13 (12.6)</td>
</tr>
<tr>
<td>Difficulties in reading and understanding the notes</td>
<td>7 (6.8)</td>
</tr>
</tbody>
</table>

<sup>a</sup>The number of patients (n, %) who mentioned each theme is indicated in parentheses.

<sup>b</sup>ORA: online record access.
Reasons for Acceptance

Overview
The stated reasons for accepting participation in ORA can be grouped into 4 main categories with a total of 13 subcategories.

Wish to Engage in Treatment
The respondents associated their agreement to participate with the wish and motivation to become more actively engaged in their treatment. They hoped that by using the portal, they could gain a better understanding of themselves and their condition, often driven by their interest in the external perspectives of their clinicians.

I am interested in knowing what they actually [think] about me here, because it’s about me, my health. Maybe I can understand everything [about why I’m feeling unwell] better. [Patient 14]

In the responses, this interest was often connected to a wish for ongoing communication and interaction.

I found it quite practical because it helps me stay in touch with my doctor and keep track of documentation. This way, I can tell the doctor when I’m not feeling well. [Patient 55]

Overall, the participants viewed their participation as an incentive to become more engaged in their treatment.

Overview of the Treatment Process
Gaining a comprehensive understanding of the treatment process is another category that emerged from the participants’ responses. While closely related to their willingness to engage in treatment, it primarily focused on the desire to have an overview of the treatment process. Many participants appreciated the ability to read their open notes through ORA, as it served as a reminder of therapy sessions and allowed them to track the chronological progression of their treatment.

I would like to have an overview of how my condition has changed over time, whether things have improved or worsened. Otherwise, you just live in the moment and with the things I tell you in this moment. But having it documented from appointment to appointment, and knowing that things might have gotten better without me realizing it, I would like to have that on paper. [Patient 89]

Additionally, embracing ORA was motivated by an interest in understanding how their own symptoms are translated in a medical context. This includes the use of specialized terminology to describe their symptoms and the subsequent treatment recommendations.

I would like to know how you medically process what I tell you during the treatment and what implications it has for the diagnosis. I’m just sharing things from my life, but what does it mean for the illness and what needs to be done now? I would like to see that. [Patient 84]

Critically Assessing Clinicians
Participants expressed their acceptance of ORA not only as a means to engage more actively in their treatment but also as an opportunity to critically assess the perspectives and approaches of clinicians. Participants valued the chance to gain more transparency about how clinicians view their patients by reading clinical notes and being able to provide feedback via comments.

To find out what therapists think about me behind my back and whether they even notice you in the hospital setting. [Patient 72]

Participants expressed concerns that clinicians might not accurately understand or document patients’ individual needs, leading to doubts about whether they can be trusted. This led to a need to verify the correctness of the notes in order to be able to trust their clinician.

I have so much mistrust towards doctors, especially regarding my psychosis and the forced medication, and how things I’ve said and done have been twisted. I can see it happening with my grandma too, how she’s being treated. That’s why I just want to see what you actually write down. [Patient 91]

From a more constructive perspective, many participants saw the possibility of viewing open notes within ORA as a way of preventing and rectifying misunderstandings that may arise during conversations.

Sharing Personal Health Data With Third Parties
This category describes aspects that are less focused on the treatment itself and its documentation, but rather on the use of this documentation with third parties. Participants expressed the hope that sharing the clinical documentation with significant others (eg, family members and friends) and other health care providers would improve the exchange of information about their illness.

For instance, my wife would like to talk to someone about how to handle my condition. That was my first thought, that she could also read what you write. I can’t remember and convey everything. This way, she could participate without me burdening her with it. [Patient 27]

Furthermore, participants viewed the ability to share their own health data with public institutions such as health insurance companies or the police through ORA as a positive aspect.

I recently had an issue with the health insurance company where they just declared me as healthy. They requested my medical records, but nobody was available at the psychiatric outpatient clinic. I could simply print out my documentation. That would be just great. [Patient 36]

Furthermore, the basic opportunity to have access to one’s own health data was also mentioned as a reason for accepting ORA.
Reasons for Rejection

Overview
The reasons for declining participation in ORA can be grouped into 5 categories with 13 subcategories.

Feeling Well-Supported in Face-To-Face Interactions
This category includes topics in which participants decline ORA because they already feel adequately cared for through the current mode of contact. For instance, they perceived their in-office appointments as sufficient for their needs.

\[\text{I am feeling satisfied with conversations. (\ldots) I'm not someone who spends a lot of time on their phone. I prefer being outdoors. (\ldots) If I don't understand something, I can simply ask for clarification. Looking at the notes would only add more information to my already busy mind.} \quad [\text{Patient 32}]\]

This feeling of being adequately cared for through in-office appointments was repeatedly associated with the desire of the respondents to address problems and inquiries through direct conversation. Moreover, they expressed a preference for discussing their own notes through verbal communication.

\[\text{We have already discussed it [the topic of today's session]. If there's anything or if I want to know more, I can always ask.} \quad [\text{Patient 12}]\]

Self-Uncertainty
Another central theme in the patients’ statements was the concern or fear of becoming unsettled by reading the notes. While several respondents mentioned being sufficiently satisfied with their own perspective on themselves, there was often an underlying fear of being burdened by reading the notes.

\[\text{I don't need to read that; I'm already experiencing all this crap myself. I don't need to see it in black and white too.} \quad [\text{Patient 78}]\]

In this context, some respondents specifically expressed fear of too much confrontation with their own condition, and some of the participants wished to leave that within the scope of the therapeutic conversation and not reactivate it through reading clinical notes.

\[\text{I wouldn't be up for that. Because, well, I unload all this stuff on you here that makes me sick, and afterwards, I actually feel better. But if I were to read through all that I've told you again, it would really bring me down all over again.} \quad [\text{Patient 67}]\]

Uncertainty in the Relationship to the Clinician
The respondents expressed concerns about not only their own self-uncertainty but also about feeling uncertain toward the health care provider and the therapeutic relationship when it comes to using ORA. Specifically, they highlighted that allowing patients to review their notes could potentially undermine the trust and rapport they have established with their clinicians.

\[\text{I am a doctor myself and I know that it harms the doctor-patient relationship when patients read what doctors write about them. It is very important to me that I trust you without constantly reviewing what you document.} \quad [\text{Patient 92}]\]

Contrary to the wish to critically assess the clinician as a justification for the use of ORA (as mentioned above), the respondents emphasized that satisfaction with treatment and a trusting relationship do not require such control.

\[\text{I trust you that everything is accurate, right? How you write it down. I'm really satisfied with the treatment, I've even had my pension extended, and all the services I need are being provided, so everything you document and how you communicate it to others must be correct, right? Others might want to know sometimes, but I also feel that what I tell you is being understood, so I don't need to read anything extra.} \quad [\text{Patient 73}]\]

Here, concerns were raised about the potential increase in workload for clinicians, which could potentially strain the therapeutic relationship due to the perceived additional workload.

Concerns About the Misuse of Health Data by Third Parties
Some respondents explained their rejection based on concerns about data security, specifically regarding the potentially insecure storage of documentation for instance on mobile devices, which could result in unauthorized access by third parties. Unlike the proponents of ORA, those who expressed opposition to it also raised concerns about unwanted control by care partners through unauthorized access to the patient portal.

\[\text{I have a curious girlfriend who doesn't necessarily need to read along. (\ldots) It's not for my family members, I would feel too controlled by my partner. She already opens my mail and goes through my bank statements.} \quad [\text{Patient 52}]\]

Practical Barriers
Finally, some respondents mentioned technical and practical challenges as reasons for their rejection. Specifically, participants over the age of 49 years highlighted the difficulty of dealing with the technology required for ORA, such as smartphones, browser apps, and 2-factor authentication. Additionally, some expressed feeling overwhelmed by the comprehension of the notes, as they encountered challenges due to the use of medical terminology and their own difficulties in reading caused by issues with concentration.

\[\text{I have really bad concentration problems, so that I can't understand anything anymore and can't fully engage in anything. I would only understand half of it, especially when reading. Additionally, I don't have internet access, and I don't understand how to set it up.} \quad [\text{Patient 97}]\]

Discussion

Principal Findings
In summary, the reasons provided by the interviewed patients with SMI for their decision to use or not use ORA are diverse.
Among those in favor, motivations range from a desire for increased engagement in treatment to critical evaluation of clinicians and using accessible health data for sharing with third parties. In contrast, those who opposed ORA perceived their therapeutic relationship as already well-established and feared that it might be jeopardized by the use of ORA. Finally, practical barriers, mostly related to digital literacy, were cited as reasons for their opposition.

Acceptance of ORA
The reasons for approval are briefly discussed as they largely align with those found in existing studies, thus providing limited implications for the further development of ORA. Those reasons include the motivation to engage more actively in treatment: by using the portal, patients hope to better understand the content of medical appointments and to obtain a clearer view of themselves, their illness, and the external perspective of their doctors [29]. Although not explicitly thematized in this study, it is reasonable to infer that this motivation also leads to increased adherence to medical treatment. This assumption is supported by another study, which found that patients with SMI when using ORA, reported an improved understanding of their medication prescriptions and described feeling more comfortable and in control throughout the therapeutic process [18,30]. However, this question requires further investigation in a follow-up study.

Moreover, many patients see ORA as a way to obtain a comprehensive overview of their treatment process. This includes accessing open notes as a reminder of therapy sessions, tracking treatment progress, and understanding how their symptoms are documented in a medical context.

Critical evaluation of clinicians is another reason for the acceptance of ORA by patients which is also reported by other studies [31]. The participants in our study reported that by reading the clinical notes, they want to evaluate the transparency and accuracy of clinicians’ perceptions and documentation of their needs. This critical view is also seen as an opportunity to address and correct potential misunderstandings that may occur during the consultation. However, this need for critical monitoring of the practitioners was ultimately linked to the desire to deepen trust in the practitioner and the treatment process. This need or desire to enhance trust is also highlighted in a study by Cromer et al [32].

In addition to patients appreciating gaining access to their health data through ORA, the ability to share personal health information with third parties, such as family, friends, other health care providers and public bodies, is also viewed positively. Again, this finding is consistent with preexisting literature on the perceived benefits of ORA by health care users [33].

Rejection of ORA
When considering the rejections of ORA among patients with SMI, differences compared to patients from general health settings become apparent. For example, I patient acknowledges having significant comprehension difficulties during direct patient-clinician interactions due to concentration problems and suspects that reading clinical notes would exacerbate the issue. This aligns with existing studies that suggest severe symptoms, which persist in daily life and tend to hinder participation in digital health interventions [34]. Furthermore, patients with SMI more often experience intersecting factors such as low educational attainment and language barriers [35], which was also repeatedly stated by study participants with regard to difficulties in dealing with technology and understanding the notes. Low educational attainment and language barriers can subject them to greater stress in patient-clinician interactions when trying to understand health care providers’ explanations [35]. Collectively, these individual limitations can lead to concerns and experiences not being adequately articulated, misunderstood, or possibly forgotten within the limited time available at an appointment. Consequently, this may result in less interest in accessing the notes made by the clinician. On the other hand, these individual limitations could also serve as an argument in favor of ORA: ORA offers the opportunity to enhance understanding of the patient-clinician relationship [36]. It can contribute to mitigating the mentioned disadvantages of inequities by extending the therapeutic interaction beyond physical encounters into the digital space where patients may feel less pressured to conform with the HCP and may also express themselves more easily than in the physical space. However, it is crucial that the clinical notes are written in a language that is relatable to everyday life and nonjudgmental [8,37]. This allows patients to reread and better understand the content discussed during previous appointments in preparation for an upcoming one. Furthermore, a messaging or commenting feature enables patients to ask questions or gather any unresolved concerns. Nonetheless, concern about being emotionally burdened by reading the notes was a common reason for deciding against ORA. Remarkably, these fears correspond with those expectations of HCPs, even though they were rarely confirmed after adopting ORA [3,7,8,12].

These arguments raise the question of whether patients with SMI, who experience daily limitations due to their symptoms, should be informed about the potential benefits of ORA in a more specific or repeated manner, and whether such adapted and improved information could potentially modify the approval rate. On the other hand, it might be that just the opposite is the case and that providers are particularly reluctant to share notes with this population and do in fact not routinely discuss open notes or encourage their clients to read them [38,39]. Unfortunately, this issue did not emerge from our data and further research is needed to clarify this question.

Then there is a group of patients who reject ORA because they already feel well taken care of in face-to-face interactions for various reasons. Studies examining the willingness to adopt digital health services explain the preference for direct patient contact, among other factors, through personality traits [34]. Extraversion, in particular, is considered a predictor of a lower likelihood of engaging with digital health services [40]. Individuals who displayed higher extraversion preferred meeting and connecting with the doctor in person. Some of the statements made by the participants convey a certain persistence in favor of nondigital means of communication (see subcategory “Sufficient oral ‘transmission’ of notes” in Table 2). In line with this, other findings indicate that personality traits associated

https://mental.jmir.org/2024/1/e51126
with resistance to change and openness to new experiences result in a lower adoption of digital health services [41]. Therefore, it would be interesting for further research to explore whether these corresponding personality traits align with the thematic trends found in our study. Beyond these considerations, the attitude of rejecting ORA seems to be explained in particular by the fact of enduring SMI. On the one hand, the hope of indirectly positively influencing one's own mental health through ORA may be reduced due to the length or severity of the course of illness [34]. On the other hand, in the examined health care system for individuals with SMI, assuming they are in a phase of predominant psychological stability, treatment contacts are rare (approximately 1-2 sessions of 15 minutes each in 3 months). As a result, the opportunities for exchange and the scope of exchangeable content through ORA are limited from the perspective of the surveyed patients [42]. However, it is worth noting that precisely because appointments are short and there are potentially many topics to be discussed (current status of well-being, medications, laboratory results, medication levels, etc.), ORA could provide patients with SMI with more space to exchange a wide range of information at a later time and overall enhance the therapeutic contact beyond the physical encounter.

Other participants expressed their rejection of ORA by explaining that the burden of their mental illness in their everyday lives was already substantial, leading them to decline any additional confrontation beyond their appointments at the outpatient clinic. This aspect seems to correspond with the preceding factor, suggesting that these patients are currently unable to dedicate any further (mental) capacity to engage with their chronic mental condition beyond medical appointments.

Comparison of Acceptance and Rejection of ORA

When comparing the reasons provided by patients for or against the use of ORA, several interesting contrasts become apparent. While some reasons for approval can be interpreted as a desire to deepen the therapeutic relationship, the opposition, in certain cases, stems from the apprehension that this therapeutic alliance may be undermined and jeopardized through the introduction of control (refer to subcategory “Trust does not require control” in Table 2). Conversely, patients who embraced ORA described a high need for control, which motivates their use of ORA. Accordingly, the use of ORA is perceived as an opportunity to critically review the HCP’s perspective and documentation, rather than blindly trusting them (see subcategory “Needing to verify the correctness of the notes in order to be able to trust the clinician” in Table 2). The disclosure of notes, in the optimal scenario, can thus be regarded as a demonstration of trust that allows for a deepening of the therapeutic relationship [42].

Another theme that underlies both the approval and rejection of ORA is the use of health data by third parties. This issue raises concerns about data security and the potential for unwanted control by family members when sharing information with significant others, health care providers, and public institutions. It is important to note that privacy and trustworthiness are among the most common reservations regarding ORA and digital (mental) health interventions in general, given the sensitive and potentially stigmatizing nature of the content involved [16,34,43]. A recent study conducted in Sweden provides evidence that these reservations are valid, as patients with mental illness experience significantly more attempts by unauthorized individuals to access their mental health records compared to patients in general health settings [44]. In our study, 1 patient expressed the misconception that health data are directly stored on their mobile phone. This misunderstanding highlights a knowledge gap where patients may not be aware that the data are actually securely stored remotely using 2-factor authentication, thereby aiming to prevent unauthorized access through the phone. However, the concerns expressed in our study once again emphasize the importance of data protection in the implementation of digital health platforms and the need for sufficient patient and provider education on this matter.

Differences Between Patient Groups

Generally, the proportion of patients willing to use ORA is approximately 60%, which is consistent with findings from previous studies [19,43,45]. However, the actual usage rate of ORA among patients is expected to be even lower [18]. The 2 groups, those in favor and those against ORA show little difference in terms of age, gender, distribution of diagnoses, and comorbidity, except for psychotic and depressive disorders. The higher levels of agreement and motivation among patients with depression align with findings from other studies [43,45], possibly due to a higher prevalence of socially desirable behavior in this patient group. The low approval rate among patients with schizophrenia is somewhat surprising compared to the existing literature. According to previous studies, patients with psychosis are generally very well able to use web-based interventions, exhibit positive attitudes toward them, and use the web-based more frequently to build their social networks compared to the general population [46-48]. The rejection of ORA in our study population could be attributed to reduced digital literacy, functional impairments caused by psychotic symptoms, as well as an approach to the illness characterized by internalized stigma and social withdrawal [49]. This social withdrawal has also been described as a protective mechanism against overly social and open interactions [50].

Strengths and Limitations

This is the first study that examines the reasons for the acceptance and rejection of ORA among patients with SMI in the German health care system. The investigation of these factors is crucial for advancing the implementation of ORA in the German-speaking region and can only be meaningful through a comparison with international research findings. Moreover, the study contributes to filling the research gap regarding the perspectives of individuals with SMI toward ORA.

One limitation of the study is that while a variety of reasons for rejecting ORA became apparent in the qualitative survey, raising further questions about factors such as digital literacy or the respondents’ social behavior, these factors were not explored in the quantitative survey. Similarly, comprehensive sociodemographic information such as educational level, socioeconomic status, or duration of mental illness was unfortunately not available in the data corpus. A follow-up study may be useful to further validate the qualitative data and to
analyze in-depth the role of other patient characteristics that contribute to the decision for or against ORA.

Another limitation of the study is that it did not present the proportion of patients who actually used ORA after having consented to do so. Since this study is based on baseline data from the PEPPPSY study [21], the analysis of usage patterns and effects of ORA is yet to be conducted. Another issue at first glance is the dichotomization of the results (see Table 2). The question arises as to how any positive attitudes of patients who reject ORA (and vice versa) were taken into account. For instance, patients who opt for ORA may still hold concerns regarding data privacy. However, this limitation was addressed by incorporating any sub-aspects of patient statements that are in opposition to their decision for or against ORA in the qualitative analysis of the data. This means that all patient statements and attitudes toward ORA were accounted for in our qualitative analysis independently from the patients’ decisions for or against ORA.

### Future Research

In addition to the research gaps identified above, further research is needed to address the unique needs of individuals with SMI in order to effectively facilitate maintained engagement with ORA. First, the extent to which patient characteristics and, in particular, psychiatric functional impairments, as well as concepts such as internalized stigma and social withdrawal, influence acceptance of ORA should be investigated further. Possible influences of personality traits such as extraversion or resistance to change on willingness to use ORA should also be considered.

Second, it should be investigated to a larger extent, whether the fear of possible adverse effects from reading the findings and clinical notes made available via ORA is confirmed in practice. Although studies to date tend to suggest otherwise, patients’ concerns should be taken seriously. In this context, research should be conducted on how to formulate clinical notes in a way that is both understandable and empathetic to patients without overburdening the available resources of practitioners. In this respect, there are preliminary indications of promising use of generative language models [51].

Third, there is no evidence on what cues, explanations, or motivations patients with SMI need from the medical team, and especially from their clinicians, to want to use ORA more. In this context, it should be investigated whether improved patient information about the benefits of ORA increases adoption rates.

At the same time, there seems to be a need to explore what skills HCPs need to acquire in order to formulate clinical notes in a way that is understood by patients and adds value, which involves adapting their communication style to align with patients’ familiar vocabulary rather than relying solely on technical medical terminology. Finally, actual rates of ORA use among patients with SMI compared to adoption rates and reasons for potential discrepancies should be explored.

### Conclusions

In general, patients with affective disorders (ICD 10, F3) showed high interest in ORA, whereas patients with schizophrenia, schizotypal and delusional disorders (ICD 10, F2) were less interested. It was mainly female patients of younger (18-39 years) and middle (50-59 years) age who agreed to receive ORA. Acceptance of ORA by patients with SMI stems primarily from a desire to be more actively involved in their care, to have a comprehensive view of their treatment process, and to evaluate the accuracy of physicians’ perception and documentation of their needs. This critical perspective is also seen as an opportunity to address and correct any misunderstandings that may have occurred during the consultation. The value placed on access to personal health information, combined with the ability to share that information with third parties, underscores the patients’ positive attitudes toward ORA.

Rejection of ORA by patients with SMI is primarily motivated by a sense of already being well supported by face-to-face interactions, as well as concerns rooted in their own insecurities. These range from fear of being unsettled by reading clinical notes to avoidance of excessive confrontation with one’s condition outside of the therapeutic conversations. Patients worry that the transparency created by ORA could undermine trust in their health care providers, especially given the additional workload for clinicians. Finally, data security risks and practical barriers such as lack of digital literacy and incomprehensible medical jargon contributed to the decision not to use ORA.

### Acknowledgments

This study was funded by the Brandenburg Medical School publication fund supported by the German Research Foundation and the Ministry of Science, Research and Cultural Affairs of the State of Brandenburg. ST received internal clinical scientist funding by the Brandenburg Medical School.

### Authors’ Contributions

JS and ST contributed to the study design and collected the data. KN, ST, and JS conducted data analysis. JS and ST wrote a first draft. EMD translated the study into the English language. Successive drafts were revised by EMD and YE. All authors critically reviewed and commented on the study.

### Conflicts of Interest

None declared.
References


33. Schwarz et al. JMIR MENTAL HEALTH 2024 | vol. 11 | e51126 | p.308 https://mental.jmir.org/2024/1/e51126


Abbreviations

COREQ: Consolidated Criteria for Reporting Qualitative research
HCP: health care professional
ICD: International Classification of Diseases
MHB: Brandenburg Medical School
NTNU: Norwegian University of Science and Technology
ORA: online record access
PEPPPSY: Piloting and Evaluation of a Participatory Patient-Accessible Electronic Health Record in Psychiatry and Somatics
SMI: severe mental illness

©Julian Schwarz, Eva Meier-Diedrich, Katharina Neumann, Martin Heinze, Yvonne Eisenmann, Samuel Thoma. Originally published in JMIR Mental Health (https://mental.jmir.org), 05.02.2024. This is an open-access article distributed under the terms
of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Mental Health, is properly cited. The complete bibliographic information, a link to the original publication on https://mental.jmir.org/, as well as this copyright and license information must be included.
Experiences of Patients With Mental Health Issues Having Web-Based Access to Their Records: National Patient Survey

Jonas Moll¹, PhD; Gunilla Myreteg², PhD; Hanife Rexhepi³*, PhD

¹Centre for Empirical Research on Information Systems, Örebro University School of Business, Örebro University, Örebro, Sweden
²Department of Business Studies, Uppsala University, Uppsala, Sweden
³School of Informatics, University of Skövde, Skövde, Sweden

*these authors contributed equally

Corresponding Author:
Jonas Moll, PhD
Centre for Empirical Research on Information Systems
Örebro University School of Business
Örebro University
Nova Building, 4th floor
Fakultetsgatan 1
Örebro, 70182
Sweden
Phone: 46 19 303878
Email: jonas.moll@oru.se

Abstract

Background: Sharing mental health notes through patient accessible electronic health records (PAEHRs) is controversial. Many psychiatric organizations and regions in Sweden have resisted the implementation, as clinicians worry about possible harms when patients are reading their notes. Despite the documented benefits of PAEHRs, there is still a lack of knowledge regarding whether patients with mental health issues could reap similar benefits of reading their notes as other patient groups.

Objective: The aim of the study is to examine the use, attitudes, and experiences of patients with mental health issues by reading their notes in the PAEHR and, moreover, whether their experiences differ from other patient groups, and if so, how.

Methods: A national patient survey was conducted with answers from 2587 patients from different patient groups. In total, 504 respondents (19.5%) indicated that they experienced a mental health disease. Answers from this patient group were compared to the answers from all other respondents. Survey questions related to attitudes, information usage, and effects on contacts with care were selected for analysis. Mann-Whitney U tests were used to detect groupwise differences.

Results: Patients with mental health issues use PAEHRs for checking that they have received the right care (mean_mental health 2.83, SD_mental health 1.39; mean_others 2.62, SD_others 1.37; P=.002) or suspected inaccuracies (mean_mental health 2.55, SD_mental health 1.34; mean_others 2.31, SD_others 1.30; P=.001), blocking access for professionals in other specialties (mean_mental health 3.43, SD_mental health 1.46; mean_others 3.04, SD_others 1.42; P<.001), and checking which care professionals have accessed their record (mean_mental health 4.28, SD_mental health 1.14; mean_others 4.05, SD_others 1.25; P<.001) to a significantly higher degree than other patients. On the other hand, the results show that a significantly lower proportion of patients with mental health issues (mean_mental health 3.38, SD_mental health 1.21; mean_others 3.52, SD_others 1.18; P=.02) believe that PAEHRs help them in shared decision-making compared to other patient groups.

Conclusions: Patients with mental health issues who took part in the survey, as a group, express some minor differences in both the use of the PAEHR and their experiences regarding its usefulness, as compared to other patients, as a group. This patient group shows a slightly higher interest in 2 types of use: checking for accuracy of care in the record and blocking access to mental health notes for professionals from other parts of the health care system. Compared to other patient groups, these patients are less likely to experience that the PAEHR is a support in shared decision-making. The study indicates that the benefits of PAEHR on a general level are the same for this patient group as for other patients. The study does not support clinicians’ worry about possible harm to this patient group. Further research is however needed.

(JMIR Ment Health 2024;11:e48008) doi:10.2196/48008
KEYWORDS
patient accessible electronic health record; patient portal; patient experiences; mental health; eHealth; national survey; digital mental health; digital health

Introduction

Patient accessible electronic health records (PAEHRs) aim to promote patients’ engagement with their care by giving patients direct access to their electronic health records (EHRs) through a national patient portal. Patients in around 20 countries worldwide, including Estonia, the Nordic countries, Australia, the United States, Canada, and England, are now offered web-based access to at least some of their EHRs. In Sweden, the PAEHR called “Journalen” was launched in 2012, when the region of Uppsala offered all citizens 18 years and older of age access to their EHRs through the national patient portal 1177 Vårdguiden. In 2015, Journalen was launched as the national system in Sweden for web-based access to clinical notes, and at the end of 2018, all regions had implemented Journalen. The PAEHR offers the patient access to his or her medical notes, prescribed medications, laboratory results, diagnosis, maternity care records, referrals, and vaccinations. Since health care in Sweden is governed by 21 autonomous regions with their own regulations, there are some regional differences concerning what type of information a patient can access and how soon (immediately or after 2 weeks). All regions offer patients access to visit notes from somatic care and test results.

Despite the documented benefits of PAEHRs [1-3], clinicians have raised concerns that patients could become confused or anxious by what they read [4]. The web-based access to mental health notes is especially controversial. The clinicians’ main argument is that in mental health, the information concerns sensitive topics that can have negative consequences for patients when they access their notes. In the study by Peck et al [5], several clinicians approved of the possibility to exclude patients from access, when they were considered too vulnerable. Different survey studies related to PAEHR and mental health suggest that clinicians worry about possible harms, and many health care professionals anticipate that patients will become confused, get angry, or decompensate when reading their notes [6,7]. Other studies report that patients with mental health can benefit from accessing their notes. Some reported benefits are increased feelings of engagement [8-10], feeling of control over their health, trusting their providers, taking better care of themselves, remembering their care plan, understanding better the rationale for medications, and being more likely to take their medications as prescribed [11,12]. A small number of studies have however found negative consequences for patients with mental health issues because of reading the mental health notes, such as feeling judged, worried, or offended [5,12-14]. A majority of studies [5,9,15] suggest that outpatient patients with mental health issues value reading their notes, that psychiatrists do not experience increased work burden or perceive negative outcomes, and that respectful, accurate mental health notes may enhance patient trust.

In Sweden, region Skåne made mental health notes accessible to adult patients from 2015. Since then, more regions have followed, and as of today, 17 of 21 regions share mental health notes through the PAEHR Journalen [16]. There are no other differences between patients with mental health issues and other patient groups regarding what information types you have access to and when. Thus, patients with mental health issues have the same access to notes from somatic care, test results, and other information types accessible in their region as all other patients. Attitudes among physicians were studied (along with other professional affiliations) before and after the implementation of PAEHR in region Skåne [7]. That study reported that some physicians were more careful with what they documented in the record, as a result of not knowing how the patient might interpret and use the information. Similar results were reported by Dobscha et al [6] and Denneson et al [17], who found that clinicians were less detailed and changed their tone of the notes when they knew that the patient might choose to read the notes. Respondents in the study by Petersson and Erlingsdottir [7] also indicated a fear of increasing tension between clinicians and the patient, which could manifest itself in threats and acts of violence. In their follow-up study, the professionals rather expressed that there were no changes in patient involvement after the implementation of PAEHRs [18].

Additionally, Denneson et al [17] reported that clinicians expressed concern that access to mental health notes "could damage the therapeutic relationship by exposing a disconnect between the patients’ in-person experience with their clinicians and the documentation they read in their notes." Mental health notes can, they argue, reveal aspects of the therapeutic process—such as clinical formulations and subjective impressions—which clinicians frequently do not communicate to their patients. Thus, a patient reading his or her notes could cause the patient to misinterpret the clinician’s notation, which could have negative effects on the patient, such as having feelings of being judged or stigmatized [17]. Moreover, patients with mental health issues are also generally considered a vulnerable patient group, which begs the question whether patients with mental health conditions can reap similar benefits of accessing their PAEHRs as other patient groups [19]. This review shows the topic to be controversial, and in practice, many psychiatric organizations resist implementation.

Empirical research is scarce, especially in Sweden. To our knowledge, there is no comparative research on how PAEHRs are perceived among patients with mental health issues as compared to other patient groups. Moreover, Petersson and Erlingsdottir [18] make a call for empirical research regarding the perspectives of patients with mental health issues toward PAEHRs. The aim of this paper is to examine the use, attitudes, and experiences of patients with mental health issues by reading their notes in the PAEHR and if their experiences differ from other patient groups. More practical knowledge is needed in this area as input to the ongoing debate regarding the possible benefits for patients in accessing their mental health notes through PAEHRs.
Methods

Ethical Considerations

The survey, which focused on attitudes toward and experiences of using Journalen, was approved by the regional ethical review board in Uppsala, Sweden (EPN 2017/045). The respondents were informed about the voluntary participation and the aim of the study as well as presented with standard consent that needed to be accepted before the survey could be started. No data were stored unless the respondent chose to submit the answers at the end of the survey. The data were anonymized by representing each respondent with a number. No incentives were offered for participation.

Study Design

This paper is based on data from an open anonymous self-completion digital national patient survey distributed to users of the Swedish PAEHR system, Journalen, through a link on the login page. Thus, all citizens who logged in to the service during the period that the questionnaire was accessible (June to October 2016) were potential respondents to the voluntary survey.

The survey included 24 questions with a combination of Likert-scale items, multiple-choice items, and free-text alternatives. The questions covered the following themes: attitudes and reactions, access to and usage of information, effects on contact with health care, information content, security and privacy, personal health information, and demographics. The theme “personal health information” included a question about which diagnosis group the respondents identified himself or herself to belong to. The respondents could choose between the alternatives of cancer, mental health, diabetes, high blood pressure, and others. The diagnoses of cancer, diabetes, and high blood pressure were specified as survey alternatives as they are the most common chronic conditions in Sweden. The alternative to mental health was included in order to address the ongoing debate in Sweden regarding whether psychiatric records should be made available and whether this patient group can benefit from accessing their medical record. Of the respondents, 504 people chose to identify themselves as belonging to the group of patients with mental health issues. This constitutes 19.5% (n=504) of the respondents who answered the survey. Globally, about 1 in every 8 people live with a mental disorder—most commonly, an anxiety or depressive disorder [20]. In Sweden, an official national health survey reported that 16% of respondents experienced severe mental difficulties, but that as many as 71% of respondents experienced feelings of anxiety or worry [21]. It is thus not remarkable or questionable that as many as 19.5% (n=504) of respondents in this study identified themselves to belong to this group of patients.

The full national survey was analyzed and presented by Moll et al [22]. In this study, 7 Likert-scale questions, including several items, related to attitudes, information usage, and contacts with care were selected for further analysis in relation to patients with mental health issues. Questions related to general attitudes and information usage were also picked out in order to shed light on any differences regarding how patients view and use the possibilities that Journalen gives. Finally, questions related to information accuracy, contact with care, and involvement in the care process were selected, since they reflect issues that mental health professionals have raised, as reported by previous studies. This set of selected questions is motivated by the need to develop knowledge that addresses the controversial question regarding access of patients with mental health issues to their records, and the set of questions relate closely to the concerns that were raised by health care professionals. The paper focuses on the answers of patients with mental health issues as a group and compares those to the answers from the other respondents as constituting another group.

During the time that the survey was distributed on the login page of Journalen, only 2 regions (Skåne and Kronoberg) had opened up web-based access to mental health notes. One consequence of this is that some of the patients with mental health issues who answered the survey could not yet access the mental health notes, while others could. Patients who lived in other Swedish regions could still access other types of health information (eg, test results and notes from primary care visits). During the survey period, 154 patients (30.6% of the mental health respondents) belonged to Skåne or Kronoberg and could thus access their mental health notes.

Analysis

Apart from descriptive analysis, Mann-Whitney U tests were used for detecting groupwise differences in answers on the 5-point Likert-scale questions between the group of patients with mental health issues and the group of all other respondents. The same test was used for detecting groupwise differences between mental health respondents from the regions Skåne and Kronoberg, who could read the mental health notes at the time the survey was open, and all other mental health respondents. This extra comparison was performed to investigate if the survey answers were affected by the fact that some of the mental health respondents could not actually access their mental health information but only information related to somatic care. Prior to the analysis, the Likert-scale options strongly agree, agree, neutral, disagree, and strongly disagree were converted to a numerical scale (1=strongly disagree and 5=strongly agree). No free-text questions were analyzed in this study. The SPSS software (version 25; IBM Corp) was used for all calculations.

Results

Result Presentation

In the tables in the following subsections, the response options “strongly agree” and “agree” have been combined for readability. For the same reason, “strongly disagree” and “disagree” were also combined. In Multimedia Appendix 1, the results for all response options are provided for completeness. As mentioned, some patients with mental health issues could access their mental health notes in Journalen at the time of the study, while others could not. Mann-Whitney U tests were used to check if there were any significant differences between patients with mental health issues who could and could not access their mental health notes. Significant differences were
only found for 2 of the survey questions in the study (regarding access to all types of record entries and access to log list). Hence, the vast majority of the results presented here were not affected by the respondents being able to access their mental health notes.

**Demographic Information**

Demographic information about the respondents is provided in Table 1 together with a comparison against demographic data from the other group of respondents. Chi-square tests were used to check for significant associations between the compared variables. The group of patients with mental health issues was a bit younger, which was expected based on current statistics [23]. Moreover, there was a female dominance that was however a bit stronger than the statistics would suggest. Additionally, the level of education is lower for the mental health group. No differences were found regarding previous work experience in health care.

**Table 1.** Demographic information for respondents who identified themselves as patients with mental health issues.

<table>
<thead>
<tr>
<th>Demographic</th>
<th>Mental health, n (%)</th>
<th>Others, n (%)</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Age</strong>&lt;sup&gt;a&lt;/sup&gt; (years)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18-25</td>
<td>75 (15)</td>
<td>96 (4.9)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>26-35</td>
<td>145 (29)</td>
<td>269 (13.8)</td>
<td></td>
</tr>
<tr>
<td>36-45</td>
<td>109 (21.8)</td>
<td>264 (13.5)</td>
<td></td>
</tr>
<tr>
<td>46-55</td>
<td>89 (17.8)</td>
<td>366 (18.8)</td>
<td></td>
</tr>
<tr>
<td>56-65</td>
<td>55 (11)</td>
<td>427 (21.9)</td>
<td></td>
</tr>
<tr>
<td>&gt;66</td>
<td>27 (5.4)</td>
<td>528 (27.1)</td>
<td></td>
</tr>
<tr>
<td><strong>Sex</strong>&lt;sup&gt;b&lt;/sup&gt;</td>
<td></td>
<td></td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Female</td>
<td>387 (78.3)</td>
<td>1242 (63.9)</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>100 (20.3)</td>
<td>698 (35.9)</td>
<td></td>
</tr>
<tr>
<td>Others</td>
<td>7 (1.4)</td>
<td>3 (0.2)</td>
<td></td>
</tr>
<tr>
<td><strong>Works or has worked in health care</strong>&lt;sup&gt;c&lt;/sup&gt;</td>
<td></td>
<td></td>
<td>.11</td>
</tr>
<tr>
<td>Yes</td>
<td>226 (45.4)</td>
<td>804 (41.4)</td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>272 (54.6)</td>
<td>1139 (58.6)</td>
<td></td>
</tr>
<tr>
<td><strong>Education</strong>&lt;sup&gt;d&lt;/sup&gt;</td>
<td></td>
<td></td>
<td>.004</td>
</tr>
<tr>
<td>Research education</td>
<td>11 (2.2)</td>
<td>64 (3.3)</td>
<td></td>
</tr>
<tr>
<td>Higher education ≥3 years</td>
<td>168 (33.8)</td>
<td>776 (39.7)</td>
<td></td>
</tr>
<tr>
<td>Higher education &lt;3 years</td>
<td>94 (18.9)</td>
<td>372 (19)</td>
<td></td>
</tr>
<tr>
<td>High school ≥3 years</td>
<td>112 (22.5)</td>
<td>296 (15.1)</td>
<td></td>
</tr>
<tr>
<td>High school &lt;3 years</td>
<td>56 (11.3)</td>
<td>192 (9.8)</td>
<td></td>
</tr>
<tr>
<td>Less than high school</td>
<td>32 (6.5)</td>
<td>127 (6.5)</td>
<td></td>
</tr>
<tr>
<td>No formal education</td>
<td>11 (2.2)</td>
<td>55 (2.8)</td>
<td></td>
</tr>
<tr>
<td>Others</td>
<td>13 (2.6)</td>
<td>72 (3.7)</td>
<td></td>
</tr>
</tbody>
</table>

<sup>a</sup>Patients with mental health issues: n=500; other patients: n=1950.

<sup>b</sup>Patients with mental health issues: n=494; other patients: n=1943.

<sup>c</sup>Patients with mental health issues: n=498; other patients: n=1943.

<sup>d</sup>Patients with mental health issues: n=497; other patients: n=1954.

**General Attitudes**

Both patients with mental health issues and all other patients are generally positive toward the Swedish PAEHR system, Journalen (Table 2). The vast majority of the respondents, no matter which of the 2 groups they belong to, believe that access to Journalen is good for them (Q3b), and that web-based access to medical records is generally a good reform (Q3a). The Mann-Whitney U tests showed significant differences between the 2 groups for questions Q3a (mean_mental health 4.73, SD_mental health 0.68; mean_others 4.80, SD_others 0.60; P=.01) and Q3b (mean_mental health 4.78, SD_mental health 0.64; mean_others 4.86, SD_others 0.51; P=.005), indicating a slightly less positive attitude among patients with mental health issues. The difference between the 2 groups is, however, very small, pointing to that the results are not that significant.

Regarding the content within the PAEHR Journalen, the respondents were asked to rate how accurate they believe the content is. This question was split into 2 aspects: whether the
information that is found in the record is correct (Q15a) and whether sufficient information was recorded (Q15b; Table 3). Both groups, patients with mental health issues and all other patients, gave a fairly high rating regarding the correctness of information (mean_mental health 3.98, SD_mental health 0.99; mean_others 4.22, SD_others 0.91; P<.001) and a lower score regarding the completeness of information (mean_mental health 3.32, SD_mental health 1.34; mean_others 2.85, SD_others 1.40; P<.001). The differences between the groups were statistically significant in both cases, indicating that patients with mental health issues were more inclined to think that information was complete but less convinced that the information was correct, as compared to all other patients. This being said, the average rating among patients with mental health issues for the question about completeness was fairly low (close to neutral), while it was higher (close to “agree”) for the question about correctness.

### Table 2. The results regarding general attitudes toward Journalen from patients with mental health issues and all other patients.

<table>
<thead>
<tr>
<th>Question</th>
<th>Mean_mental health (SD)</th>
<th>Mean_others (SD)</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>I believe that access to medical records online is generally a good reform</td>
<td>4.73 (0.68)</td>
<td>4.80 (0.60)</td>
<td>.01</td>
</tr>
<tr>
<td>I believe that access to “Journalen” is good for me</td>
<td>4.78 (0.64)</td>
<td>4.86 (0.51)</td>
<td>.005</td>
</tr>
</tbody>
</table>

### Table 3. The results regarding information accuracy in Journalen from patients with mental health issues and all other patients.

<table>
<thead>
<tr>
<th>Question</th>
<th>Mean_mental health (SD)</th>
<th>Mean_others (SD)</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>The content in the record reflects the information I think that health care has about me</td>
<td>3.98 (0.99)</td>
<td>4.22 (0.91)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>There is information about me that is missing in the record which I think should be there and that the staff should know</td>
<td>3.32 (1.34)</td>
<td>2.85 (1.40)</td>
<td>&lt;.001</td>
</tr>
</tbody>
</table>

### Accessing Patient Information

Regarding the respondents’ answers to why a patient uses Journalen (Q4a-h), the results show that the most common reasons for accessing Journalen, among patients with mental health issues, are to get an overview of the medical history and treatment (Q4b), to follow-up what has been said during a health care visit (Q4e), and to become more involved in the care (Q4h). The same holds true for all other survey respondents. The least common reason for access, for both groups, was to get an overview of relatives’ medical history and treatment (Q4c). Results for all these questions are presented in Table 4.

There were 4 reasons for access, where the analysis showed significant differences between patients with mental health issues and the other respondents (Q4a,b,d, and f; also in Table 4; detailed results for all items related to this question (Q4) are presented in Multimedia Appendix 1). Of these 4, to get an overview of medical history and treatment, Q4b got the highest mean score, where patients with mental health issues gave slightly lower ratings (mean_mental health 4.58, SD_mental health 0.78) than other patients (mean_others 4.65, SD_others 0.80; P=.001). Still, the results show that patients with mental health issues see this as one of the most important reasons for accessing the PAEHR. On the other hand, compared to the other respondents, patients with mental health issues gave slightly higher, significant, ratings for the following reasons to use: general interest (mean_mental health 3.86, SD_mental health 1.19; mean_others 3.66, SD_others 1.29; P=.002), insecurity of whether the care is right (mean_mental health 2.83, SD_mental health 1.39; mean_others 2.62, SD_others 1.37; P=.002), and suspicion of inaccuracies (mean_mental health 2.55, SD_mental health 1.34; mean_others 2.31, SD_others 1.30; P<.001). The differences are, however, not very large, and none of these 3 were marked as one of the most common reasons for access by any of the groups.

Respondents were also asked to rate items of their importance, in connection to being able to access Journalen (Q5). The respondents in the mental health group rated the following benefits of being able to access Journalen the highest: it makes me feel informed (Q5e), it makes me feel safe (Q5d), and it improves communication between medical staff and me (Q5a). The other respondents’ ratings gave similar results aside from that they rated Q5c (it improves the understanding of the condition) as one of the top 3 benefits instead of Q5d. It is also of interest to note that both groups of respondents gave a very low rating to the item it has no relevance (Q5j), indicating that the respondents generally see clear benefits from accessing Journalen. See Table 5 for detailed results.
### Table 4. Respondents’ answers to the question “Why do you use Journalen?” by patients with mental health issues and all other patients.

<table>
<thead>
<tr>
<th>Question</th>
<th>Mean_mental health (SD)</th>
<th>Mean_others (SD)</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mostly general interest</td>
<td>3.86 (1.19)</td>
<td>3.66 (1.29)</td>
<td>.002b</td>
</tr>
<tr>
<td>To get an overview of my medical history and treatment</td>
<td>4.58 (0.78)c</td>
<td>4.65 (0.80)c</td>
<td>.001b</td>
</tr>
<tr>
<td>To get an overview of my relatives’ medical history and treatment</td>
<td>2.06 (1.45)</td>
<td>2.21 (1.52)</td>
<td>.05</td>
</tr>
<tr>
<td>Because I am not sure if I got the right care</td>
<td>2.83 (1.39)</td>
<td>2.62 (1.37)</td>
<td>.002b</td>
</tr>
<tr>
<td>To follow up what has been said during a health care visit</td>
<td>4.47 (0.91)c</td>
<td>4.39 (0.99)c</td>
<td>.10</td>
</tr>
<tr>
<td>Because I suspect inaccuracies</td>
<td>2.55 (1.34)</td>
<td>2.31 (1.30)</td>
<td>&lt;.001b</td>
</tr>
<tr>
<td>To prepare for my health care visit</td>
<td>3.40 (1.35)</td>
<td>3.51 (1.33)</td>
<td>.11</td>
</tr>
<tr>
<td>To become more involved in my care</td>
<td>4.21 (1.10)c</td>
<td>4.28 (1.02)c</td>
<td>.39</td>
</tr>
</tbody>
</table>

aThe Mann-Whitney U test was used for statistical analysis.
bSignificant P values.
cThe most highly ranked options by mental health respondents and other respondents.

### Table 5. Respondents’ answers to the question “How important is it for you to be able to access patient information?” by patients with mental health issues and all other patients. Some items related to Q5 are only shown in Multimedia Appendix 1.

<table>
<thead>
<tr>
<th>Question</th>
<th>Mean_mental health (SD)</th>
<th>Mean_others (SD)</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>It improves communication between medical staff and me</td>
<td>4.21 (0.96)b</td>
<td>4.36 (0.92)b</td>
<td>&lt;.001c</td>
</tr>
<tr>
<td>It improves the understanding of the condition</td>
<td>4.18 (1.02)</td>
<td>4.26 (0.98)b</td>
<td>.12</td>
</tr>
<tr>
<td>It makes me feel safe</td>
<td>4.24 (0.99)b</td>
<td>4.22 (0.97)</td>
<td>.44</td>
</tr>
<tr>
<td>It makes me feel informed</td>
<td>4.59 (0.76)b</td>
<td>4.62 (0.72)b</td>
<td>.42</td>
</tr>
<tr>
<td>It leads to that I can take care of my health better</td>
<td>3.40 (1.16)</td>
<td>3.54 (1.13)</td>
<td>.02c</td>
</tr>
<tr>
<td>It leads to that I can take care of my relatives health better</td>
<td>2.39 (1.38)</td>
<td>2.52 (1.36)</td>
<td>.04c</td>
</tr>
<tr>
<td>It is essential that I am able to actively participate in decisions</td>
<td>3.31 (1.44)</td>
<td>3.47 (1.41)</td>
<td>.03c</td>
</tr>
<tr>
<td>It has no relevance</td>
<td>1.49 (0.96)</td>
<td>1.50 (0.93)</td>
<td>.52</td>
</tr>
</tbody>
</table>

aThe Mann-Whitney U test was used for statistical analysis.
bThe most highly ranked options by mental health respondents and other respondents.
cSignificant P values.

Some significant differences were however found between the groups (also in Table 5). Patients with mental health issues rated the following benefits of Journalen as of lower importance than did the other patients: improves communication with healthcare providers (<.001), enables better self-care (P=.02), improves the possibility to take better care of relatives (P=.04), and enables the essential possibility to participate in health decisions (P=.03). None of the stated items in Q5 were rated of higher importance by patients with mental health issues compared to other patients, with a significant difference between the groups. Among the items where significant differences could be identified, only item Q5a about improved communication was one of the most highly rated benefits of accessing Journalen.

Moreover, respondents were asked how important different information types and functions in Journalen are to them. The 3 functions or information types that the group of patients with mental health issues rated to be of highest importance to have access to were results of tests (Q17d), being able to read all types of record entries (Q17g), and overview of all healthcare contacts (Q17e). This rating corresponds well to the ratings from the group of other patients. Both respondent groups gave the lowest ratings to the importance of being able to communicate electronically with other patients (Q17o).

There were 18 items included in this survey question. For most of the items, no significant difference could be found between the 2 groups of respondents. The 6 items for which significant differences could be identified are presented in Table 6. Patients with mental health issues gave higher ratings to the importance of the following information types or functions: psychiatry records (<.001), all types of medical notes or record entries (P=.02), blocking professionals from access to certain information (<.001), and access to the log list (<.001). Of these, the differences were largest regarding the importance of having access to the psychiatry record (mean_mental health
For 2 of the information types or functions listed in the survey (access to all record entries and access to the log list), the responses from patients with mental health issues who could or could not access their mental health records differed significantly. In the case of access to all types of record entries, those who could access their mental health notes gave significantly higher ratings ($P=.049$), and in the case of the log list, this group of respondents gave significantly lower ratings ($P<.001$).

### Table 6. Respondents’ answers to the question “How important is it for you to have access to the following information which is wholly or partly based on information contained in “Journalen”? by patients with mental health issues and all other patients. Some items related to Q17 are only shown in Multimedia Appendix 1.

<table>
<thead>
<tr>
<th>Question</th>
<th>Mean_mental health (SD)</th>
<th>Mean_others (SD)</th>
<th>$P$ value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Referral (content and how it is handled in care)</td>
<td>4.50 (0.88)</td>
<td>4.62 (0.74)$^b$</td>
<td>.02$^c$</td>
</tr>
<tr>
<td>Results of tests</td>
<td>4.69 (0.76)$^b$</td>
<td>4.78 (0.61)$^b$</td>
<td>.01$^c$</td>
</tr>
<tr>
<td>Overview of all health care contacts</td>
<td>4.59 (0.79)$^b$</td>
<td>4.61 (0.77)</td>
<td>.60</td>
</tr>
<tr>
<td>Being able to read record entries from psychiatry</td>
<td>4.47 (1.09)</td>
<td>3.65 (1.38)</td>
<td>&lt;.001$^c$</td>
</tr>
<tr>
<td>Being able to read all types of record entries</td>
<td>4.68 (0.83)$^b$</td>
<td>4.64 (0.79)$^b$</td>
<td>.02$^c$</td>
</tr>
<tr>
<td>Ability to communicate electronically with other patients</td>
<td>2.15 (1.36)</td>
<td>2.02 (1.24)</td>
<td>.14</td>
</tr>
<tr>
<td>Ability to block certain medical records from access by other medical staff</td>
<td>3.43 (1.46)</td>
<td>3.04 (1.42)</td>
<td>&lt;.001$^c$</td>
</tr>
<tr>
<td>See which care units and staff groups have been inside “Journalen” (see log data)</td>
<td>4.28 (1.14)</td>
<td>4.05 (1.25)</td>
<td>&lt;.001$^c$</td>
</tr>
</tbody>
</table>

$^a$The Mann-Whitney $U$ test was used for statistical analysis.  
$^b$The most highly ranked options by mental health respondents and other respondents.  
$^c$Significant $P$ values.

### Relationship With Health Care and Patient Involvement

Two of the questions in the survey (Q7 and Q16) covered aspects of the patient’s relationship with health care and his or her involvement in the care process. Regarding possible changes in the patient’s relationship to health care (in general) after using Journalen (Q7a) and to health care professionals (more specific) due to communication about Journalen (Q7b and c) and its content (Q7d), no significant differences could be found between respondents in the 2 groups (Table 7). Here, it is again important to remember that not all of the respondents who answered in the role of patients with mental health issues, as of the date of the survey, had access to his or her psychiatry record. The results however show that all patients, regardless of group, experience at least a moderate positive effect on the relationship with health care. Furthermore, patients (regardless of group) and health care professionals generally do not talk about the possibility for the patient to use Journalen nor do they discuss its content.

Regarding patient involvement in the care process, some significant differences were found between the 2 groups (Table 8). Patients with mental health issues gave significantly lower ratings when it came to Journalen’s potential to support communication with medical staff ($P=.02$) and its potential to enable shared decision-making ($P=.02$). No significant differences were found regarding support for following prescription of treatment ($P=.53$) or support for self-care ($P=.69$). Overall, the results indicate that Journalen had at least a moderate positive effect in the involvement in the care process for patients with mental health issues, and the same holds true for the other respondents as well.
The aim of this study was to, through a national patient survey, investigate the experiences of patients with mental health issues with the Swedish PAEHR Journalen, as well as possible differences between patients with mental health issues and other patients, related to experiences with and attitudes toward the eHealth service. The paper contributes most and foremost to a much-needed knowledge about the effects of Journalen for a specific patient group—patients with mental health issues—several years after the launch of the service. Several important conclusions about aspects that patients with mental health issues value with regard to Journalen were identified in this study, and some interesting differences between the groups of patients with mental health issues and all other patients were brought to light in the comparative analysis. The results also reveal that, in most cases, patients with mental health issues see the same values in Journalen as other patients.

First, and on an overall level, it is clear from the results of the survey that respondents in the mental health group, as well as all other respondents, were positive toward being able to access personal health information in Journalen, and that there are no big differences between patients with mental health issues and other patients. These results are in accordance with earlier research [9,12]. These results are important, as health care professionals have raised concerns that patients with mental health issues in particular would become confused and agitated from reading their PAEHR [6,7]. Moreover, the results reveal that the group of patients with mental health issues is somewhat more critical toward the accuracy of the content compared to other respondents. A possible explanation for why more patients with mental health issues find inaccuracies could be that mental health conditions are more subjective and difficult to quantify and therefore may give rise to disagreements in how they should be described and documented. Since patients with mental health issues find more inaccuracies in the record, and since current research [6,17] has reported that some clinicians change the way they document as a result of patients reading their notes, it is of utmost importance that we open up a discussion regarding how notes could or should be written, and if, how, and when the patients should be involved.

Second, regarding the reasons for using the service as well as what information types were considered to be important to patients, there were no big differences between the 2 respondent groups. The reasons for use that were rated highest among mental health respondents are using Journalen for receiving an overview of one’s treatments, following up on what was said during a health care visit, and becoming more involved in the care process. This should be seen as an important result, since it shows that the reasons for implementing Journalen in the first place are as relevant for patients with mental health issues as they are for all other patient groups represented in the survey. Nevertheless, this patient group has been treated differently during the implementation process in that, for example, mental health notes were excluded in all regions during the first 3 years and were only accessible in 2 regions 6 years after launch. Results from the comparison between mental health respondents who could access their mental health notes and those who could not, show, interestingly enough, that there are no significant differences for any of the results related to the reasons for using Journalen. Hence, this study does not show any indication that access to these particular notes makes a difference in the reasons

### Discussion

**Principal Findings**

The aim of this study was to, through a national patient survey, investigate the experiences of patients with mental health issues with the Swedish PAEHR Journalen, as well as possible differences between patients with mental health issues and other patients, related to experiences with and attitudes toward the eHealth service. The paper contributes most and foremost to a much-needed knowledge about the effects of Journalen for a specific patient group—patients with mental health issues—several years after the launch of the service. Several important conclusions about aspects that patients with mental health issues value with regard to Journalen were identified in this study, and some interesting differences between the groups of patients with mental health issues and all other patients were brought to light in the comparative analysis. The results also reveal that, in most cases, patients with mental health issues see the same values in Journalen as other patients.

First, and on an overall level, it is clear from the results of the survey that respondents in the mental health group, as well as all other respondents, were positive toward being able to access personal health information in Journalen, and that there are no big differences between patients with mental health issues and other patients. These results are in accordance with earlier research [9,12]. These results are important, as health care professionals have raised concerns that patients with mental health issues in particular would become confused and agitated from reading their PAEHR [6,7]. Moreover, the results reveal that the group of patients with mental health issues is somewhat more critical toward the accuracy of the content compared to

### Table 7. Respondents’ answers to the question “How important is “Journalen” to make you feel that you are involved in your own care?” by patients with mental health issues and all other patients.

<table>
<thead>
<tr>
<th>Question</th>
<th>Mean_mental health (SD)</th>
<th>Mean_others (SD)</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>To take part of the patient information via “Journalen” has affected the relationship with health care system positively</td>
<td>3.80 (1.11)</td>
<td>3.88 (1.07)</td>
<td>.13</td>
</tr>
<tr>
<td>Medical staff has informed me about the possibility to read “Journalen”</td>
<td>1.82 (1.29)</td>
<td>1.87 (1.24)</td>
<td>.17</td>
</tr>
<tr>
<td>Medical staff has encouraged me to use the “Journalen”</td>
<td>1.70 (1.12)</td>
<td>1.72 (1.09)</td>
<td>.40</td>
</tr>
<tr>
<td>I discuss the content of “Journalen” with medical staff</td>
<td>2.53 (1.49)</td>
<td>2.52 (1.41)</td>
<td>.95</td>
</tr>
</tbody>
</table>

*a* The Mann-Whitney U test was used for statistical analysis.

### Table 8. Respondents’ answers to the question “To what extent do you agree with the following statements regarding your relationship with health care?” by patients with mental health issues and all other patients.

<table>
<thead>
<tr>
<th>Question</th>
<th>Mean_mental health (SD)</th>
<th>Mean_others (SD)</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Information in “Journalen” has helped me in communication with medical staff</td>
<td>3.57 (1.19)</td>
<td>3.71 (1.13)</td>
<td>.02</td>
</tr>
<tr>
<td>Information in “Journalen” had a positive impact on the ability to work together with medical staff making decisions about care and treatment</td>
<td>3.38 (1.21)</td>
<td>3.52 (1.18)</td>
<td>.02</td>
</tr>
<tr>
<td>Information in “Journalen” had a positive impact on the ability to follow the prescription of treatment</td>
<td>3.71 (1.22)</td>
<td>3.83 (1.16)</td>
<td>.05</td>
</tr>
<tr>
<td>Information in “Journalen” had a positive impact on the ability to take own steps to improve health</td>
<td>3.56 (1.22)</td>
<td>3.60 (1.18)</td>
<td>.69</td>
</tr>
</tbody>
</table>

*a* The Mann-Whitney U test was used for statistical analysis.

https://mental.jmir.org/2024/1/e48008
for using Journalen or the attitudes that patients with mental health issues have toward it.

There were some significant differences indicating specific needs that are related to patients with mental health issues, but most of these differences are still small. Hence, even though there were statistically significant differences between the groups, most of the results do not support that there, in practice, would be any big differences between patients with mental health issues and other patients. Patients with mental health issues gave a somewhat higher rating for possibilities of reviewing which of the health care professionals have read the content in the EHR (access to the log list) and for blocking professionals from other health specialties from accessing all information. They also reported somewhat higher feelings of insecurity regarding having received the right care.

The results showed that some patients might use Journalen because of insecurity about receiving the right care. This can be related to some of the concerns raised by mental health professionals. Both Petersson and Erlingsdóttir [7], from the Swedish perspective, and Dobscha et al [6], from the US perspective, reported the general concern that patients would request changes in the health record both due to found inaccuracies and notes that can be considered sensitive and that the patient might not agree with. These studies, from the perspective of health care professionals, report on health care concerns regarding the consequences of patients’ access to mental health notes. This survey’s results show that patients from this group indeed use Journalen to check whether they have received the right care. The perceived importance of the blocking feature in the system as well as the log list of who has accessed the record points toward an insecurity in who can access the information. This could possibly be due to the sensitive nature of the information related to mental health.

Third, when it comes to communication with health care, no big differences between patients with mental health issues and all other patients could be observed. Respondents with mental health conditions, as well as all other respondents, were generally positive regarding the effects on communication with health care, which is in line with existing research [22], but they gave lower ratings when it comes to communicating with health care professionals about the existence of Journalen. The fact that health care professionals generally do not inform patients about Journalen or encourage them to use the service has also been reported in earlier research but then concerning patients with cancer [24]. A reasonable interpretation of this neglect is that the earlier-mentioned concerns raised by health care professionals function as an obstacle.

Finally, with regard to effects on involvement in care, the results were also similar between the 2 groups. However, patients with mental health issues gave significantly lower ratings to actual effects on the relationship with health care and regarding shared decision-making. These results could possibly be related to the concerns on the effects of the therapeutic alliance that mental health professionals have raised [17]. In contrast to these results, a previous study showed that patients with cancer gave significantly higher ratings than all other patient groups and on all items regarding the effects of involvement in care [24].

Earlier studies, not focusing on specific patient groups [25], have shown that patients’ web-based access to medical records has improved the possibility for patients to engage in shared decision-making, something that Rexhepi et al [24] also showed for patients with cancer in Sweden. Similar studies of patients with mental health issues are very few to date. Petersson and Erlingsdóttir [18] showed that most of their answering professionals did not experience a higher patient involvement. This survey indicates that patients with mental health issues, regarding participation in decisions related to their care, do not experience the same positive effect as other patient groups. This conclusion is thus in line with the health professionals’ view [18]. This issue is clearly worthy of additional exploration.

**Limitations**

As already mentioned, at the time of the survey, only patients from the regions Skåne and Kronoberg could access mental health notes, since the introduction in 2013. Consequently, some of the patients with mental health issues who answered the survey could not yet access the mental health notes, while others in fact could. All patients with mental health issues could, however, access information on somatic care. An additional Mann-Whitney U test was used to compare answers between patients with mental health issues who could and could not access mental health notes. For most of the areas covered in this paper, no significant differences could be found between these 2 groups. In light of this study limitation, when it comes to actual experiences, the study is focusing more on experiences of the PAEHR among patients with mental health issues in general than mental health notes in particular. However, regarding attitudes, the study captures the ideas of patients with mental health issues regarding access to his or her mental health notes. It is also important to recognize that answers were gathered through self-report. A unique patient might have had contact with health care for numerous reasons. We cannot be sure whether there are patients with mental health issues in the “others” group who did not disclose their mental health status in the survey. Moreover, one could discuss how homogenous the mental health group is, given the diversity in both type of diagnosis and severity in symptoms that might be visible within the group. The survey did not capture this diversity.

**Conclusions**

The study was based on a national patient survey where 19.5% (n=504) of respondents indicated that they experienced a mental health disease. The objectives of the paper were to examine the use, attitudes, and experiences of patients with mental health issues by reading their notes in the PAEHR and, moreover, whether their experiences differ from other patient groups, and if so, how.

A first conclusion, on an overall level, is that patients with mental health issues are as positive in their attitudes toward the access of personal health information in Journalen as other patient groups. This conclusion agrees with previous research. A second conclusion is that patients with mental health issues use Journalen differently in 2 manners, as compared to other patient groups: they check the record for inaccuracies regarding care and information content and they tend to block access to mental health notes for professionals from other parts of the
health care system. These differences in usage were not known from previous research. A third conclusion is that patients with mental health issues have somewhat other experiences from Journalen than other patient groups, in that they are less likely to find it supportive of shared decision-making between themselves and their doctor. This was not known from previous research. A final conclusion is that the clinicians' worry about possible harm to this patient group does not find support by the current empirical evidence. Patients with mental health issues with access to their mental health notes reported the same positive attitudes toward Journalen as did patients with mental health issues with only access to their somatic health notes.

There are many previous studies on how patients access PAEHR and their attitudes to the introduction of such eHealth services. This study contributes with knowledge, through its comparative research design, on how PAEHRs are perceived by patients with mental health issues as compared to other patients. Further research is however needed in this area. For example, the study contributed with insights regarding different usage patterns. It would be valuable with empirical insights and explanations of why patients with mental health issues are somewhat more critical regarding the accuracy of the information content. From the perspective of professionals, previous research has predicted that patient access to mental health notes will have consequences on what and how the professionals write the notes. This study indicates interesting paths for further investigation of that issue. A practical implication is, however, that professionals do not need to be overly concerned about potential harm to the patients. Patients with mental health issues use Journalen and its information by the same reasons as other patient groups. Patients with mental health issues are as positive to the effects on communication with health care as other patient groups, which is in line with previous research.

Patients with mental health issues are a vulnerable group, where professionals anticipate that patients may get confused, judged, worried, or angry when reading their notes. This study did not find support for that. Other studies have reported benefits from accessing the mental health notes, such as feelings of increased engagement, control over their health, and trust toward the professionals. Finally, this study contributes with the insight that the group of patients with mental health issues finds it less possible to engage in shared decision-making as compared to other patient groups. Further research could help us better understand why we need to know more about the obstacles to patient participation and how Journalen can be used to better address practical issues related to feelings of engagement, control, and trust.

**Acknowledgments**

The authors would like to thank the colleagues in the national Development of Online Medical Records and E-Health Services consortium who participated in the data collection. The authors would also like to thank INERA and Evry for distributing the survey through the national patient portal.

**Authors' Contributions**

JM led the work and analyzed all the data. JM and GM did most of the study design, and all authors contributed equally in writing and editing the paper.

**Conflicts of Interest**

None declared.

Multimedia Appendix 1
Detailed results from all the survey questions covered in this study. [DOCX File, 53 KB - mental_v11i1e48008_app1.docx ]

**References**


**Abbreviations**

<table>
<thead>
<tr>
<th>EHR</th>
<th>electronic health record</th>
</tr>
</thead>
<tbody>
<tr>
<td>PAEHR</td>
<td>patient accessible electronic health record</td>
</tr>
</tbody>
</table>
Predictors of Use and Drop Out From a Web-Based Cognitive Behavioral Therapy Program and Health Community for Depression and Anxiety in Primary Care Patients: Secondary Analysis of a Randomized Controlled Trial

Background: A previously reported study examined the treatment of primary care patients with at least moderate severity depressive or anxiety symptoms via an evidence-based computerized cognitive behavioral therapy (CCBT) program (Beating the Blues) and an online health community (OHC) that included a moderated internet support group. The 2 treatment arms proved to be equally successful at 6-month follow-up.

Objective: Although highly promising, e-mental health treatment programs have encountered high rates of noninitiation, poor adherence, and discontinuation. Identifying ways to counter these tendencies is critical for their success. To further explore these issues, this study identified the primary care patient characteristics that increased the chances patients would not initiate the use of an intervention, (ie, not try it even once), initiate use, and go on to discontinue or continue to use an intervention.

Methods: The study had 3 arms: one received access to CCBT (n=301); another received CCBT plus OHC (n=302), which included a moderated internet support group; and the third received usual care (n=101). Participants in the 2 active intervention arms of the study were grouped together for analyses of CCBT use (n=603) because both arms had access to CCBT, and there were no differences in outcomes between the 2 arms. Analyses of OHC use were based on 302 participants who were randomized to that arm.

Results: Several baseline patient characteristics were associated with failure to initiate the use of CCBT, including having worse physical health (measured by the Short Form Health Survey Physical Components Score, \(P=0.01\)), more interference from pain (by the Patient-Reported Outcomes Measurement Information System Pain Interference score, \(P=0.048\)), less formal education (\(P=0.02\)), and being African American or another US minority group (\(P=0.006\)). Characteristics associated with failure to initiate use of the OHC were better mental health (by the Short Form Health Survey Mental Components Score, \(P=0.04\)), lower use of the internet (\(P=0.005\)), and less formal education (\(P=0.001\)). Those who initiated the use of the CCBT program but went on to complete less of the program had less formal education (\(P=0.01\)) and lower severity of anxiety symptoms (\(P=0.03\)).

Conclusions: This study found that several patient characteristics predicted whether a patient was likely to not initiate use or discontinue the use of CCBT or OHC. These findings have clear implications for actionable areas that can be targeted during
initial and ongoing engagement activities designed to increase patient buy-in, as well as increase subsequent use and the resulting success of eHealth programs.

**Trial Registration:** ClinicalTrials.gov NCT01482806; https://clinicaltrials.gov/study/NCT01482806

*JMIR Ment Health 2024;11:e52197*  doi:10.2196/52197

**KEYWORDS**
e-mental health; user engagement; initiation; discontinue; depression; anxiety; cognitive behavioral therapy; computerized CBT; online health community; collaborative care; internet support group

## Introduction

### Background

Currently, there is a significant gap in mental health treatment. For example, almost 75% of US patients who screen positive for depressive symptoms do not participate in treatment [1]. One reason for this gap is the reliance on in-person delivery models [2]. In-person services have inherent barriers, including challenges scheduling sessions, transportation limitations, symptom exacerbations limiting the ability to travel, illness-reduced motivation, stigma, and difficulties affording out-of-pocket costs [2]. Availability of and timely access to in-person services are also limited, with >90% of psychologists and psychiatrists and 80% of professionals with a master’s degree in social work practicing exclusively in metropolitan areas [3]. In addition, treatments may have limited effectiveness. Randomized clinical trials (RCTs) of pharmacotherapy and in-person psychotherapy for mental health disorders yield effect sizes that are typically small, with 0.30 standardized mean difference [4]. This contributes to the difficulty of finding effective treatments. A study on patient-reported helpfulness of mental health treatments found that only 26.1% were helped by the first treatment they tried [5]. Persisting to a second treatment resulted in a cumulative probability of feeling helped to 51.2%. After experiencing unhelpful treatment, patients would need to persist through up to 8 providers to reach a cumulative probability of 91%, obtaining a treatment that they found helpful [5].

These characteristics of in-person treatment contribute to high rates of noninitiation, poor adherence, and discontinuation. Even in RCTs, noninitiation and premature discontinuation average approximately 30% [6]. In practice settings, almost one-third of patients referred to psychotherapies do not initiate treatment, and those who do typically discontinue during the initial sessions (up to 44% after the first session and 82% by the fifth session) [7].

eHealth approaches offer promise in addressing many of these limitations by increasing convenience, reducing the level of effort and associated motivation involved in traveling to in-person treatments, overcoming limits on the availability of practitioners and treatments, reducing the obstacles involved in initiating and switching treatments, lowering relative costs, and reducing stigma. Despite their great potential and many successes [8-10], there are commonly encountered barriers, including the levels of user digital and health literacy required, lack of confidence [11], poor designs, lack of attention to cognitive design needs [12], limited technology access, and a lack of evidence-based methods to engage patients with eHealth programs [13-15]. Engagement has several different meanings in the literature [16]. The view taken in this paper is that a user is engaged when he or she (1) believes that using the intervention will result in positive changes that she or he values; (2) has the motivation, confidence, and ability to initiate use; and (3) can efficiently use the technologies and eHealth program. The 3 implications of this conceptualization are that engagement is a process, usability (actual and perceived) of the technologies and their information architecture influence engagement, and intervention methods should be designed to develop sufficient engagement of users initially and going forward. Evidence shows that many, eHealth programs too often fail to engage users sufficiently.

Consequently, noninitiation, discontinuation, and lack of adherence can be quite high [15,17]. In recognition of the significance of engagement in the success of eHealth, and how understudied this has been, Eysenbach [17] developed the “Law of Attrition.” This law highlights inadequate buy-in by users as a fundamental methodological challenge that must be addressed if eHealth programs are to be successful in practice. In one of his examples, 99.5% of the participants discontinued MoodGym, an online, evidence-based program for depression. A recent meta-analysis of RCTs on the treatment of depressive symptoms found an estimated dropout rate of 47.8% (after adjusting for publication bias) [18]. A meta-analysis of the real-world uptake of interventions for depression, mood enhancement, and anxiety found that 12% to 79% of those who downloaded an intervention did not use the intervention once; of the remaining who did initiate use, 58% to 93% had poor adherence; and 71.4% to 99.5% discontinued before completing 40% of the intervention [19]. Eysenbach [17] argued that there is a need to develop scientific theories that provide a better understanding of the causes of such phenomena and can form the basis for creating best practices to counter them.

To gain insights into possible influences on engagement, this study used data from the Online Treatment for Mood and Anxiety Disorders in Primary Care (Online Treatment) Trial [20] to describe the characteristics of participants who did not initiate versus initiated the use of each of the 2 e-mental health interventions and discontinued versus continued using one or the other intervention. Participants were given access to a self-help computerized cognitive behavioral therapy (CCBT) program (Beating the Blues). It was provided in a collaborative care framework that included support from a “care manager.” A subgroup of participants also had access to a password-protected online health community (OHC) platform.
which included moderated asynchronous internet support groups (ISGs).

Beating the Blues (CCBT), a web-based program, has been shown to be effective in treating depressive and anxiety symptoms [21-23]. The program is self-administered and consists of a brief 10-minute introductory video followed by eight 50-minute-long interactive modules. The modules’ topics are (1) problem definition and pleasurable events; (2) automatic thoughts; (3) thinking errors and distraction; (4 and 5) challenging unhelpful thinking; (6) core beliefs; (7) attributional style; and (8) review, action planning, and conclusion. Each module uses text, audio, audiovisual clips, and “homework” assignments designed to impart basic cognitive behavioral therapy techniques that target depression and anxiety. These 8 modules must be completed sequentially.

ISGs can improve users’ illness knowledge, coping skills, emotional support, connectedness, self-efficacy, and mental health [24-27]. Studies on users of web-based peer support have found that more frequent use results in more mental health benefits [28]. To explore the potential of peer support to enhance the effectiveness of CCBT, which is fully self-guided and does not offer the option for interaction with peers or providers, we included OHC. This provided access to several ISGs that the participants could use. However, nontherapeutic interactions can occur in these groups, which may result in negative consequences for users [29]. Therefore, it is important to use methods to keep the interactions supportive. The ISGs in this study had a moderator to manage content, facilitate supportive interactions, and amplify potential benefits [24].

Objective
The objective of this study was to identify patient baseline characteristics that were associated with noninitiation, poor adherence, discontinuing, or continuing to use each of 2 eHealth treatments for anxiety and depression, CCBT and OHC. This study conducted secondary analyses of data collected from the parent randomized controlled trial, Online Treatment for Mood and Anxiety Disorders in Primary Care [20]. The parent study provided CCBT and OHC to primary care patients with elevated depressive or anxiety symptoms.

Methods

Ethics Approval
The institutional review board of the University of Pittsburgh approved all study procedures (reference number: 20030187), and all participants provided written informed consent.

Overview
This study is a secondary analysis of data from the parent randomized controlled trial, the Online Treatment for Mood and Anxiety Disorders in Primary Care (Online Treatment) trial that examined the impact of access to 2 e-mental health interventions or their primary care physician’s (PCP’s) usual care (UC) on clinical outcomes [20]. The institutional review board of the University of Pittsburgh approved the study protocol, and all participants provided written informed consent.

Briefly, the parent trial recruited patients from 26 primary care offices affiliated with the University of Pittsburgh Medical Center, which shared a common electronic medical record system. A PCP received an electronic medical record system reminder about the study at the time of the clinical encounter for all patients aged 18 to 75 years with a diagnosis of depression, anxiety, generalized anxiety, or panic disorder. If the patient agreed to a study referral, they were contacted by a study recruiter via telephone to confirm protocol eligibility.

Participants
Eligible patients needed to have at least a moderate level of depressive or anxiety symptoms with a score 10 or greater on the 9-item Patient Health Questionnaire (PHQ-9) [30] or the 7-item Generalized Anxiety Disorder Scale [31]; reliable access to the internet, email, and telephone; and no alcohol dependence (as determined by the Alcohol Use Disorders Identification Test) [32], active suicidality, or other serious mental illness. Research assessors then administered via telephone the baseline battery and collected information on patients’ self-reported race, sex, and other sociodemographic characteristics. This study was registered to ClinicalTrials.gov under trial registration number NCT01482806.

Randomization Procedure
Following the baseline assessment, patients were randomized in a 3:3:1 ratio to (1) access to a self-guided CCBT program (CCBT-only [Beating the Blues]; n=301); (2) CCBT plus access to a password-protected OHC platform that included moderated ISGs (CCBT+OHC; n=302); or (3) UC of their PCP (n=101). The present report only used data from participants randomized to the CCBT-only and CCBT+OHC intervention groups (N=603).

Interventions

Computerized Cognitive Behavioral Therapy
The Beating the Blues web-based program (described earlier) served as the CCBT intervention [33]. CCBT was provided in a collaborative care framework via support from a care manager. The care managers encouraged participants to complete a module every 1 to 2 weeks, but the participants were free to proceed at their own pace.

Online Health Community
The OHC was password-protected and featured collections of links to external resources (eg, crisis hotlines, “find-a-therapist” sites, local US $4 generic pharmacy programs) and brief YouTube videos on stress management, sleep hygiene, meditation, exercise benefits, and nutrition. In addition, the participants could interact with one another on a variety of moderated ISGs. Each ISG was dedicated to a specific topic [34]. Unlike the CCBT, the OHC was designed so that participants had freedom of choice regarding the resources they used.

A study investigator logged into the ISGs daily to review new posts and monitor for the presence of suicidal thoughts or potentially inappropriate content. In addition, the OHC had a moderator who oversaw all communications by participants. The moderator was a care manager (described below) for this
study. To promote participants’ ongoing involvement with the OHC a variety of strategies were used: (1) weekly emails from the moderator that highlighted new discussions, new content, or self-management tips; (2) status indicators on participants’ profiles and comments they posted to a discussion (eg, stars and “likes”); (3) automatically generated email notifications of new ISG activities, for example, when someone replied to a participant’s ISG post or comment; (4) automated highlighting of recent comments on participants’ home pages, which were personalized to their ISG profile and past comments; (5) invited participants as guest moderators; and (6) initiating contests (eg, scavenger hunts, respond to emails or posts). During regular care management contacts, the care manager directed the participants to pertinent content on the OHC. For safety, at least one study clinician logged into the ISGs daily to monitor user-generated posts and comments, and the participants were able to “flag” potentially inappropriate content for removal.

Collaborative Care Framework: Care Managers Promote Engagement, Ongoing Involvement, and Effectiveness

Both eHealth programs were delivered via a collaborative care strategy. The 2 care managers had a bachelor’s or master’s degree in the psychology field and had worked on conducting mental health research with human participants. After randomization, a care manager dedicated to each intervention arm contacted the patient for an introductory telephone call and provided guidance in the setup of access to the CCBT program and the OHC, if applicable. During the 6-month intervention, participants were encouraged to complete a CCBT module every 1 to 2 weeks and were provided reminders, if necessary, encouragement on their progress, and personalized feedback. Participants had the option to contact a care manager with questions or for assistance. Care managers, irrespective of eHealth program usage, monitored patients’ progress, symptoms, use of the eHealth programs, and telephoned those who either were not doing well, as indicated by their scores on the PHQ-9, or had failed to log onto the CCBT or OHC regularly to inquire why. They entered all contacts and information gathered into an electronic registry that was used to track each patient’s progress and guide care managers through their contacts. If a patient’s symptoms did not improve or worsen, they contacted her or him via email or telephone (depending on the patient’s preference) to discuss additional treatment options, depending on the patient’s preference. At weekly case review meetings, with the assistance of the electronic registry, the care managers presented their patients to the study clinical team, which consisted of a PCP, psychiatrist, and psychologist-study coordinator (study team). In addition to providing patients with suggestions for general lifestyle adjustments, including social engagement, exercise, adequate sleep, and nutrition suggestions, the study team also recommended initiation or modification of antidepressant or anxiolytic pharmacotherapy based on patients’ symptom response and treatment preferences and referral to a mental health specialist when a patient had either complex psychosocial issues or was not responding to treatment. Following the case review, the care managers discussed recommendations with the patient and then notified the patient’s PCPs of progress and treatment suggestions. PCPs remained responsible for the treatment and were free to continue or modify the patients’ treatments.

Assessments

Blinded telephone assessors administered the assessment battery at baseline, 3 months, 6 months (end of interventions), and 12 months to assess the durability of the interventions. It included the administration of the 12-Item Short Form Health Survey (SF-12) assessment of SF-12 Mental Component Summary (MCS; SF-12 MCS) and Physical Component Summary (PCS; SF-12 PCS) scores, composed of 6 items each from the SF-12 to measure these 2 aspects of health-related quality of life [35]; the PHQ-9, a fixed-length Patient-Reported Outcomes Measurement Information System (PROMIS) measure to assess the severity of depressive symptoms [30]; the Generalized Anxiety Disorder 7-item scale to assess anxiety severity [31]; the 8-item fixed-length PROMIS-PI to assess pain interference, that is, the degree to which pain interferes with an individual’s physical, mental, and social daily activities [36]; and the Primary Care Evaluation of Mental Disorders to provide diagnoses of depressive and anxiety disorders [37]. In addition, an 11-item shortened version of the Pew Internet Use questionnaire was administered at baseline to assess participants’ use of the internet; server logs were abstracted to measure the use of the CCBT and OHC programs; and care managers’ electronic registry was used to assess the number of intervention emails and telephone contacts.

Statistical Analyses

As the purpose of this study was to identify subgroups of participants based on their CCBT or OHC use patterns, we limited the analyses to the CCBT alone and CCBT+OHC treatment arms (N=603). As documented in a previous study, both were more effective than PCPs’ UC for depression and anxiety but were similarly effective to each other (ie, offering access to the OHC produced no added reduction in depressive or anxiety symptoms over CCBT alone) [20]. Consequently, for CCBT analyses, we combined CCBT users from both arms (N=603).

We first classified the participants according to whether they initiated the use of the CCBT program. Participants who logged in at least once were categorized as having initiated CCBT use, whereas participants who never logged in comprised the noninitiation group. We compared baseline sociodemographic, clinical, and functional status measures by initiation status using t tests for continuous data and chi-square tests for categorical data. Fisher exact test was used for categorical measures when the expected cell counts were less than 5. Participants were then stratified by whether they “continued” the use of CCBT (yes or no), as those who completed at least one CCBT module. Those who did not complete the first CCBT module were categorized as “discontinued” use. Baseline characteristics were compared between these 2 groups (ie, continued and discontinued use) using t tests and chi-squared tests (or Fisher exact test, as appropriate). On the basis of previous work with this data set that found differences in the effects of CCBT between African American participants and White participants [38], we also examined whether there might be differences in the use patterns.
Participants in the CCBT+OHC study arm were then classified as initiation or noninitiation of OHC. Those who logged in at least once were considered to have initiated OHC use, whereas those who never logged in were categorized as noninitiators of OHC use. The baseline measurements were compared between the 2 groups. Continuation of OHC use was evaluated by stratifying OHC participants into 3 mutually exclusive categories: those who logged in only once, those who logged in 2 to 3 times, and those who logged in more than 3 times during the 6-month intervention phase. The baseline characteristics were compared across these groups. Finally, continuation of OHC use was classified by the number of months during which participants logged in to the OHC: logged in during 1 month only, logged in during 2 or 3 months, and logged in over a period of more than 3 months during the intervention phase. The participant characteristics were compared between the groups. All statistical analyses were performed using R version 3.6.3 (R Foundation for Statistical Computing). A significance level of $\alpha$ of .05 was assumed, and no adjustments were made for multiplicity.

### Results

#### Participant Description

Participants (N=603) in this study had a mean age of 42.8 (SD 14.2) years, were 79.6% (n=480) female, 82.8% (n=499) were White, 82.4% (n=497) had some college education, and 69.8% (n=421) were employed (Table 1). At baseline, they reported a mean SF-12 PCS of 50.9 (SD 12.3), which is similar to that of the general US population, 50 (SD 10) [35]. The mean SF-12 MCS was 31.5 (SD 8.9), which is considerably lower than the 50 (SD 10) for the general US population [35]. The participants’ mean PROMIS depression and anxiety scores were 62.3 (SD 12.3), and 65.9 (SD 12.3), respectively. Both indicate a higher severity of symptoms than the general US population, which had a mean score of 50 [39]. The 2 intervention arms did not differ in their baseline sociodemographic or clinical characteristics by random treatment assignment (all $P$$\geq$0.4, previously published) [20].
Table 1. Baseline characteristics for intervention initiators and noninitiators.

<table>
<thead>
<tr>
<th>Measure and category</th>
<th>Total (n=603)</th>
<th>Computerized cognitive behavioral therapy</th>
<th>Online health community</th>
<th>$P$ value</th>
<th>$P$ value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Initiation (n=521)</td>
<td>Noninitiation (n=82)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Demographics</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age (y), mean (SD)</td>
<td>42.8 (14.2)</td>
<td>42.8 (14.2)</td>
<td>43.1 (14.4)</td>
<td>.84</td>
<td>.38</td>
</tr>
<tr>
<td>Sex, n (%)</td>
<td></td>
<td></td>
<td></td>
<td>.94</td>
<td>.31</td>
</tr>
<tr>
<td>Female</td>
<td>480 (79.6)</td>
<td>415 (79.7)</td>
<td>65 (79.3)</td>
<td>182 (79.8)</td>
<td>63 (85.1)</td>
</tr>
<tr>
<td>Male</td>
<td>123 (20.4)</td>
<td>106 (20.3)</td>
<td>17 (20.7)</td>
<td>46 (20.2)</td>
<td>11 (14.9)</td>
</tr>
<tr>
<td>Race, n (%)</td>
<td></td>
<td></td>
<td></td>
<td>.006</td>
<td>.09</td>
</tr>
<tr>
<td>White</td>
<td>499 (82.8)</td>
<td>441 (84.6)</td>
<td>58 (70.7)</td>
<td>188 (82.5)</td>
<td>54 (73)</td>
</tr>
<tr>
<td>African American</td>
<td>91 (15.1)</td>
<td>69 (13.2)</td>
<td>22 (26.8)</td>
<td>34 (14.9)</td>
<td>19 (25.7)</td>
</tr>
<tr>
<td>Other</td>
<td>13 (2.2)</td>
<td>11 (2.1)</td>
<td>2 (2.4)</td>
<td>6 (2.6)</td>
<td>1 (1.4)</td>
</tr>
<tr>
<td>Living situation, n (%)</td>
<td></td>
<td></td>
<td></td>
<td>.06</td>
<td>.60</td>
</tr>
<tr>
<td>Alone with child</td>
<td>82 (13.6)</td>
<td>64 (12.3)</td>
<td>18 (22)</td>
<td>23 (10.1)</td>
<td>11 (14.9)</td>
</tr>
<tr>
<td>Alone no child</td>
<td>114 (18.9)</td>
<td>100 (19.2)</td>
<td>14 (17.1)</td>
<td>45 (19.7)</td>
<td>15 (20.3)</td>
</tr>
<tr>
<td>Living together with child</td>
<td>106 (17.6)</td>
<td>89 (17.1)</td>
<td>17 (20.7)</td>
<td>45 (19.7)</td>
<td>11 (14.9)</td>
</tr>
<tr>
<td>Living together with no child</td>
<td>300 (49.8)</td>
<td>267 (51.3)</td>
<td>33 (40.2)</td>
<td>115 (50.4)</td>
<td>37 (50)</td>
</tr>
<tr>
<td>Working, n (%)</td>
<td></td>
<td></td>
<td></td>
<td>.85</td>
<td>.15</td>
</tr>
<tr>
<td>Employed</td>
<td>421 (69.8)</td>
<td>363 (69.7)</td>
<td>58 (70.7)</td>
<td>159 (69.7)</td>
<td>45 (60.8)</td>
</tr>
<tr>
<td>Other</td>
<td>182 (30.2)</td>
<td>158 (30.3)</td>
<td>24 (29.3)</td>
<td>69 (30.3)</td>
<td>29 (39.2)</td>
</tr>
<tr>
<td>Education, n (%)</td>
<td></td>
<td></td>
<td></td>
<td>.02</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>High school or less</td>
<td>106 (17.6)</td>
<td>84 (16.1)</td>
<td>22 (26.8)</td>
<td>25 (11)</td>
<td>23 (31.1)</td>
</tr>
<tr>
<td>Attended college but did not receive a 4-year degree (also business or technical school)</td>
<td>216 (35.8)</td>
<td>184 (35.3)</td>
<td>32 (39)</td>
<td>86 (37.7)</td>
<td>24 (32.4)</td>
</tr>
<tr>
<td>College degree or higher</td>
<td>281 (46.6)</td>
<td>253 (48.6)</td>
<td>28 (34.1)</td>
<td>117 (51.3)</td>
<td>27 (36.5)</td>
</tr>
<tr>
<td>Clinical characteristics</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Short Form Health Survey Physical Components Score, mean (SD)</td>
<td>50.9 (12.3)</td>
<td>51.4 (12.1)</td>
<td>47.7 (12.8)</td>
<td>.01</td>
<td>51.3 (12.5)</td>
</tr>
<tr>
<td>PROMIS® pain interference, mean (SD)</td>
<td>31.5 (8.9)</td>
<td>31.3 (8.8)</td>
<td>32.4 (10)</td>
<td>.30</td>
<td>31.1 (8.8)</td>
</tr>
<tr>
<td>Short Form Health Survey Mental Components Score, mean (SD)</td>
<td>17.8 (9.3)</td>
<td>17.5 (9.3)</td>
<td>19.7 (9.4)</td>
<td>.048</td>
<td>17.2 (9.2)</td>
</tr>
<tr>
<td>PROMIS depression, mean (SD)</td>
<td>62.3 (6.2)</td>
<td>62.2 (6.4)</td>
<td>62.6 (4.8)</td>
<td>.55</td>
<td>62.2 (6.3)</td>
</tr>
<tr>
<td>PROMIS anxiety, mean (SD)</td>
<td>65.9 (6.1)</td>
<td>65.9 (6.6)</td>
<td>65.7 (6.4)</td>
<td>.82</td>
<td>66 (6.1)</td>
</tr>
<tr>
<td>PROMIS sleep impairment, mean (SD)</td>
<td>24.1 (5.8)</td>
<td>24 (5.8)</td>
<td>25.1 (5.6)</td>
<td>.12</td>
<td>24.2 (6)</td>
</tr>
<tr>
<td>Mobile and internet use, n (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mobile phone use</td>
<td></td>
<td></td>
<td></td>
<td>.82</td>
<td>.32</td>
</tr>
<tr>
<td>Yes</td>
<td>405 (67.2)</td>
<td>349 (67)</td>
<td>56 (68.3)</td>
<td>162 (71.1)</td>
<td>48 (64.9)</td>
</tr>
<tr>
<td>No</td>
<td>198 (32.8)</td>
<td>172 (33)</td>
<td>26 (31.7)</td>
<td>66 (28.9)</td>
<td>26 (35.1)</td>
</tr>
<tr>
<td>Nonwork internet use</td>
<td></td>
<td></td>
<td></td>
<td>.13</td>
<td>.005</td>
</tr>
<tr>
<td>Never or rare</td>
<td>46 (7.6)</td>
<td>36 (6.9)</td>
<td>10 (12.2)</td>
<td>12 (5.3)</td>
<td>11 (14.9)</td>
</tr>
<tr>
<td>Occasional</td>
<td>90 (14.9)</td>
<td>75 (14.4)</td>
<td>15 (18.3)</td>
<td>30 (13.2)</td>
<td>15 (20.3)</td>
</tr>
<tr>
<td>Consistent</td>
<td>467 (77.4)</td>
<td>410 (78.7)</td>
<td>57 (69.5)</td>
<td>186 (81.6)</td>
<td>48 (64.9)</td>
</tr>
</tbody>
</table>

https://mental.jmir.org/2024/1/e52197
Noninitiation Versus Initiation of CCBT Use

When pooling CCBT users from the CCBT-only and CCBT+OHC arms, 603 participants had access to the CCBT program. During the 6-month intervention, 13.6% (82/603) did not initiate the use of the program (Table 1). Those who did not initiate use were more likely to have a lower (worse) SF-12 PCS ($P=0.01$), higher (worse) PROMIS Pain Interference measure ($P=0.048$), and less likely to have a 4-year college degree ($P=0.02$). The rate of noninitiation was 24% (22/91) for African American participants, 11.6% (58/499) for White participants, and 15% (2/13) for participants of “other” races ($P=0.006$).

Discontinue Versus Continue Use of CCBT

Of those who initiated use (n=521), 97% (504/521) completed the first module (Table 2), and 12.9% (65/504) discontinued on completing module 1. If a patient was going to complete the first module, she or he did so during the first 3 months (495/504, 98%). Only an additional 9 (1.8%) participants completed the first module in the subsequent 3 months (data not shown).

<table>
<thead>
<tr>
<th>Measure and category</th>
<th>Total (n=603)</th>
<th>Computerized cognitive behavioral therapy</th>
<th>Online health community</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>initiation (n=521)</td>
<td>noninitiation (n=82)</td>
</tr>
<tr>
<td>Work internet use</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Never or rare</td>
<td>282 (46.8)</td>
<td>239 (45.9)</td>
<td>43 (52.4)</td>
</tr>
<tr>
<td>Occasional</td>
<td>29 (4.8)</td>
<td>24 (4.6)</td>
<td>5 (6.1)</td>
</tr>
<tr>
<td>Consistent</td>
<td>292 (48.4)</td>
<td>258 (49.5)</td>
<td>34 (41.5)</td>
</tr>
</tbody>
</table>

Table 2. Baseline characteristics for sample, intervention users, and nonusers.

<table>
<thead>
<tr>
<th>Categories, measure, and category</th>
<th>Total (n=603)</th>
<th>CCBT user(^b) (n=504)</th>
<th>CCBT nonuser(^c) (n=99)</th>
<th>(P) value</th>
<th>OHC user(^d) (n=228)</th>
<th>OHC nonuser (n=74)</th>
<th>(P) value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Demographics</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age (y), mean (SD)</td>
<td>42.8 (14.2)</td>
<td>43 (14.2)</td>
<td>42 (14.4)</td>
<td>.54</td>
<td>42.2 (14)</td>
<td>43.9 (15.5)</td>
<td>.38</td>
</tr>
<tr>
<td>Sex, n (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>480 (79.6)</td>
<td>403 (80)</td>
<td>77 (77.8)</td>
<td>.62</td>
<td>182 (79.8)</td>
<td>63 (85.1)</td>
<td>.31</td>
</tr>
<tr>
<td>Male</td>
<td>123 (20.4)</td>
<td>101 (20)</td>
<td>22 (22.2)</td>
<td></td>
<td>46 (20.2)</td>
<td>11 (14.9)</td>
<td></td>
</tr>
<tr>
<td>Race, n (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>499 (82.8)</td>
<td>429 (85.1)</td>
<td>70 (70.7)</td>
<td>.001</td>
<td>188 (82.5)</td>
<td>54 (73)</td>
<td>.09</td>
</tr>
<tr>
<td>African American</td>
<td>91 (15.1)</td>
<td>64 (12.7)</td>
<td>27 (27.3)</td>
<td></td>
<td>34 (14.9)</td>
<td>19 (25.7)</td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>13 (2.2)</td>
<td>11 (2.2)</td>
<td>2 (2)</td>
<td></td>
<td>6 (2.6)</td>
<td>1 (1.4)</td>
<td></td>
</tr>
<tr>
<td>Living situation, n (%)</td>
<td></td>
<td></td>
<td></td>
<td>.048</td>
<td></td>
<td></td>
<td>.60</td>
</tr>
<tr>
<td>Alone with child</td>
<td>82 (13.6)</td>
<td>60 (11.9)</td>
<td>22 (22.2)</td>
<td></td>
<td>23 (10.1)</td>
<td>11 (14.9)</td>
<td></td>
</tr>
<tr>
<td>Alone with no child</td>
<td>114 (18.9)</td>
<td>98 (19.5)</td>
<td>16 (16.2)</td>
<td></td>
<td>45 (19.7)</td>
<td>15 (20.3)</td>
<td></td>
</tr>
<tr>
<td>Living together with child</td>
<td>106 (17.6)</td>
<td>88 (17.5)</td>
<td>18 (18.2)</td>
<td></td>
<td>45 (19.7)</td>
<td>11 (14.9)</td>
<td></td>
</tr>
<tr>
<td>Living together with no child</td>
<td>300 (49.8)</td>
<td>257 (51.1)</td>
<td>43 (43.4)</td>
<td></td>
<td>115 (50.4)</td>
<td>37 (50)</td>
<td></td>
</tr>
<tr>
<td>Working, n (%)</td>
<td></td>
<td></td>
<td></td>
<td>.65</td>
<td></td>
<td></td>
<td>.15</td>
</tr>
<tr>
<td>Employed</td>
<td>421 (69.8)</td>
<td>350 (69.4)</td>
<td>71 (71.7)</td>
<td></td>
<td>159 (69.7)</td>
<td>45 (60.8)</td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>182 (30.2)</td>
<td>154 (30.6)</td>
<td>28 (28.3)</td>
<td></td>
<td>69 (30.3)</td>
<td>29 (39.2)</td>
<td></td>
</tr>
<tr>
<td>Education, n (%)</td>
<td></td>
<td></td>
<td></td>
<td>&lt;.001</td>
<td></td>
<td></td>
<td>&lt;.001</td>
</tr>
<tr>
<td>High school or less</td>
<td>106 (17.6)</td>
<td>76 (15.1)</td>
<td>30 (30.3)</td>
<td></td>
<td>25 (11)</td>
<td>23 (31.1)</td>
<td></td>
</tr>
<tr>
<td>Attended college but did not receive a 4-year degree (also business or technical school)</td>
<td>216 (35.8)</td>
<td>177 (35.1)</td>
<td>39 (39.4)</td>
<td>.86 (37.7)</td>
<td>24 (32.4)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>College degree or higher</td>
<td>281 (46.6)</td>
<td>251 (49.8)</td>
<td>30 (30.3)</td>
<td></td>
<td>117 (51.3)</td>
<td>27 (36.5)</td>
<td></td>
</tr>
<tr>
<td><strong>Clinical characteristics, mean (SD)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Short Form Health Survey Physical Components Score</td>
<td>50.9 (12.3)</td>
<td>51.4 (12.1)</td>
<td>48.4 (12.7)</td>
<td>.03</td>
<td>51.3 (12.5)</td>
<td>50.3 (11.8)</td>
<td>.56</td>
</tr>
<tr>
<td>PROMIS(^e) pain interference</td>
<td>17.8 (9.3)</td>
<td>17.5 (9.3)</td>
<td>19.5 (9.4)</td>
<td>.049</td>
<td>17.2 (9.2)</td>
<td>18.8 (9.7)</td>
<td>.19</td>
</tr>
<tr>
<td>Short Form Health Survey Mental Components Score</td>
<td>31.5 (8.9)</td>
<td>31.3 (8.7)</td>
<td>32.2 (9.9)</td>
<td>.36</td>
<td>31.1 (8.8)</td>
<td>33.6 (10.9)</td>
<td>.04</td>
</tr>
<tr>
<td>PROMIS depression</td>
<td>62.3 (6.2)</td>
<td>62.2 (6.5)</td>
<td>62.6 (5)</td>
<td>.56</td>
<td>62.2 (6.3)</td>
<td>61.5 (6.1)</td>
<td>.41</td>
</tr>
<tr>
<td>PROMIS anxiety</td>
<td>65.9 (6.1)</td>
<td>65.8 (6.1)</td>
<td>65.9 (6.2)</td>
<td>.92</td>
<td>66 (6.1)</td>
<td>65.2 (6.2)</td>
<td>.35</td>
</tr>
<tr>
<td>PROMIS sleep impairment</td>
<td>24.1 (5.8)</td>
<td>24 (5.8)</td>
<td>24.8 (5.6)</td>
<td>.22</td>
<td>24.2 (6)</td>
<td>24.6 (5.5)</td>
<td>.68</td>
</tr>
<tr>
<td><strong>Mobile and internet use</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mobile phone use, n (%)</td>
<td></td>
<td></td>
<td></td>
<td>.56</td>
<td></td>
<td></td>
<td>.32</td>
</tr>
<tr>
<td>Yes</td>
<td>405 (67.2)</td>
<td>336 (66.7)</td>
<td>69 (69.7)</td>
<td></td>
<td>162 (71.1)</td>
<td>48 (64.9)</td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>198 (32.8)</td>
<td>168 (33.3)</td>
<td>30 (30.3)</td>
<td></td>
<td>66 (28.9)</td>
<td>26 (35.1)</td>
<td></td>
</tr>
<tr>
<td>Nonwork internet use, n (%)</td>
<td></td>
<td></td>
<td></td>
<td>.12</td>
<td></td>
<td></td>
<td>.005</td>
</tr>
<tr>
<td>Never or rare</td>
<td>46 (7.6)</td>
<td>35 (6.9)</td>
<td>11 (11.1)</td>
<td></td>
<td>12 (5.3)</td>
<td>11 (14.9)</td>
<td></td>
</tr>
<tr>
<td>Occasional</td>
<td>90 (14.9)</td>
<td>71 (14.1)</td>
<td>19 (19.2)</td>
<td></td>
<td>30 (13.2)</td>
<td>15 (20.3)</td>
<td></td>
</tr>
<tr>
<td>Consistent</td>
<td>467 (77.4)</td>
<td>398 (79)</td>
<td>69 (69.7)</td>
<td></td>
<td>186 (81.6)</td>
<td>48 (64.9)</td>
<td></td>
</tr>
</tbody>
</table>
Appendix 1). Those who logged in fewer times tended to be logged in 4 or more times (Figure S2 in Multimedia discontinued use after logging in 2 to 3 times, and 56.6% discontinued use after logging in only once, 25% (56/228) OHC during the 6-month intervention, 19% (43/228) Discontinue Versus Continue Use of the OHC PROMIS pain interference with life score (sometime after the first month of having access had a higher (54/242, 22.3%; White participants (19/53, 36%) than for White participants =.04). The rate of noninitiation was higher for African (P=.04). Patients who waited to initiate use until (P=.005), poorer SF-12 PCS (P=.04), and were less likely to use the internet outside of work (P=.003). Discontinue Versus Continue Use of the OHC Of the 75% (n=228) of patients who initiated the use of the OHC during the 6-month intervention, 19% (43/228) discontinued use after logging in only once, 25% (56/228) discontinued use after logging in 2 to 3 times, and 56.6% (129/228) logged in 4 or more times (Figure S2 in Multimedia Appendix 1). Those who logged in fewer times tended to be younger (P=.03) and African American (P=.005).

### Discussion

#### Principal Findings

In this study, the vast majority of patients initiated the use of the eHealth programs. Over half of the participants discontinued using CCBT before completion, and 43% (99/228) logged into the OHC 3 or fewer times. These latter groups provide an opportunity to explore areas where improvements in engagement methods and eHealth interventions could be made.

#### Noninitiation of CCBT and OHC

Patients who did not initiate the use of the CCBT or OHC programs were more likely to have less formal education, and those who did not initiate the use of the OHC also tended to be less frequent internet users. These findings are consistent with the conclusion that those who have less experience with technology may be less savvy or confident with technology, and that there is more reluctance to initiate use of web-based programs. Reluctance due to one’s level of digital or health literacy is addressable at the start by helping to develop greater confidence in the technologies involved in the interventions and the ability of the contents to be effective. In this study, all participants needed to have access to the necessary technologies for participation; however, the care managers also offered telephone assistance with technical difficulties with the e-mental health programs. Some may have felt stigmatized asking for help, particularly with their personal technologies; felt that phone guidance would not work; or felt that this was not an offer to assist with using the programs or their personal technologies but for technical issues outside of their control.

Poor physical well-being, as indicated by lower SF-12 PCS scores or higher pain interference with life, also reduced the likelihood of patients initiating use. This raises several possibilities. First, it may be that these patients had more difficulty using the technologies and websites due to physical limitations and associated pain or they may have had a reduced ability to concentrate on the contents. Another possibility is that they were less enthusiastic about the interventions because they were not directly focused on physical health issues; thus, they were perceived as not addressing their high-priority need. However, it has been well established that chronic pain, depression, anxiety, and somatic amplification co-occur [40]. Such issues would be important to address during early engagement activities designed to both develop confidence in

---

### Table: Work internet use, n (%) & Other measures of interest, n (%) 

<table>
<thead>
<tr>
<th>Categories, measure, and category</th>
<th>Total (n=603)</th>
<th>CCBT user (n=504)</th>
<th>CCBT nonuser (n=99)</th>
<th>P value</th>
<th>OHC user (n=228)</th>
<th>OHC nonuser (n=74)</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Work internet use, n (%)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Never or rare</td>
<td>282 (46.8)</td>
<td>231 (45.8)</td>
<td>51 (51.5)</td>
<td>.22</td>
<td>102 (44.7)</td>
<td>42 (56.8)</td>
<td>.19</td>
</tr>
<tr>
<td>Occasional</td>
<td>29 (4.8)</td>
<td>22 (4.4)</td>
<td>7 (7.1)</td>
<td></td>
<td>10 (4.4)</td>
<td>3 (4.1)</td>
<td></td>
</tr>
<tr>
<td>Consistent</td>
<td>292 (48.4)</td>
<td>251 (49.8)</td>
<td>41 (41.4)</td>
<td></td>
<td>116 (50.9)</td>
<td>29 (39.2)</td>
<td></td>
</tr>
</tbody>
</table>

**Notes:**

aCCBT: computerized cognitive behavioral therapy.
bCompleted the first CCBT module.
cDid not complete the first CCBT module.
dOHC: online health community.
ePROMIS: Patient-Reported Outcomes Measurement Information System.
the ability of the interventions to help and to assess and address reasons for patients’ hesitancy to use the interventions.

Those who did not initiate the use of the interventions during the first month of the trial were unlikely to later. If the methods designed to develop engagement with the interventions did not sufficiently influence a patient during the initial weeks, they were unlikely to have a subsequent effect. This may help identify an area where certain patients could benefit from earlier or additional methods to increase their engagement.

**Discontinuation of CCBT**

Patients who did not complete the first CCBT module were likely to be less technology savvy, have poorer physical functioning, have higher pain interference with life, and have less severe anxiety symptoms at baseline. Mental health burden, comfort with technology, and physical functioning were recurring influences on patient initiation and discontinuation. Others have also found that an increased mental health burden is associated with greater use of e-mental health interventions for anxiety and depressive symptoms [41]. A counterintuitive finding was that those less likely to be mobile phone users were more likely to complete a higher number of CCBT modules. This could be because they had less access to web-based resources, which made the offered treatments a more unique and valued opportunity to receive mental health services.

**Discontinuation of CCBT and OHC**

Given that over half of the participants discontinued the CCBT program before completion, this may indicate the need for additional ongoing methods to maintain users’ buy-in. It should be noted that CCBT did have a positive effect on clinical outcomes compared with UC [20], indicating that to receive benefit, at least partial benefit, it was not necessary to complete the program. Some patients may have felt that the program was no longer addressing their needs or that they improved sufficiently and discontinued [42]. This raises the issue of whether setting patients’ treatment expectations, for example, for when full benefit has been gained, would influence discontinuation and improve outcomes for those who leave before completion. Examining patients’ reasons for discontinuing would provide patient-centered insights on this issue and potentially how the interventions could be tailored as needs change during use.

African American participants who initiated the use were more likely than White participants to discontinue OHC use after only one login and more likely to discontinue CCBT use before completing the program. Taken together, these findings indicate that more could be done to facilitate the engagement of African Americans, and possibly other minority groups. This may also indicate that content adaptations could improve the intervention’s fit for diverse groups [38]. Digital literacy and internet connectivity, which include skills, confidence, connectivity and its ongoing affordability, level of access, devices, training, and technical support, have been called the “super social determinants of health” because they influence all other social determinants of health [43]. The study examined only a few social determinants of health, for example, employment and education. It is likely that other social determinants of health influenced these findings, and this is a key area for future investigations.

**The Importance of Strategies to Facilitate Engagement**

Only a minority of patients did not initiate the use of e-mental eHealth interventions, and almost half of them completed CCBT. This argues for the effectiveness of the engagement offered by the ongoing collaborative care strategies used. This is supported by the findings from 2 earlier studies on CCBT. The original study that established the efficacy of CCBT provided in-person treatment at a research office [21]. One module was completed during each visit. Each user was provided with 1:1 supervision from a practice nurse. The nurse ensured that each participant interacted successfully with the computer and treatment program. A subsequent study examined the web-based home delivery of CCBT. Investigators found no effect of CCBT when compared with UC [44]. In this study, although participants were called weekly at their homes by a research staff member who encouraged continued program use, the amount of contact over the 4-month study period was an average of only 23 seconds per week.

These findings support the potential of providing eHealth programs via collaborative care and supported and “guided” models [45]. They also point to the potential of including additional and improved methods to facilitate engagement. Highly effective engagement practices, research findings, and theoretical models have identified several best practices to engage patients with interventions. These identify at least 5 characteristics of an intervention model that facilitate strong engagement [14,17,46-49]. One is compatibility. Engagement activities should establish how the intervention will be personally meaningful and able to meet the needs a user finds important. For example, it will facilitate better illness management or an improved ability to socialize or participate in work. In this study, it may have been helpful if a patient’s PCP or care manager had introduced the programs and engaged in shared decision-making to identify specific benefits to the patient. In addition, participants may have considered the interventions more of a study than personalized treatment because they were not presented by their PCP. Another characteristic is relative advantage. This is the level of belief that the effort and resources involved will provide sufficient added value over alternatives. In total, 2 alternatives that patients commonly consider are not initiating the treatment and/or discontinuing. The study’s CCBT program may have been too rigid for some patients, as they could not skip a module or tailor the program to their immediate needs. This feedback was provided to the care managers. Several existing and recent eHealth programs have developed more flexible approaches that allow users to move through the program and tailor the presentation according to their current needs and previous experience with mental health treatment [50-52]. However, many eHealth interventions with fixed presentations are highly effective [53]. Creating approaches that can accommodate diverse and even changing needs of patients may help support broader engagement. A third characteristic is the complexity of the intervention. This is the extent to which the intervention and required technologies are intuitive to use either before or after training. This can be especially important for users without
sufficient digital or health literacy, and in-person training may be beneficial. Fourth characteristic is observability. This is the extent to which it will be easy for users to see improvements in issues that are important to them because of their use of the intervention. This may have manifested in 2 opposite ways. Some may have perceived initial improvement and thought that it was all that could be expected. Others may have felt hampered by the rigidity of the CCBT program because patients may have had to complete modules that did not seem relevant to them; thus, they did not see added benefit in continuing the program. Both these factors could have led to discontinuation. The fifth characteristic is supportive accountability. This includes support from a human coach or care manager who can provide guidance and develop accountability on the part of users. This can increase adherence. In this study, the personalized and continuous support provided via the collaborative care strategy, and knowing that care managers would call to check on progress and issues around completing homework, likely contributed to the high initiation and ongoing use rates. The ability of such “guidance” to improve initiation and adherence by some users has been documented [45,51]. This argues for the considerable advantage of guided collaborative care models such as was used in this study.

Conclusions
These analyses produced several findings: those with greater mental health needs had a greater predisposition to initiate and continue use, those with poorer physical well-being were less likely to initiate use, and comfort with technologies seemed to influence patients’ likelihood of initiating and continuing use. These findings support the conclusion that methods are needed to build engagement with eHealth interventions that can be tailored to these as well as other specific needs of patients. Although discontinuation has been explored to a somewhat greater extent in the eHealth literature, less attention has been devoted to noninitiation, and neither has been thoroughly investigated. The findings highlight a potentially important need for additional studies of both noninitiation and discontinuation and the relevance of these 2 phenomena to the types of personalized engagement and ongoing support methods that could benefit users and help address the gaps that lead to Eisenbach’s [17] Law of Attrition.

Acknowledgments
This trial was supported by grant R01 MH093501 from the National Institute of Mental Health.

Conflicts of Interest
None declared.

Multimedia Appendix 1
Number of computerized cognitive behavioral therapy program modules completed and number of logins to the online health community platform.

References


Abbreviations

CCBT: computerized cognitive behavioral therapy
ISG: internet support group
MCS: Mental Components Score
OHC: online health community
PCP: primary care physician
PCS: Physical Components Score
PHQ-9: 9-item Patient Health Questionnaire
PROMIS: Patient-Reported Outcomes Measurement Information System
RCT: randomized clinical trial
SF-12: Short Form Health Survey
UC: usual care

©Armando J Rotondi, Bea Herbeck Belnap, Scott Rothenberger, Robert Feldman, Barbara Hanusa, Bruce L Rollman. Originally published in JMIR Mental Health (https://mental.jmir.org), 17.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Mental Health, is properly cited. The complete bibliographic information, a link to the original publication on https://mental.jmir.org/, as well as this copyright and license information must be included.
Developing and Implementing a Web-Based Branching Logic Survey to Support Psychiatric Crisis Evaluations of Individuals With Developmental Disabilities: Qualitative Study and Evaluation of Validity

Deborah A Bilder1, MD; Mariah Mthembu2, MD; Whitney Worsham2, MD; Patricia Aguayo1, MPH, MD; Jacob R Knight2, BS; Steven W Deng2, HBS; Tejinder P Singh2, BDS, MDS, MPH; John Davis3, PhD

1University of Utah Huntsman Mental Health Institute, Salt Lake City, UT, United States
2University of Utah School of Medicine, Salt Lake City, UT, United States
3University of Utah, Salt Lake City, UT, United States

Corresponding Author:
Deborah A Bilder, MD
University of Utah Huntsman Mental Health Institute
383 Colorow Drive, Room 360
Salt Lake City, UT, 84108
United States
Phone: 1 801 410 0852
Email: deborah.bilder@hsc.utah.edu

Abstract

Background: Individuals with developmental disabilities (DD) experience increased rates of emotional and behavioral crises that necessitate assessment and intervention. Psychiatric disorders can contribute to crises; however, screening measures developed for the general population are inadequate for those with DD. Medical conditions can exacerbate crises and merit evaluation. Screening tools using checklist formats, even when designed for DD, are too limited in depth and scope for crisis assessments. The Sources of Distress survey implements a web-based branching logic format to screen for common psychiatric and medical conditions experienced by individuals with DD by querying caregiver knowledge and observations.

Objective: This paper aims to (1) describe the initial survey development, (2) report on focus group and expert review processes and findings, and (3) present results from the survey’s clinical implementation and evaluation of validity.

Methods: Sources of Distress was reviewed by focus groups and clinical experts; this feedback informed survey revisions. The survey was subsequently implemented in clinical settings to augment providers’ psychiatric and medical history taking. Informal and formal consults followed the completion of Sources of Distress for a subset of individuals. A records review was performed to identify working diagnoses established during these consults.

Results: Focus group members (n=17) expressed positive feedback overall about the survey’s content and provided specific recommendations to add categories and items. The survey was completed for 231 individuals with DD in the clinical setting (n=161, 69.7% men and boys; mean age 17.7, SD 10.3; range 2-65 years). Consults were performed for 149 individuals (n=102, 68.5% men and boys; mean age 18.9, SD 10.9 years), generating working diagnoses to compare survey screening results. Sources of Distress accuracy rates were 91% (95% CI 85%-95%) for posttraumatic stress disorder, 87% (95% CI 81%-92%) for anxiety, 87% (95% CI 81%-92%) for episodic expansive mood and bipolar disorder, 82% (95% CI 75%-87%) for psychotic disorder, 79% (95% CI 71%-85%) for unipolar depression, and 76% (95% CI 69%-82%) for attention-deficit/hyperactivity disorder. While no specific survey items or screening algorithm existed for unspecified mood disorder and disruptive mood dysregulation disorder, these conditions were caregiver-reported and working diagnoses for 11.7% (27/231) and 16.8% (25/149) of individuals, respectively.

Conclusions: Caregivers described Sources of Distress as an acceptable tool for sharing their knowledge and insights about individuals with DD who present in crisis. As a screening tool, this survey demonstrates good accuracy. However, better differentiation among mood disorders is needed, including the addition of items and screening algorithm for unspecified mood disorder and disruptive mood dysregulation disorder. Additional validation efforts are necessary to include a more geographically
diverse population and reevaluate mood disorder differentiation. Future study is merited to investigate the survey’s impact on the psychiatric and medical management of distress in individuals with DD.

(Referred to: doi:10.2196/50907)

KEYWORDS

developmental disabilities; disruptive behavior; psychiatric comorbidity; web-based; psychiatric crisis; disability; mental health; behavioral crises; intervention; general population; screening; accuracy; mood disorder; sources of distress; autism; intellectual disability

Introduction

Background

Individuals with developmental disabilities (DD) such as autism and intellectual disability (ID) experience mental health crises more frequently than the general population [1,2]. A broad range of psychiatric and medical conditions can contribute to the agitation, aggression, and self-injury that often characterize these crises [3-10]. Rates of anxiety (20%-77%), depression (10%-20%), expansive mood and bipolar disorder (5%-11%), and psychosis (5%-10%) among individuals with autism exceed those in neurotypical individuals [11-19]. Elevated rates of psychiatric disorders have also been identified in individuals with ID, notably for unspecified psychosis (4.8%), schizophrenia (3.9%), and bipolar disorder (8%) [20-22]. A history of trauma or abuse should also be considered in individuals with DD presenting in crisis [23].

When psychiatric and medical conditions are recognized as factors contributing to a person’s mental health crisis, clear long-term treatment targets emerge. Nevertheless, for those with DD, co-occurring medical and psychiatric conditions are often unrecognized, leaving them vulnerable to experiencing diagnostic overshadowing. Diagnostic overshadowing occurs when disruptive behaviors in individuals with DD are attributed to their disability without consideration of other potential medical or psychiatric conditions that could contribute to their behavioral presentation [24].

Self-, parent-, and caregiver-report mental health questionnaires provide an efficient means of screening for common psychiatric conditions in the neurotypical population. However, for those with DD, self-report questionnaires may be impeded by communication deficits or a limited capacity to reflect on internal experiences. Parent- and caregiver-report questionnaires normed in typically developing children may also provide inadequate mental health screening for those with ID because they often include items that are inapplicable to children with minimal language ability, exclude severe conditions that disproportionately affect children with DD (e.g., mania and psychosis), and overlook the individualized manner in which psychiatric symptoms manifest in this population [20,25-27].

The American Psychiatric Association and the National Association for the Dually Diagnosed published the Diagnostic Manual–Intellectual Disability in 2007, and subsequently, in 2016, the second edition (Diagnostic Manual–Intellectual Disability–Second Edition; DM-ID-2) [28,29]. These texts adapt the Diagnostic and Statistical Manual of Mental Disorders criteria to reflect their presentation in individuals with ID. The Psychopathology Instrument for Mentally Retarded Adults and the Psychiatric Assessment Schedule for Adults with Developmental Disabilities (PAS-ADD) operationalize adapted diagnostic criteria into structured interviews to provide a framework through which to identify psychiatric conditions in this population [30,31]. These interviews are quite lengthy and require training to administer. Even as an abbreviated semistructured interview, the Mini PAS-ADD Clinical Interview takes approximately 45 minutes to complete [32]. Existing parent- and caregiver-report psychiatric screening tools for individuals with ID create a more efficient and practical means of collecting information [33-36]; yet, the checklist format of parent- and caregiver-report questionnaires limits depth and scope, both of which are necessary when evaluating crises in a population with complex medical and mental health needs. In addition, there is a great need for the inclusion of items that query symptoms of common medical conditions (e.g., epilepsy, gastrointestinal disorders, and poor dentition) that manifest with agitation and aggression and occur more frequently in individuals with DD [3,37,38].

Sources of Distress is a survey developed for parents and caregivers (hereinafter collectively referred to as caregivers) that uses a web-based branching logic format to screen for mental health and medical conditions among individuals with DD who present in crisis. This tool informs the care of individuals experiencing distress and is intended for use when the severity or persistence of disruptive behavior prompts the consideration of medication intervention. Screening information endorsed by caregivers is organized into relevant psychiatric and medical categories within a report. This report (Multimedia Appendix 1 [39]) is developed for the caregiver and can subsequently facilitate their shared decision-making process with health care providers as specific underlying conditions are evaluated. Sources of Distress aims to minimize diagnostic overshadowing and optimize the ability of the caregiver and the provider to recognize the presence of psychiatric and medical conditions that merit targeted intervention. The web-based branching logic format is adaptive in nature—optimizing caregiver and health care provider convenience and efficiency and minimizing caregiver burden for survey completion [40].

Objectives

This paper aims to (1) describe the initial development of Sources of Distress; (2) report on the findings from focus group evaluations and expert reviews and indicate how this feedback shaped the subsequent version of the survey; and (3) present the results from the evaluation of validity for Sources of Distress after its implementation in the clinical setting. The Methods and Results sections are divided into 3 subsections (apart from the
Ethical Considerations section in Methods) corresponding to the development, initial evaluation, and clinical implementation phases of Sources of Distress.

Methods

Ethical Considerations

The University of Utah Institutional Review Board approved focus group activities for Sources of Distress content validation (IRB_00111975). Focus group participants provided informed consent and received compensation for their time in the form of an Amazon gift card worth US $50. The University of Utah Institutional Review Board approved with a waiver of consent for the retrospective records review, data collection, and subsequent deidentified data analysis for individuals for whom Sources of Distress was completed as part of their clinical care (IRB_00170868).

Early Survey Development

Funding for the development of Sources of Distress was provided by the Autism Council of Utah based in Murray, Utah, United States [41]. The development team comprised a triple board physician (pediatrics, general psychiatry, and child and adolescent psychiatry), an educational psychologist, a medical student, and a business consultant grandparent of a child with autism and ID. In the initial development phase, Sources of Distress was built in Qualtrics (Qualtrics International Inc) using a branching logic format to approximate the history-taking component of a DD psychiatric evaluation. This evaluation queries psychiatric symptom clusters, physical complaints, and psychiatric medical history to support the development of a diagnostic impression for which treatment recommendations could be made.

Multiple expert opinion sources were reviewed to identify pertinent screening categories and corresponding items to include in Sources of Distress. The expert sources included published literature, the DM-ID-2, the Mini PAS-ADD Clinical Interview, and the screening interview for the Kiddie Schedule for Affective Disorders and Schizophrenia–Present and Lifetime (a semistructured psychiatric diagnostic interview for children and adolescents) [28,32,42]. As Sources of Distress is intended for use in the context of distress, the presence of at least 1 manifestation of a behavioral or emotional crisis must be endorsed to initiate survey questions.

Initial Survey Evaluation

Focus Group Evaluation

In 2018 and early 2019, focus group participants were recruited from (1) a university-based outpatient program that provides medical and psychiatric care for individuals with DD across the lifespan and (2) the Autism Council of Utah (a community stakeholder organization for individuals and families affected by autism). Six focus groups were conducted that consisted of parents (6/17, 35%), professional caregivers (6/17, 35%), and adults with both DD and the ability to provide verbal feedback (5/17, 29%). Participants completed Sources of Distress before attending the focus group and reported on specific items, missing items, item wording, and attribution of items to corresponding conditions. Interviews and discussions were transcribed and analyzed following the framework analysis of Ritchie and Spencer [43]. Inductive reasoning and the constant comparative method put forth by Strauss and Corbin [44] were used to compare statements by parents, professional caregivers, and individuals with disability within and across focus groups.

Expert Review Evaluation

Revisions were made to Sources of Distress based on focus group feedback. Experts reviewed the revised survey version, and additional changes were made. The experts included a pediatrician and 2 child psychiatrists, all with national recognition for their clinical and research work in DD.

Clinical Implementation

Overview

Sources of Distress was implemented in various clinical settings to augment the clinical history-taking process—outpatient (primary care, neurology, developmental pediatrics, and psychiatry), emergency department, psychiatric inpatient, and residential care. Caregivers were given a link to the survey when their health care provider identified the need for expert support in managing severe agitation and aggression. All caregivers (231/231, 100%) completed the survey outside of the clinical setting. An informal or formal consult followed survey completion for a subset of individuals. In August 2020, the survey was transitioned from the Qualtrics platform to the REDCap (Research Electronic Data Capture; Vanderbilt University) platform to automate the Sources of Distress report generation using the custom template engine [45]. This external REDCap module was developed and has been maintained by the Integrated Research Informatics Services of British Columbia Children’s Hospital Research Institute [46].

Survey Data Collection

Sources of Distress responses were collected from its first use in a clinical setting from February 2019 through June 2022. The following information was obtained: respondent type, individual characteristics, caregiver-reported diagnoses, current medications, distress manifestations, psychiatric symptoms, and medical symptoms, conditions, or concerns. When multiple caregivers reported on the same individual, responses were used from the caregiver closest to where the individual lived (eg, parent for a child living at home and professional caregiver for an individual living in a residential setting). Psychotropic medications were organized within the following mutually exclusive categories: antipsychotics, antidepressants, non-antidepressant anxiolytics, anticonvulsants, lithium, alpha-2 agonists, stimulants, and atomoxetine.

Consults

A medical decision-making support consultation took place after survey completion as either an informal or a formal consult for a subset of individuals. This consult was conducted by a clinical team led by the triple board physician member of the survey’s development team. The consult team used DM-ID-2 criteria as the basis for establishing psychiatric diagnoses. At a minimum (as an informal consult), the consult involved a

https://mental.jmir.org/2024/1/e50907 JMIR Ment Health 2024 | vol. 11 | e50907 | p.339 (page number not for citation purposes)
discussion between a DD clinical expert and the referring provider. This discussion resulted in a collective determination of working diagnoses and treatment plan. A formal consult included the additional components of medical records review, caregiver interview, and direct participant evaluation. Psychiatric diagnoses that were not reported in the survey but discussed by the provider or documented in the medical record were included among preexisting diagnoses.

Working diagnoses were abstracted from formal and informal consult documentation and served as the standard to define true case status.

**Mood Disorder Classification**

The presence of a mood disorder among preexisting and working diagnoses was classified into mutually exclusive categories such that there was no overlap among individuals across mood disorder categories to allow for direct comparisons across preexisting diagnoses, survey screening status results, and working diagnoses. The following mood disorder classification hierarchy was used from highest to lowest: (1) episodic expansive mood, hypomania, mania, and bipolar disorder, hereafter collectively referred to as bipolar disorder, (2) disruptive mood dysregulation disorder (DMDD) and unspecified mood disorder, and (3) unipolar depression. If an individual had a diagnosis of bipolar disorder, regardless of what other mood disorder diagnosis was reported or identified, their mood disorder classification would be bipolar disorder. An individual was only classified with unipolar depression if (1) they had a depression diagnosis and (2) they had no other mood disorder diagnosis.

**Statistical Analyses and Evaluation of Validity**

Descriptive statistics and chi-square tests were conducted in SPSS (version 28.0; IBM Corp) with an α of .05 selected to assess statistical significance. Differences between surveys with an accompanying consult and those without were measured. Positive predictive value (PPV), negative predictive value (NPV), and accuracy rates were calculated for (1) preexisting diagnoses and (2) survey screening results with working diagnoses used as the determinant of true case status. We calculated 95% CIs for the binomial distribution of accuracy rates.

**Results**

**Early Survey Development**

Table 1 lists the modules and corresponding items initially selected as the categories, characteristics, and symptoms to be queried by Sources of Distress. The initial version of the survey included scoring algorithms to determine positive screen status for the following conditions: anxiety, unipolar depression, bipolar disorder, psychosis, and attention-deficit/hyperactivity disorder (ADHD).
<table>
<thead>
<tr>
<th>Module</th>
<th>Original items</th>
<th>Added in response to feedback</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction and demographics</td>
<td>• Respondent’s relationship to the individual who is affected</td>
<td>• For professional caregivers: how long have you known the affected individual?</td>
</tr>
<tr>
<td></td>
<td>• Distress symptoms</td>
<td>• Added “increased fixation on certain things” and “changes in behavior such as increased isolation, social withdrawal” to distress symptoms</td>
</tr>
<tr>
<td></td>
<td>• Language ability</td>
<td>• Is there a difference in language ability at the physician’s office? If so, is there something the provider can do to improve the individual’s ability to speak for themselves?</td>
</tr>
<tr>
<td></td>
<td>• Age</td>
<td>• Added “increased fixation on certain things” and “changes in behavior such as increased isolation, social withdrawal” to distress symptoms</td>
</tr>
<tr>
<td></td>
<td>• Known diagnoses</td>
<td>• Is there a difference in language ability at the physician’s office? If so, is there something the provider can do to improve the individual’s ability to speak for themselves?</td>
</tr>
<tr>
<td></td>
<td>• Current medications</td>
<td>• Added “increased fixation on certain things” and “changes in behavior such as increased isolation, social withdrawal” to distress symptoms</td>
</tr>
<tr>
<td>Behavior patterns and triggers</td>
<td>• Circumstances of disruptive behavior (recognized triggers, patterns, motivation and reinforcement, and location)</td>
<td>• Query perceived function to behavior surrounding distress</td>
</tr>
<tr>
<td>Sleep</td>
<td>• Time of sleep onset and awakening</td>
<td>• Food seeking as an activity interfering with sleep</td>
</tr>
<tr>
<td></td>
<td>• Middle-of-the-night interruptions</td>
<td>• Sleep apnea diagnosis and symptoms</td>
</tr>
<tr>
<td></td>
<td>• Naps</td>
<td>• Discomfort precipitating sleep disturbance</td>
</tr>
<tr>
<td>Anxiety</td>
<td>• Leading to significant outbursts or discomfort: transitioning activities, getting stuck on certain topics or things, and minor changes in daily activities</td>
<td>• Panic and nightmares</td>
</tr>
<tr>
<td></td>
<td>• Less energy than usual, increased crying spells, sadness, irritability, isolative, loss of interest in activities typically enjoyed, and excess sleep</td>
<td>• Sensory sensitivity that leads to discomfort</td>
</tr>
<tr>
<td></td>
<td>• Injures self on purpose; if yes: location of injury and whether self-injury is causing discomfort?</td>
<td>• Repeated checking or rituals, which interferes with daily activities</td>
</tr>
<tr>
<td>Mania</td>
<td>• Establish baseline energy</td>
<td>• Whether self-injury is concerning to parent or caregiver</td>
</tr>
<tr>
<td></td>
<td>• Query discrete periods out of the blue lasting ≥ 2 days of increased energy compared to baseline, laughing or vocalizing for no clear reason, particularly happy or giddy, risk taking, sexually acting out, increased impulsivity, and decreased need for sleep</td>
<td>• Whether self-injury could be perpetuated by attention seeking or avoidance</td>
</tr>
<tr>
<td>Psychosis</td>
<td>• Appearing to be responding to internal auditory or visual stimuli</td>
<td>• No changes made</td>
</tr>
<tr>
<td></td>
<td>• Yelling angrily in a room where no one else is present as if yelling at someone who is not there</td>
<td>• No changes made</td>
</tr>
<tr>
<td>ADHD&lt;sup&gt;a&lt;/sup&gt;</td>
<td>• Difficulty following through on instructions, avoiding task demands, easily distractible, fidgety or restless, high activity when expected to remain in 1 place, constantly moving, blurring into other people’s conversations, and demanding attention or desired items</td>
<td>• Excessive talking</td>
</tr>
<tr>
<td>General medical problems</td>
<td>• Query history of headaches, seizures, injuries that can be causing discomfort, thyroid abnormalities, and tooth pain</td>
<td>• Are there unusual ways of responding to physical discomfort?</td>
</tr>
<tr>
<td></td>
<td>• Could any of these issues be contributing to distress?</td>
<td>• Added joint pain; ear, nose, or throat pain; and seasonal allergies</td>
</tr>
<tr>
<td>Trauma&lt;sup&gt;b&lt;/sup&gt;</td>
<td>• N/A&lt;sup&gt;c&lt;/sup&gt;</td>
<td>• History of trauma</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Related to trauma: avoidance, flashbacks, and nightmares</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Hypervigilance</td>
</tr>
<tr>
<td>Gastrointestinal concerns&lt;sup&gt;b&lt;/sup&gt;</td>
<td>• N/A</td>
<td>• Bowel movement frequency</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Query history of constipation, stool accidents, frequent stomachaches, food allergies, and acid reflux.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Subsequent additions: changes in appetite, nausea, and variable bowel movements</td>
</tr>
</tbody>
</table>
Added in response to feedback

Module | Original items | Added in response to feedback
---|---|---
Menstrual concerns \( ^b \) (for female patients only) | N/A | - Query presence of mood changes during menses, endometriosis, polycystic ovary syndrome, significant menstrual pain, excess bleeding during or between cycles, and anxiety surrounding periods. Could any of these conditions be leading to distress?
- Birth control: oral contraceptives, hormonal IUD\( ^d \), nonhormonal IUD, and Depo-Provera (a contraceptive injection).
- When was the last dental visit?
- Query presence of changes in eating patterns: texture preference, sensitivity to hot or cold food or drink preference for eating on 1 side of the mouth, and reduced oral intake
- Grinding teeth

Dental concerns \( ^b \) | N/A |  

\( ^a \) ADHD: attention-deficit/hyperactivity disorder.
\( ^b \) Module added in response to focus group feedback.
\( ^c \) N/A: not applicable.
\( ^d \) IUD: intrauterine device.

**Initial Survey Evaluation**

**Focus Group Feedback**

During the focus groups, 3 main themes emerged in this analysis.

- Theme A: respondents gave overall positive feedback regarding existing content and specific feedback regarding areas where there was room to expand content. Table 1 describes the modules and items added in response to this feedback. Notably, a posttraumatic stress disorder (PTSD) module was added along with a PTSD scoring algorithm to determine positive screen status.
- Theme B: most of the respondents (15/17, 88%) agreed that the symptoms queried matched their understanding of the psychiatric and medical conditions to which they are attributed.
- Theme C: all participant groups reported positive acceptability of the branching logic format and time required to complete the measure.

**Expert Review**

Overall, the expert review supported the Sources of Distress categories and respective items attributed to each condition.

One expert recommended adding items that query gender and replacing sex as the basis for pronoun selection within the tool and its report. This expert also suggested that the report include screening results for each psychiatric condition. The former recommendations were implemented when Sources of Distress was transitioned to the REDCap platform. The latter recommendation was deferred until after screening algorithms are validated in a clinical setting.

**Clinical Implementation**

**Sample Characteristics**

Surveys (N=264) were completed by parents or guardians (n=200, 75.8%), professional caregivers (n=43, 16.3%), and other caregivers (n=21, 8%) of 231 individuals (n=161, 69.7% men and boys; n=69, 29.9% women and girls; and n=1, 0.4% other; mean age 17.7, SD 10.3; range 2-65 years). Informal (n=62, 41.6%) and formal (n=87, 58.4%) consults were performed for 149 individuals collectively. Table 2 presents sample characteristics, the manifestations of distress, and a comparison between individuals with a consult and those without.
### Table 2. Sample characteristics and distress manifestations.

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>With consult&lt;sup&gt;a&lt;/sup&gt; (n=149), n (%)</th>
<th>Without consult (n=82), n (%)</th>
<th>Total (N=231), n (%)</th>
<th>Chi-square (df)</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender&lt;sup&gt;b&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Man or boy</td>
<td>102 (68.5)</td>
<td>59 (72)</td>
<td>161 (69.7)</td>
<td>2.3 (2)</td>
<td>.34</td>
</tr>
<tr>
<td>Woman or girl</td>
<td>47 (31.5)</td>
<td>22 (25.5)</td>
<td>69 (29.9)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other&lt;sup&gt;b&lt;/sup&gt;</td>
<td>0 (0)</td>
<td>1 (1.2)</td>
<td>1 (0.4)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Caregiver&lt;sup&gt;c&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parent or guardian</td>
<td>108 (72.5)</td>
<td>71 (86.6)</td>
<td>179 (77.5)</td>
<td>7.8 (2)</td>
<td>.02</td>
</tr>
<tr>
<td>Professional caregiver</td>
<td>32 (21.5)</td>
<td>6 (7.3)</td>
<td>38 (16.5)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>9 (6)</td>
<td>5 (6.1)</td>
<td>14 (6.1)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age range (y)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;13</td>
<td>46 (30.9)</td>
<td>38 (46.3)</td>
<td>84 (36.4)</td>
<td>7.4 (2)</td>
<td>.03</td>
</tr>
<tr>
<td>13-22</td>
<td>57 (38.3)</td>
<td>30 (36.6)</td>
<td>87 (37.7)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&gt;22</td>
<td>46 (30.9)</td>
<td>14 (17.1)</td>
<td>60 (26)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Language ability</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Full verbal ability</td>
<td>76 (51)</td>
<td>42 (51.2)</td>
<td>118 (51.1)</td>
<td>0.0 (2)</td>
<td>.99</td>
</tr>
<tr>
<td>Limited use of words</td>
<td>46 (30.9)</td>
<td>25 (30.5)</td>
<td>71 (30.7)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nonverbal</td>
<td>27 (18.1)</td>
<td>15 (18.3)</td>
<td>42 (18.2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Manifestation of distress</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Agitation</td>
<td>130 (87.2)</td>
<td>70 (85.4)</td>
<td>200 (86.6)</td>
<td>0.2 (1)</td>
<td>.69</td>
</tr>
<tr>
<td>Aggression</td>
<td>97 (65.1)</td>
<td>50 (61)</td>
<td>147 (63.6)</td>
<td>0.4 (1)</td>
<td>.53</td>
</tr>
<tr>
<td>Change in sleep</td>
<td>86 (57.7)</td>
<td>38 (46.3)</td>
<td>124 (53.7)</td>
<td>2.8 (1)</td>
<td>.10</td>
</tr>
<tr>
<td>Moodiness</td>
<td>122 (81.9)</td>
<td>65 (79.3)</td>
<td>187 (81)</td>
<td>0.2 (1)</td>
<td>.63</td>
</tr>
<tr>
<td>Increased fixation</td>
<td>115 (77.2)</td>
<td>57 (69.5)</td>
<td>172 (74.5)</td>
<td>1.6 (1)</td>
<td>.20</td>
</tr>
<tr>
<td>Change in eating patterns</td>
<td>52 (34.9)</td>
<td>19 (23.2)</td>
<td>71 (30.7)</td>
<td>3.4 (1)</td>
<td>.07</td>
</tr>
<tr>
<td>Change in personality</td>
<td>99 (66.4)</td>
<td>59 (72.0)</td>
<td>158 (68.4)</td>
<td>0.7 (1)</td>
<td>.39</td>
</tr>
<tr>
<td>Change in behavior</td>
<td>96 (64.4)</td>
<td>45 (54.9)</td>
<td>141 (61)</td>
<td>2.0 (1)</td>
<td>.15</td>
</tr>
<tr>
<td>Self-injurious behavior</td>
<td>73 (49)</td>
<td>39 (47.6)</td>
<td>112 (48.5)</td>
<td>0.0 (1)</td>
<td>.84</td>
</tr>
<tr>
<td>Type of disability</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Autism without ID&lt;sup&gt;d&lt;/sup&gt;</td>
<td>52 (34.9)</td>
<td>47 (57.3)</td>
<td>99 (42.9)</td>
<td>10.9 (1)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>ID without autism</td>
<td>15 (10.1)</td>
<td>7 (8.5)</td>
<td>22 (9.5)</td>
<td>0.1 (1)</td>
<td>.71</td>
</tr>
<tr>
<td>ID and autism</td>
<td>74 (49.7)</td>
<td>18 (22.0)</td>
<td>92 (39.8)</td>
<td>17.0 (1)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Genetic syndrome&lt;sup&gt;e&lt;/sup&gt;</td>
<td>17 (11)</td>
<td>17 (20.7)</td>
<td>34 (14.7)</td>
<td>3.7 (1)</td>
<td>.06</td>
</tr>
</tbody>
</table>

<sup>a</sup>Includes informal and formal consults.

<sup>b</sup>One participant reported other as gender: no participants reported non-binary as gender.

<sup>c</sup>When multiple caregivers completed Sources of Distress, the report from the caregiver with whom the participant spends the most time was used in this table.

<sup>d</sup>ID: intellectual disability.

<sup>e</sup>Genetic syndrome includes some individuals who also populate the autism or ID categories.

### Preexisting Psychiatric Diagnoses

The presence of at least 1 preexisting psychiatric diagnosis was reported in 65.4% (151/231) of the individuals. Individuals who received a consult compared to those without a consult were more likely to have a caregiver-reported history of psychotic disorder (14/149, 9.4% vs 1/82, 1%; P=.02; Table 3).

---

https://mental.jmir.org/2024/1/e50907 JMIR Ment Health 2024 | vol. 11 | e50907 | p.343

(page number not for citation purposes)
Table 3. Medical conditions, preexisting psychiatric diagnoses, and psychiatric screening results.

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>With consult (n=149), n (%)</th>
<th>Without consult (n=82), n (%)</th>
<th>Total (N=231), n (%)</th>
<th>Chi-square (df)</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Medical conditions</strong>&lt;sup&gt;b&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gastrointestinal concerns</td>
<td>82 (55)</td>
<td>37 (45.1)</td>
<td>119 (51.5)</td>
<td>2.0 (1)</td>
<td>.15</td>
</tr>
<tr>
<td>Dental concerns</td>
<td>32 (21.5)</td>
<td>25 (30.5)</td>
<td>57 (24.7)</td>
<td>2.3 (1)</td>
<td>.13</td>
</tr>
<tr>
<td>Menstrual concerns&lt;sup&gt;c&lt;/sup&gt;</td>
<td>16 (53.3)</td>
<td>5 (27.8)</td>
<td>21 (43.8)</td>
<td>3.0 (1)</td>
<td>.13</td>
</tr>
<tr>
<td><strong>General</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Headache</td>
<td>26 (17.4)</td>
<td>8 (9.9)</td>
<td>34 (14.7)</td>
<td>2.5 (1)</td>
<td>.11</td>
</tr>
<tr>
<td>Ear, nose, and throat concerns</td>
<td>17 (11.4)</td>
<td>8 (9.9)</td>
<td>25 (10.8)</td>
<td>0.2 (1)</td>
<td>.70</td>
</tr>
<tr>
<td>Seasonal allergies</td>
<td>34 (22.8)</td>
<td>13 (15.9)</td>
<td>47 (20.3)</td>
<td>1.6 (1)</td>
<td>.21</td>
</tr>
<tr>
<td>Injury pain</td>
<td>14 (9.4)</td>
<td>9 (11)</td>
<td>23 (10)</td>
<td>0.2 (1)</td>
<td>.70</td>
</tr>
<tr>
<td>Thyroid abnormalities</td>
<td>5 (3.4)</td>
<td>6 (7.3)</td>
<td>11 (4.8)</td>
<td>1.8 (1)</td>
<td>.18</td>
</tr>
<tr>
<td>Joint pain</td>
<td>9 (6)</td>
<td>3 (3.7)</td>
<td>12 (5.2)</td>
<td>0.6 (1)</td>
<td>.44</td>
</tr>
<tr>
<td>Seizures</td>
<td>29 (19.5)</td>
<td>16 (19.5)</td>
<td>45 (19.5)</td>
<td>0.0 (1)</td>
<td>.99</td>
</tr>
<tr>
<td>Seizure History</td>
<td>40 (26.8)</td>
<td>18 (22)</td>
<td>58 (25.1)</td>
<td>0.7 (1)</td>
<td>.41</td>
</tr>
<tr>
<td>Sleep disturbance</td>
<td>124 (83.2)</td>
<td>68 (82.9)</td>
<td>192 (83.1)</td>
<td>0.0 (1)</td>
<td>.95</td>
</tr>
<tr>
<td><strong>Preexisting psychiatric diagnoses</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Any psychiatric condition</td>
<td>103 (69.1)</td>
<td>48 (58.5)</td>
<td>151 (65.4)</td>
<td>2.6 (1)</td>
<td>.11</td>
</tr>
<tr>
<td>Depression&lt;sup&gt;d&lt;/sup&gt;</td>
<td>20 (13.4)</td>
<td>14 (17.1)</td>
<td>34 (14.7)</td>
<td>0.6 (1)</td>
<td>.46</td>
</tr>
<tr>
<td>Bipolar disorder&lt;sup&gt;d&lt;/sup&gt;</td>
<td>21 (14.1)</td>
<td>9 (11)</td>
<td>30 (13)</td>
<td>0.5 (1)</td>
<td>.50</td>
</tr>
<tr>
<td>Unspecified mood disorder or DMDD&lt;sup&gt;d,e&lt;/sup&gt;</td>
<td>20 (13.4)</td>
<td>7 (8.5)</td>
<td>27 (11.7)</td>
<td>1.2 (1)</td>
<td>.27</td>
</tr>
<tr>
<td>Anxiety&lt;sup&gt;f&lt;/sup&gt;</td>
<td>62 (41.6)</td>
<td>31 (37.8)</td>
<td>93 (40.3)</td>
<td>0.3 (1)</td>
<td>.57</td>
</tr>
<tr>
<td>PTSD&lt;sup&gt;g&lt;/sup&gt;</td>
<td>11 (7.4)</td>
<td>4 (4.9)</td>
<td>15 (6.5)</td>
<td>0.6 (1)</td>
<td>.46</td>
</tr>
<tr>
<td>Psychotic disorder</td>
<td>14 (9.4)</td>
<td>1 (1.2)</td>
<td>15 (6.5)</td>
<td>5.8 (1)</td>
<td>.02</td>
</tr>
<tr>
<td>ADHD&lt;sup&gt;b&lt;/sup&gt;</td>
<td>50 (33.6)</td>
<td>26 (31.7)</td>
<td>76 (32.9)</td>
<td>0.1 (1)</td>
<td>.78</td>
</tr>
<tr>
<td><strong>Psychiatric screening status</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Any psychiatric condition</td>
<td>146 (98)</td>
<td>80 (97.6)</td>
<td>226 (97.8)</td>
<td>0.1 (1)</td>
<td>.83</td>
</tr>
<tr>
<td>Unipolar depression&lt;sup&gt;d&lt;/sup&gt;</td>
<td>61 (40.9)</td>
<td>30 (36.6)</td>
<td>91 (39.4)</td>
<td>0.4 (1)</td>
<td>.52</td>
</tr>
<tr>
<td>Episodic expansive mood and bipolar disorder&lt;sup&gt;d&lt;/sup&gt;</td>
<td>60 (40.3)</td>
<td>22 (34.1)</td>
<td>88 (38.1)</td>
<td>0.8 (1)</td>
<td>.36</td>
</tr>
<tr>
<td>Anxiety</td>
<td>130 (87.2)</td>
<td>71 (86.6)</td>
<td>201 (87)</td>
<td>0.0 (1)</td>
<td>.89</td>
</tr>
<tr>
<td>PTSD</td>
<td>37 (24.8)</td>
<td>15 (18.3)</td>
<td>52 (22.5)</td>
<td>1.3 (1)</td>
<td>.26</td>
</tr>
<tr>
<td>Psychosis</td>
<td>52 (34.9)</td>
<td>15 (18.3)</td>
<td>67 (29)</td>
<td>7.1 (1)</td>
<td>.008</td>
</tr>
<tr>
<td>ADHD&lt;sup&gt;b&lt;/sup&gt;</td>
<td>102 (68.5)</td>
<td>56 (68.3)</td>
<td>158 (68.4)</td>
<td>0.0 (1)</td>
<td>.98</td>
</tr>
</tbody>
</table>

<sup>a</sup>Includes informal and formal consults.

<sup>b</sup>Medical conditions perceived by the caregiver as contributing to the current presentation of distress.

<sup>c</sup>Analysis for menstrual concerns restricted to female patients aged >12.

<sup>d</sup>Unipolar depression, unspecified mood disorder and disruptive mood dysregulation disorder, and episodic expansive mood and bipolar disorder are mutually exclusive categories.

<sup>e</sup>DMDD: disruptive mood dysregulation disorder.

<sup>f</sup>Preexisting diagnosis of obsessive-compulsive disorder is included within the anxiety disorder category.

<sup>g</sup>PTSD: posttraumatic stress disorder.

<sup>h</sup>ADHD: attention-deficit/hyperactivity disorder.
Caregiver-Reported Medical Conditions

Table 3 describes medical conditions reported by caregivers. Caregivers of 73.2% (169/231) of the individuals identified at least 1 physical concern that they perceived as contributing to distress. The most common conditions were gastrointestinal concerns (119/231, 51.5%), menstrual concerns (21/48, 44% of female patients aged >12 y), seasonal allergies (47/231, 20.3%), and seizures (45/231, 19.5%).

Psychiatric Screening Results

Table 3 lists the frequency of positive psychiatric screening results. All but 2% (5/231) of the individuals screened positive for a psychiatric condition, with a mean of 2.8 (SD 1.1; range 0-5) conditions per individual. Of those who were classified as having bipolar disorder, 89% (78/88) screened positive for a recent depressive episode. Positive screen status for psychiatric conditions were similar between those with a consult and those without, except in the case of psychosis (52/149, 34.9% vs 15/82, 18%; *P* = .008).

Psychotropic Medication Use

Table 4 reports on the frequency of medication use. Most of the individuals (194/231, 84%) were taking psychotropic medication, and the majority were receiving antipsychotics (142/231, 61.5%) and antidepressants (129/231, 55.8%).

<table>
<thead>
<tr>
<th>Medication</th>
<th>With consult (n=149), n (%)</th>
<th>Without consult (n=82), n (%)</th>
<th>Total (N=231), n (%)</th>
<th>Chi-square (df)</th>
<th><em>P</em> value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Any medication</td>
<td>146 (98)</td>
<td>68 (82.9)</td>
<td>214 (92.6)</td>
<td>17.6 (1)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Any psychotropic medication</td>
<td>136 (91.3)</td>
<td>58 (70.7)</td>
<td>194 (84)</td>
<td>16.6 (1)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Antipsychotic</td>
<td>102 (68.5)</td>
<td>40 (48.8)</td>
<td>142 (61.5)</td>
<td>8.7 (1)</td>
<td>.003</td>
</tr>
<tr>
<td>Antidepressant(^a)</td>
<td>87 (58.4)</td>
<td>42 (51.2)</td>
<td>129 (55.8)</td>
<td>1.1 (1)</td>
<td>.29</td>
</tr>
<tr>
<td>Anxiolytic(^b)</td>
<td>66 (44.3)</td>
<td>23 (28.0)</td>
<td>89 (38.5)</td>
<td>5.9 (1)</td>
<td>.02</td>
</tr>
<tr>
<td>Anticonvulsant(^c)</td>
<td>45 (30.2)</td>
<td>12 (14.6)</td>
<td>57 (24.7)</td>
<td>6.9 (1)</td>
<td>.009</td>
</tr>
<tr>
<td>Lithium</td>
<td>10 (6.7)</td>
<td>7 (8.5)</td>
<td>17 (7.4)</td>
<td>0.3 (1)</td>
<td>.61</td>
</tr>
<tr>
<td>Alpha-2 agonist</td>
<td>72 (48.3)</td>
<td>27 (32.9)</td>
<td>100 (43.3)</td>
<td>5.1 (1)</td>
<td>.02</td>
</tr>
<tr>
<td>Stimulant and atomoxetine</td>
<td>30 (20.1)</td>
<td>14 (17.1)</td>
<td>44 (19)</td>
<td>0.3 (1)</td>
<td>.57</td>
</tr>
</tbody>
</table>

\(^a\)Selective serotonin reuptake inhibitors, duloxetine, tricyclics, mirtazapine, and trazodone were included exclusively within the antidepressant category.

\(^b\)Benzodiazepines, buspirone, hydroxyzine, beta-blockers, and prazosin were included exclusively within the anxiolytic category.

\(^c\)Anticonvulsant medication use in the absence of a reported seizure history.

Working Psychiatric Diagnoses

Of the 149 individuals who received a consult, 148 (99.3%) were diagnosed with at least 1 psychiatric condition with a mean of 2.7 (SD 1.0; range 0-5) diagnoses per individual. The conditions identified were anxiety (129/149, 86.6%), ADHD (84/149, 56.4%), bipolar disorder (67/149, 45%), unipolar depression (33/149, 22.1%), PTSD (35/149, 23.5%), and psychosis (31/149, 20.8%). Furthermore, 25 (16.8%) of the 149 individuals were diagnosed with either unspecified mood disorder or DMDD. Nearly all individuals identified with psychosis (29/31, 94%) had a co-occurring mood disorder diagnosis: bipolar disorder (22/31, 71%), unipolar depression (5/31, 16%), and unspecified mood disorder or DMDD (2/31, 6%).

Evaluation of Validity

Sources of Distress accuracy rates ranged from 76% (95% CI 69%-82%) for ADHD to 91% (95% CI 85%-95%) for PTSD and exceeded those of preexisting diagnoses, except in the case of psychosis, for which the accuracy rates were equivocal (82%, 95% CI 75%-87%; Table 5). The survey demonstrated higher NPVs (81%-98%) than PPVs (51%-78%) for all conditions, with the exceptions of anxiety (53% and 92%, respectively) and episodic expansive mood bipolar disorder (85% and 90%, respectively). Low PPVs were notable for depression (51%) and psychosis (54%).
Table 5. Association between consult diagnoses after completing Sources of Distress with preexisting psychiatric diagnoses and Sources of Distress screening status (n=149).

<table>
<thead>
<tr>
<th>Working diagnosis</th>
<th>Preexisting psychiatric diagnosis</th>
<th>Sources of Distress screening status</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accuracy rate</td>
<td>NPV</td>
</tr>
<tr>
<td></td>
<td>(%)</td>
<td>(%)</td>
</tr>
<tr>
<td>Unipolar depression</td>
<td>78 (71-84)</td>
<td>51</td>
</tr>
<tr>
<td>Case negative</td>
<td>106</td>
<td>10</td>
</tr>
<tr>
<td>Case positive</td>
<td>23</td>
<td>10</td>
</tr>
<tr>
<td>Episodic expansive mood and bipolar disorder</td>
<td>62 (55-70)</td>
<td>90</td>
</tr>
<tr>
<td>Case negative</td>
<td>77</td>
<td>5</td>
</tr>
<tr>
<td>Case positive</td>
<td>51</td>
<td>16</td>
</tr>
<tr>
<td>DMDD and unspecified mood disorder</td>
<td>82 (75-88)</td>
<td>N/A</td>
</tr>
<tr>
<td>Case negative</td>
<td>113</td>
<td>11</td>
</tr>
<tr>
<td>Case positive</td>
<td>16</td>
<td>9</td>
</tr>
<tr>
<td>Anxiety disorder</td>
<td>52 (44-60)</td>
<td>92</td>
</tr>
<tr>
<td>Case negative</td>
<td>18</td>
<td>2</td>
</tr>
<tr>
<td>Case positive</td>
<td>69</td>
<td>60</td>
</tr>
<tr>
<td>Posttraumatic stress disorder</td>
<td>84 (77-89)</td>
<td>78</td>
</tr>
<tr>
<td>Case negative</td>
<td>114</td>
<td>0</td>
</tr>
<tr>
<td>Case positive</td>
<td>24</td>
<td>11</td>
</tr>
<tr>
<td>Psychotic disorder</td>
<td>82 (75-87)</td>
<td>54</td>
</tr>
<tr>
<td>Case negative</td>
<td>113</td>
<td>5</td>
</tr>
<tr>
<td>Case positive</td>
<td>22</td>
<td>9</td>
</tr>
<tr>
<td>Attention-deficit/hyperactivity disorder</td>
<td>66 (59-74)</td>
<td>74</td>
</tr>
<tr>
<td>Case negative</td>
<td>57</td>
<td>8</td>
</tr>
<tr>
<td>Case positive</td>
<td>42</td>
<td>42</td>
</tr>
</tbody>
</table>

*Preexisting diagnoses included caregiver-reported diagnoses in Sources of Distress and diagnoses in the medical record before survey completion.

**PPV:** positive predictive value = ☐

***NPV:** negative predictive value = ☐

**Accuracy rate = ☐

**Depression, episodic expansive mood and bipolar disorder, and disruptive mood dysregulation disorder and unspecified mood disorder are mutually exclusive categories. There is no Sources of Distress screening algorithm for disruptive mood dysregulation disorder or unspecified mood disorder.

**Preexisting and working diagnoses included schizoaffective disorder when hypomanic, manic, or mixed episode was specified.

**DMDD:** disruptive mood dysregulation disorder.

**N/A:** not applicable.

**Preexisting and working diagnoses of anxiety disorder and obsessive-compulsive disorder are combined to coincide with anxiety disorder screening status.

**Preexisting and working diagnoses were schizophrenia, schizoaffective disorder, unspecified psychotic disorder, and psychotic features associated with a mood disorder.

**Exploration of Mood Disorder Categories**

Figure 1 demonstrates the distribution of mood disorder diagnoses among individuals based on (1) preexisting mood disorder diagnosis and (2) Sources of Distress mood disorder screening status. The majority of the individuals (18/25, 72%) who received a working diagnosis of unspecified mood disorder and DMDD screened positive for either unipolar depression or bipolar disorder.
Figure 1. Comparison of mood disorder categorization between working diagnosis established during consultation and (A) preexisting diagnosis and (B) Sources of Distress positive screen. DMDD: disruptive mood dysregulation disorder.

Discussion

Principal Findings

The focus group feedback indicates that Sources of Distress provides an acceptable means for caregivers to share their knowledge and insights about individuals with DD who present in crisis. As a screening tool, this survey demonstrates good accuracy, although additional work is needed to differentiate among mood disorders. The purpose of this survey is to screen individuals with DD for mental health and common medical concerns in health care settings when they present in crisis. By querying what symptom clusters and physical conditions coincide with their patient’s crisis, providers can direct their evaluation toward specific psychiatric and medical conditions that have established treatment protocols in the general population. This approach aims to reduce diagnostic overshadowing and improve medical decision-making surrounding the management of agitation and aggression in individuals with DD. Focus group participants validated the survey content and provided recommendations that prompted the inclusion of additional modules and items. Despite its length (ie, 15-20 min), participants reported positive acceptability of the survey’s format and duration. This feedback may reflect the convenience of completing a web-based survey at home versus in the medical setting and highlights caregivers’ motivation toward understanding potential factors contributing to the person’s distress. After incorporating caregiver recommendations, Sources of Distress content was also reviewed and supported by clinical and research experts.

Caregivers of most of the individuals (200/231, 86.6%) identified agitation as a presenting concern. The Food and Drug Administration has approved short-term antipsychotic
medication for treating irritability in individuals with autism [47]; 61.5% (142/231) of the individuals were taking antipsychotics at the time of presenting in crisis. This frequency exceeds previously reported estimates of antipsychotic use in the population with DD (ie, 10%-48%) and reflects the high acuity and potentially treatment-resistant nature of individuals for whom the survey was completed [48,49]. This study group’s acuity is further supported by the high frequency in which severe mental health conditions were diagnosed in those receiving a consultation (eg, bipolar disorder and psychosis).

Anxiety was the most common condition to screen positive (201/231, 87%) and be established as a working diagnosis (129/149, 86.6%). These rates exceeded measured anxiety prevalence rates in the population with DD (ie, 20%-77%), indicating a higher propensity toward experiencing anxiety among those presenting in crisis [12,15,16]. As a precipitant of distress, prior studies have identified aggression, disruptive behavior, sleep disturbance, and self-injurious behavior as symptoms of anxiety among individuals with DD [4,8,50]. To reduce overclassification among individuals whose autism core features overlap with some anxiety symptoms [51], the Sources of Distress anxiety scoring algorithm was set at a higher threshold than the generalized anxiety disorder criteria described in DM-ID-2. The survey’s low NPV (53%) and high PPV (92%) for anxiety likely reflect this adaptation.

Sources of Distress captured well the presence of a mood disturbance; however, the type of mood disorder was not. Study results report a diagnosis frequency of 16.8% (25/149) for unspecified mood disorder and DMDD and indicate the need to add items and a screening algorithm for this condition. The low PPV (51%) for depression primarily resulted from individuals screening positive for depression who were subsequently diagnosed with unspecified mood disorder and DMDD. The DM-ID-2, survey data, and records review will inform new items and algorithm development as well as revisions for the depression screening algorithm. In the interim, the Sources of Distress report will replace the “depression” category label with “depression and unspecified mood disorder” to broaden the range of conditions which it currently captures.

Caregivers of the majority of the individuals (169/231, 73.2%) identified at least 1 physical concern that they perceived as contributing to distress. As agitation may be one of the few visible indicators of pain in an individual with limited expressive language ability and DD, sources of pain should be considered when unexplained agitation is present [3,9,52]. Limited access to medical care by the population with DD further reduces the likelihood that pain and other underlying physical causes of agitation are recognized [53]. Through Sources of Distress, caregivers demonstrated their ability to provide meaningful insight into the potential presence of physical discomfort. This attention was directed most frequently to gastrointestinal, menstrual, dental, and seizure concerns.

**Limitations**

The generalizability of study results is limited to the geographic, racial, and ethnic diversity of Utah. While survey access requires internet or smartphone access, it has been completed by parents without this access through the assistance of state-sponsored support coordinators and medical assistants. Sources of Distress has a Spanish translation available (Causas de Aficción); however, these data were not included because its content has not yet been validated by Spanish-speaking caregivers and individuals who are affected. The expert leading the consult team was a member of the survey’s development team, which introduces the inherent bias of evaluating for the presence of mental health conditions through the lens of DM-ID-2 criteria on which survey components were also based. While the DM-ID-2 is well recognized and accepted in the ID provider community, few autism specialty providers are familiar with its use.

**Future Directions**

Edits and additions to mood disorder items and scoring algorithms are being made to improve differentiation across mood disorders. Branching logic that incorporates the individual’s language ability has recently been added to the psychosis module to improve question clarity and scoring algorithm accuracy. The most updated version of the Sources of Distress can be accessed through the Utah Department of Health and Human Services Autism Systems Development Program webpage [39]. Reevaluation of the survey’s PPVs, NPVs, and accuracy will follow the completion of these changes. Additional studies of this survey are needed to measure its acceptability and validity in clinical settings outside of Utah and by other DD specialty providers. REDCap has also demonstrated capacity to integrate digital mental health screening results into electronic medical records, significantly improving provider adoption of the screening tools [54]. The integration of Sources of Distress into electronic medical records could further enhance its impact on provider efficiency. This survey has already been used during medical evaluations to facilitate the consideration of potential discordance between medications prescribed and conditions present [55]. Prospective studies are merited to determine the survey’s impact on treatment approaches, hospital and emergency department use, and outcomes for individuals with DD who experience crisis.

**Conclusions**

Individuals with DD presenting in crisis experience high rates of psychiatric disorders and medical concerns that may contribute to, or manifest as, distress. Sources of Distress is a valuable screening tool for psychiatric and medical conditions that commonly accompany treatment-resistant agitation in individuals with DD. When systematically queried, caregivers’ knowledge provides essential information to minimize diagnostic overshadowing and support an evaluation focused on the individual rather than their disability when persistent agitation is assessed in the population with DD.
Acknowledgments

The authors thank the individuals with developmental disabilities and their parents and caregivers whose support and participation shaped the development, validation, and publication of the Sources of Distress. In particular, the authors appreciate the insight and assistance provided by Jaye Olafson. The Autism Council of Utah funded the development of this survey, and access to REDCap (Research Electronic Data Capture) is supported through U1TR004409 NCATS/NIH.

Conflicts of Interest

The Sources of Distress survey is copyrighted by the University of Utah; DAB, JD, and WW are listed as inventors. DAB consults for BioMarin Pharmaceutical Inc, Encoded Therapeutics, Taysha Gene Therapies, and Synlogic Therapeutics and attended an advisory board meeting for Sanofi. All other authors declare no other conflicts of interest.

Multimedia Appendix 1

A sample of the Sources of Distress report.

References


Abbreviations

- **ADHD**: attention-deficit/hyperactivity disorder
- **DD**: developmental disabilities
- **DDMD**: disruptive mood dysregulation disorder
- **ID**: intellectual disability
- **NPV**: negative predictive value
- **PAS-ADD**: Psychiatric Assessment Schedule for Adults with Developmental Disabilities
- **PPV**: positive predictive value
- **PTSD**: posttraumatic stress disorder
- **REDCap**: Research Electronic Data Capture

https://mental.jmir.org/2024/1/e50907

JMIR Ment Health 2024 | vol. 11 | e50907 | p.351

(page number not for citation purposes)
Developing and Implementing a Web-Based Branching Logic Survey to Support Psychiatric Crisis Evaluations of Individuals With Developmental Disabilities: Qualitative Study and Evaluation of Validity

Bilder DA, Mthembu M, Worsham W, Aguayo P, Knight JR, Deng SW, Singh TP, Davis J

JMIR Ment Health 2024;11:e50907

URL: https://mental.jmir.org/2024/1/e50907
doi:10.2196/50907
PMID:38551644

©Deborah A Bilder, Mariah Mthembu, Whitney Worsham, Patricia Aguayo, Jacob R Knight, Steven W Deng, Tejinder P Singh, John Davis. Originally published in JMIR Mental Health (https://mental.jmir.org), 29.03.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Mental Health, is properly cited. The complete bibliographic information, a link to the original publication on https://mental.jmir.org/, as well as this copyright and license information must be included.
Understanding Public Perceptions of Virtual Reality Psychological Therapy Using the Attitudes Towards Virtual Reality Therapy (AVRT) Scale: Mixed Methods Development Study

Aislinn D Gomez Bergin1,2,3,4*, PhD; Aoife M Allison3*, MSc; Cassie M Hazell5*, PhD

1National Institute of Health and Care Research MindTech MedTech Co-operative, Institute of Mental Health, University of Nottingham, Nottingham, United Kingdom
2National Institute of Health and Care Research Nottingham Biomedical Research Centre, University of Nottingham, Nottingham, United Kingdom
3Mental Health and Clinical Neurosciences, School of Medicine, University of Nottingham, Nottingham, United Kingdom
4School of Computer Science, University of Nottingham, Nottingham, United Kingdom
5Department of Psychological Interventions, School of Psychology, University of Surrey, Guilford, United Kingdom

*all authors contributed equally

Corresponding Author:
Aislinn D Gomez Bergin, PhD
National Institute of Health and Care Research MindTech MedTech Co-operative
Institute of Mental Health
University of Nottingham
Jubilee Campus
Triumph Rd
Nottingham, NG7 2TU
United Kingdom
Phone: 44 115 82 30431
Email: aislinn.bergin@nottingham.ac.uk

Abstract

Background: Virtual reality (VR) psychological therapy has the potential to increase access to evidence-based mental health interventions by automating their delivery while maintaining outcomes. However, it is unclear whether these more automated therapies are acceptable to potential users of mental health services.

Objective: The main aim of this study was to develop a new, validated questionnaire to measure public perceptions of VR therapy (VRT) guided by a virtual coach. We also aimed to explore these perceptions in depth and test how aspects such as familiarity with VR and mental health are associated with these perceptions, using both quantitative and qualitative approaches.

Methods: We used a cross-sectional mixed methods design and conducted an exploratory factor analysis of a questionnaire that we developed, the Attitudes Towards Virtual Reality Therapy (AVRT) Scale, and a qualitative content analysis of the data collected through free-text responses during completion of the questionnaire.

Results: We received 295 responses and identified 4 factors within the AVRT Scale, including attitudes toward VRT, expectation of presence, preference for VRT, and cost-effectiveness. We found that being more familiar with VR was correlated with more positive attitudes toward VRT (factor 1), a higher expectation of presence (factor 2), a preference for VRT over face-to-face therapy (factor 3), and a belief that VRT is cost-effective (factor 4). Qualitative data supported the factors we identified and indicated that VRT is acceptable when delivered at home and guided by a virtual coach.

Conclusions: This study is the first to validate a scale to explore attitudes toward VRT guided by a virtual coach. Our findings indicate that people are willing to try VRT, particularly because it offers increased access and choice, and that as VR becomes ubiquitous, they will also have positive attitudes toward VRT. Future research should further validate the AVRT Scale.

(JMIR Ment Health 2024;11:e48537) doi:10.2196/48537

KEYWORDS
psychological interventions; digital; virtual reality; virtual agent; mental health; presence

https://mental.jmir.org/2024/1/e48537

JMIR Ment Health 2024 | vol. 11 | e48537 | p.353
Background

Virtual reality (VR) is an immersive environment where people can interact using either computer equipment, such as a screen and mouse, or VR-enabled headsets and controllers, where additional sensors can track the users’ actions in real time. The latter application provides people with a greater sense of presence, a term used to describe how closely a virtual environment is interpreted as real [1]. In recent years, VR has been used successfully in a range of health care settings to improve and increase access to treatment [2]. In particular, VR has been used in the delivery of psychological therapies for a range of mental health problems, with several decades of evidence demonstrating its clinical efficacy in the treatment of psychosis, depression, anxiety, and eating disorders [3-8].

VR therapies (VRTs) were initially developed to be used by therapists as an adjunct or tool in their delivery of therapy. However, the need for a real-world therapist to deliver VRTs presents a key challenge for their widespread implementation [9]. Researchers have shown that the automation of some therapeutic elements may overcome this barrier to meet the increase in demand for treatment globally [10,11]. Emerging evidence demonstrates that VRTs can be successfully delivered with little to no therapist involvement, with virtual coaches supporting people receiving therapy for fear of heights and agoraphobia in the context of psychosis [10,11]. Virtual coaches are also known as virtual agents [12]. These characters are not under human control and therefore offer automation of therapies, in which dialogue and responses are scripted instead of the formulation that is offered by real-world therapists.

There are financial and resource incentives for mental health services to offer more automated therapies [13]. Clinicians also appear to be in support of VRTs. For example, cognitive behavior therapists [14] and psychiatric health care staff [15] reported positive attitudes toward VRT, particularly when they were more familiar with VR. However, these studies do not consider how staff feel about VRT guided by a virtual coach and, notably, do not explore patient and public perceptions of VRT. VRT dropout rates have been used as a proxy measure of patient experience, and these figures show similar dropout rates to therapies delivered without VR [16]. However, dropout rates from research do not provide us with a clear picture of whether people will engage in therapies delivered using VR, including those guided by a virtual coach. A content analysis of social media posts by the public appears to suggest an interest in the application of VR in mental health care [17]. Staff and service users also have positive views toward their use in mental health inpatient facilities [18]. However, these studies still do not directly ask potential users of mental health services whether they would be willing to try VRT guided by a virtual coach or the factors that relate to this willingness.

Health care staff, when asked for their views regarding service users’ opinions of VRT, had concerns regarding patients’ willingness to accept their use as part of their mental health care package [19]. Furthermore, the literature has highlighted a lack of personalization as a barrier to engagement with digital mental health interventions [20]. It is unclear whether this indicates that automated VRTs can be sufficient when scripts are relevant to the experience of the individual. It is possible that the presence of a virtual coach may encourage more positive attitudes and a willingness to try VRTs. There is a need to understand service user and public perspectives on the use and delivery of VRTs and those guided by a virtual coach and how different factors may affect the uptake of such interventions.

Aim

The main aim of this study was to establish a new, validated questionnaire to measure the perceptions of VRT guided by a virtual coach. Second, we aimed to explore how these perceptions are associated with familiarity with VR and mental health, using both quantitative and qualitative approaches.

Methods

Study Design

This study used mixed methods with a cross-sectional design. Data were collected from a web-based questionnaire using Jisc software [21].

Participants

The participants were recruited via social media to complete the web-based questionnaire. We aimed to recruit a minimum of 200 participants in line with sample size recommendations for exploratory factor analysis (EFA) [22]. To be eligible to participate, persons were required to be a resident of the United Kingdom or Ireland and aged ≥18 years. A link to the web-based survey was included in all promotional materials.

Measures

Demographics

The participants were asked to provide basic demographic information, including their age and sex, as well as whether they were identified as having a mental health condition, had ever experienced therapy, or had supported anyone with a mental health condition. Furthermore, they were asked about their experience of VR (from never to ≥10 times) and their familiarity with VR, VRT, and mental health conditions.

The Attitudes Towards Virtual Reality Therapy Scale

The Attitudes Towards Virtual Reality Therapy (AVRT) Scale was developed by AMA and ADGB. Items were based on themes identified in previous literature that contribute to perceptions of VRT and digital mental health interventions [14,23-29]. Items surrounding the virtual coach drew on the literature related to therapeutic alliance and focused on trust, comfort, and need [30].

We designed 54 items all assessing different aspects of attitudes toward VRT, including 9 items related to attitudes toward VRT delivered by a virtual coach. Each item used a 7-point Likert scale where participants rated their agreement from “strongly agree” to “strongly disagree.” Strong agreement or disagreement with 16 of these items triggered a free-text question for participants to provide context using free-text responses. Responses were scored from 1 (strongly disagree) to 7 (strongly agree).
agree). A higher score indicates more positive perceptions of VRT. A total of 27 items were reverse-worded and therefore reverse-coded.

Furthermore, participants were invited to respond to 3 additional free-text questions asking what they would like to know more about, how they think their level of VR experience has influenced their perceptions, and what they think the best setting for VRT would be.

**Procedure**

Upon opening the questionnaire, participants were first shown the information sheet, followed by a consent statement. After consenting, participants were asked to enter a unique identification code so that their anonymized responses could be identified later. Participants were then asked to provide basic demographic information, followed by an explanatory paragraph (Multimedia Appendix 1) about VRT and the virtual coach, which was described as “a computer-generated avatar” that “guides the patients through the scenarios and offers advice and encouragement.” This was followed by items on experience with VR and mental health, the AVRT Scale, and the 3 free-text questions. After completing these questionnaires, participants were presented with a debrief statement.

**Ethics Approval**

Ethics approval was granted by the Division of Psychiatry and Applied Psychology Ethics Subcommittee of the University of Nottingham (Project ID 1534).

**Statistical Analysis**

Raw data were downloaded from Jisc [21] into SPSS Statistics software (version 25; IBM Corp) [31]. We removed responses from participants who did not meet the inclusion criteria, did not provide consent, or had missing data. Sample characteristics were summarized using descriptive statistics.

To validate our new questionnaire, we conducted an EFA using principal component analysis with a varimax rotation. We assessed the suitability of the data for factor analysis using Bartlett’s test of sphericity and the Kaiser-Meyer-Olkin measure of sampling adequacy (acceptable adequacy ≥0.6). All 54 items using Likert scales were included in the EFA. Items were first screened to check for multicollinearity and poor correlations with the other items. We operationalized this screening by assessing the determinant and searching for any interitem correlations of ≥0.7 or where most coefficients were nonsignificant or <0.4. Any items that failed this initial screening were removed, and the EFA was rerun. Factors were derived using eigenvalues ≥1, where the Kaiser criterion [32] was met, and in combination with the point of inflection on the scree plot, where they were not. For factors to be retained, they must comprise at least 3 items. Where items loaded on >1 factor, the item was assigned to the factor with which it made the most thematic sense.

The questionnaire included both positively and negatively worded items. Once the factors were established, we reverse-scored negatively worded items so that a higher score indicated a more favorable attitude. We then assessed the internal consistency of the final factor structure using Cronbach α, with an acceptable internal consistency of ≥0.7 [33]. Further items may be removed at this point, where the scale reliability can be substantially improved if the item is removed. The factor scores were computed using the mean and SD of the scale sum.

We conducted a series of Pearson r correlations to assess whether there was a relationship among the scale totals of the derived factors and lived experience of VR, VRT, and mental health problems.

**Qualitative Analysis**

All responses to the free-text response questions (ie, 16 free-text boxes triggered by extreme responses to survey questions and 3 additional free-text questions) were uploaded to NVivo (version 12 for Mac; QSR International). Qualitative content analysis [34] was used to quantify and summarize the qualitative data within the broader context of the AVRT Scale. All data were coded inductively by a qualitative researcher (ADGB), where several codes could be applied to a single response. The codes were collated by questions or items. The study team met to review and revise any discrepancies or discuss any questions. Findings were then summarized according to each question or item and were presented within the factors of the AVRT Scale.

**Results**

**Sample Characteristics**

We collected 295 responses to the survey. Our sample reflected a range of age groups. The majority were female, had used VR at least once, and had no personal or professional experience with mental health problems. However, most participants had supported a friend or family member with poor mental health (Table 1).
Table 1. Sample characteristics (N=295).

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Age group (years), n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>18-24</td>
<td>89 (30.2)</td>
</tr>
<tr>
<td>25-29</td>
<td>39 (13.2)</td>
</tr>
<tr>
<td>30-39</td>
<td>45 (15.3)</td>
</tr>
<tr>
<td>40-49</td>
<td>49 (16.6)</td>
</tr>
<tr>
<td>50-64</td>
<td>64 (21.7)</td>
</tr>
<tr>
<td>&gt;65</td>
<td>9 (3.1)</td>
</tr>
<tr>
<td><strong>Sex, n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>83 (28.1)</td>
</tr>
<tr>
<td>Female</td>
<td>209 (70.8)</td>
</tr>
<tr>
<td>Prefer not to say</td>
<td>3 (1)</td>
</tr>
<tr>
<td><strong>Frequency of experiencing VR^a, n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>Never</td>
<td>112 (38)</td>
</tr>
<tr>
<td>Once</td>
<td>48 (16.3)</td>
</tr>
<tr>
<td>&lt;5 times</td>
<td>88 (29.8)</td>
</tr>
<tr>
<td>5-9 times</td>
<td>20 (6.8)</td>
</tr>
<tr>
<td>≥10 times</td>
<td>27 (9.2)</td>
</tr>
<tr>
<td><strong>Participants identifying as having a mental health condition, n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>102 (34.6)</td>
</tr>
<tr>
<td>No</td>
<td>181 (61.4)</td>
</tr>
<tr>
<td>Prefer not to say</td>
<td>12 (4.1)</td>
</tr>
<tr>
<td><strong>Participants with experience in therapy for a mental health condition, n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>131 (44.4)</td>
</tr>
<tr>
<td>No</td>
<td>160 (54.2)</td>
</tr>
<tr>
<td>Prefer not to say</td>
<td>4 (1.4)</td>
</tr>
<tr>
<td><strong>Participants who have supported a friend or family member or colleague with a mental health condition, n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>253 (85.8)</td>
</tr>
<tr>
<td>No</td>
<td>39 (13.2)</td>
</tr>
<tr>
<td>Prefer not to say</td>
<td>3 (1)</td>
</tr>
<tr>
<td><strong>Participants who have worked in a caring role for people with mental health conditions, n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>108 (36.6)</td>
</tr>
<tr>
<td>No</td>
<td>186 (63.1)</td>
</tr>
<tr>
<td>Prefer not to say</td>
<td>1 (0.3)</td>
</tr>
<tr>
<td>Participants familiar with VR, mean (SD)</td>
<td>2.33 (1.02)</td>
</tr>
<tr>
<td>Participants familiar with VR therapy, mean (SD)</td>
<td>1.35 (0.75)</td>
</tr>
<tr>
<td>Participants familiar with mental health conditions, mean (SD)</td>
<td>3.65 (0.96)</td>
</tr>
</tbody>
</table>

^aVR: virtual reality.

**Quantitative Analysis**

**Item Screening**

The data were found to be appropriate for EFA (Kaiser-Meyer-Olkin=0.93; $\chi^2_{1431}=10,205.5$, $P<.001$). The determinant suggested that there was multicollinearity, and inspection of the correlation coefficients revealed 2 pairs of items that were highly correlated (items 7 and 8=0.83; items 36 and 37=0.87); therefore, we removed 1 item from each pair of correlations (items 8 and 36). Items 13, 32, and 33 were removed, as the majority of interitem correlations were
nonsignificant ($P > .05$). Furthermore, we removed items 6, 19, 25, 26, 28, 31, and 47 as either all or all but one of the correlation coefficients was <0.4. In total, we removed 12 items and then reran the EFA on the remaining 42 items.

The determinant again indicated that multicollinearity was an issue. We identified 3 pairs of correlations with coefficients >0.7 (items 10 and 11=0.76; items 29 and 49=0.76; items 35 and 37=0.71); therefore, we removed 1 item from each pair (items 10, 29, and 35) and reran the EFA on the remaining 39 items.

**Exploratory Factor Analysis**

The Kaiser criterion was met (n=295; average communalities 0.64) [32]. Therefore, the factor structure was determined based on eigenvalues >1. The rotated factor solution suggested 7 factors, which explained 63.64% of the variance. However, 3 factors were not retained because they contained <3 items. The removal of these factors resulted in the removal of items 7, 23, 24, 27, 48, and 49. The resulting 33 items were entered into a final EFA. A 4-factor solution was suggested based on the eigenvalues and the scree plot, which explained 58.61% of the variance.

Factor 1 had 13 items that assessed respondents’ support for VRT, including 6 reverse-worded items (factor 1: attitude toward VRT). Factor 2 had 9 items. These items, including 7 reverse-worded items, assessed the extent to which the respondents expected VRT to be immersive (factor 2: the expectation of presence). Factor 3 had 7 items asking respondents to compare VRT to aspects of face-to-face therapies (factor 3: preference for VRT). Factor 4 had 4 items each assessing different aspects of the cost-effectiveness of VRT (factor 4: cost-effectiveness). Refer to Multimedia Appendix 2 for the final factor structure.

**Scale Reliability**

After reverse-scoring the reverse-worded items, we computed Cronbach $\alpha$ values for each scale. All scales had strong internal consistency (all Cronbach $\alpha > 0.82$). The scale reliabilities could not be improved by removing any of the items. A higher score on each of the subscales suggested a more favorable attitude (factor 1), increased perceived presence (factor 2), a preference for VRT over traditional therapies (factor 3), and agreement that VRT is cost-effective (factor 4). The desired direction for each subscale to demonstrate support for VRT was high for factors 1, 3, and 4 and low for factor 2.

**Relationship Between Scales and Lived Experience**

There was a significant relationship between the participants’ familiarity with VR and their scores on all the factors. Familiarity with VR was positively associated with a more favorable attitude toward VRT (factor 1), higher expectations of presence (factor 2), a preference over face-to-face therapy (factor 3), and a belief that VRT is cost-effective (factor 4). We also found significant positive correlations between factors 1, 2, and 3, but not factor 4, and familiarity with the VRT. There was no significant correlation between mental health familiarity and the scores for any factors. Multimedia Appendix 2 presents the correlation coefficients and associated significance scores.

**Qualitative Results**

**Qualitative Questions**

Table 2 presents the initial qualitative questions that all participants were asked.

<table>
<thead>
<tr>
<th>Type of data</th>
<th>Question</th>
<th>Response, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Qualitative question 1</td>
<td>Which aspects of virtual reality therapy guided by a virtual coach would you like to know more about?</td>
<td>226 (76.6)</td>
</tr>
<tr>
<td>Qualitative question 2</td>
<td>How has your previous experience of virtual reality (minimal or extensive) influenced your perceptions of virtual reality therapy guided by a virtual coach?</td>
<td>236 (80)</td>
</tr>
<tr>
<td>Qualitative question 3</td>
<td>If you were offered virtual reality therapy guided by a virtual coach, where do you think the best place to do the therapy would be?</td>
<td>245 (83)</td>
</tr>
</tbody>
</table>

**Which Aspects of VRT Guided by a Virtual Coach Would You Like to Know More About?**

Of the 226 participants, 19 (8.4%) indicated that they did not want to know anything more. Those who provided reasons indicated that they did not want to try or did not know enough. Of them, 22 (9.7%) participants indicated that, as they did not know enough, they would like to find out more; 5 (2%) indicated that they would like to try out VRT; and 9 (3.9%) asked regarding its cost. In total, 13 (5.8%) participants were curious about the conditions that could be targeted with the VRT, specifically regarding its use for anxiety disorders, depression, and emotion regulation.

Many participants asked how it could be tailored or personalized for them (29/226, 12.8%). This meant thinking about their position within the interaction, asking about safety or how much control they would have, and whether VRT might have a negative effect and how this would be monitored. Of 226 participants, 47 (20.8%) asked about the virtual coach, wanting to know how real it would be, how much of the language would be generic or responsive to them, and how they could build a relationship with the virtual coach. Of these, many wanted more information about whether there was a real therapist involved and how involved they would be (11/47, 23%), whether they would be able to meet them in person, whether they would deliver the therapy live or preprogram the coach, or whether the virtual coach would be completely artificially intelligent. The realism of the virtual coach and the VRT (16/226, 7.1%) was also an important question posed by participants, including asking whether it would be realistic enough and comparing it to “real” or face-to-face therapy.
Most additional responses indicated that participants would like to know more about the process (67/226, 29.6%). This included practical questions regarding the frequency of use, the length of sessions, and how it would be delivered (eg, in which location). Furthermore, many participants asked about the content (26/226, 11.5%), particularly not only the scenarios that could be represented but also other aspects including the appearance, the script, and how the content could link with face-to-face therapy. Of the 226 participants, 11 (4.9%) asked about the technical aspects including the development of the coach (eg, whether an algorithm or artificial intelligence was used) and what equipment would be used to deliver the VRT.

**How Has Your Previous Experience of VR (Minimal or Extensive) Influenced Your Perceptions of VRT Guided by a Virtual Coach?**

The largest group of those answering this question indicated that they had no previous experience (79/236, 33.5%). A few without experience were positive or curious (19/236, 8.1%) whereas others (8/236, 3.1%) expressed more negative perceptions about its effectiveness as a therapeutic tool, anticipating that it would not feel real or tailored enough to the individual. The second largest group (59/236, 25%) felt that their previous experience had helped them to be more positive and linked it to their own experience of mental health and how it could be used for treatment. Although several mentioned using VR for gaming, they felt that it was effective at producing a level of presence that would be conducive to therapy and help to invoke real emotions and responses. They felt it was easy to use, could potentially lower costs, make therapy more accessible, and even with negative experiences, such as motion sickness or technical difficulties, they still had a positive perception of VRT.

However, 34 (14.4%) of the 236 participants with a more negative perception reported nausea or dizziness, whereas others perceived VR as more suited to games. This included problems with the quality of their experience, feeling that the VRT had not offered enough presence. However, those with a negative experience comprised the smallest group (8/236, 3.1%). Finally, the third largest group felt that their previous experience would not influence how they felt about VRTs guided by a virtual coach (40/236, 16.9%). For some, their previous experiences could not inform their perception of VRTs as it had been for entertainment purposes or they had too little experience to be able to make a judgment (13/236, 5.5%).

**If You Were Offered VRT Guided by a Virtual Coach, Where Do You Think the Best Place to Do the Therapy Would Be?**

The largest group of respondents who identified a single location felt that it would be best delivered within the home (84/245, 34.3%), whereas the second largest group felt that it would be best delivered in a more professional location (43/245, 17.6%). Several felt that it could be offered in both settings (38/245, 15.5%), whereas others suggested that access could first be through a clinic (12/245, 4.9%), where they could access technical or therapeutic support, or from home (4/245, 1.6%), where they would feel more comfortable. When respondents highlighted delivery from home, they described it as being safe, comfortable, and familiar. They felt that they might feel susceptible or disorientated when coming out of a VRT session and that being at home would be preferable. More professional locations, such as physician surgeries or clinics, were also described as safe and familiar, although by fewer people. Professional or clinical settings were often viewed as a better location because of the presence of support. Other reasons included the level of cleanliness offered and that there would be fewer distractions. Those without a preference identified elements of the location that were necessary to optimize the experience, including having a space to move, feeling safe and secure (eg, in an enclosed space), having privacy and quiet, and having few distractions. They also felt it would need to consider the condition being treated (including severity) and the individual’s preferences.

**Factors With Item Responses**

Table 3 presents those items where either strong agreement or disagreement elicited a qualitative response.
Table 3. Qualitative responses to items.

<table>
<thead>
<tr>
<th>Question</th>
<th>Strongly agree, n (%)</th>
<th>Strongly disagree, n (%)</th>
</tr>
</thead>
</table>

**Factor 1: attitudes toward VRT**  
(42 and 43)

<table>
<thead>
<tr>
<th>Item</th>
<th>Description</th>
<th>Strongly agree</th>
<th>Strongly disagree</th>
</tr>
</thead>
<tbody>
<tr>
<td>41</td>
<td>If the virtual coach encouraged me to do something between sessions, I would try to do it. (n=19)</td>
<td>18 (95)</td>
<td>1 (5)</td>
</tr>
<tr>
<td>51</td>
<td>I would never be willing to try virtual reality therapy. (n=63)</td>
<td>3 (5)</td>
<td>60 (95)</td>
</tr>
<tr>
<td>50</td>
<td>I would be willing to try virtual reality therapy if I had more information about it. (n=24)</td>
<td>21 (87)</td>
<td>3 (13)</td>
</tr>
<tr>
<td>52</td>
<td>I would encourage the people I care about to try virtual reality therapy, if it was offered to them. (n=16)</td>
<td>14 (87)</td>
<td>2 (13)</td>
</tr>
<tr>
<td>53</td>
<td>I would discourage the people I care about to try virtual reality therapy, if it was offered to them. (n=25)</td>
<td>1 (4)</td>
<td>24 (96)</td>
</tr>
<tr>
<td>54</td>
<td>I cannot imagine virtual reality therapy being useful for someone with mental health problems. (n=24)</td>
<td>2 (8)</td>
<td>22 (92)</td>
</tr>
<tr>
<td>42</td>
<td>I would feel comfortable interacting with the virtual coach. (n=14)</td>
<td>10 (71)</td>
<td>4 (29)</td>
</tr>
</tbody>
</table>

**Factor 2: expectation of presence**  
(47, 49, 50, 52, and 40)

<table>
<thead>
<tr>
<th>Item</th>
<th>Description</th>
<th>Strongly agree</th>
<th>Strongly disagree</th>
</tr>
</thead>
<tbody>
<tr>
<td>43</td>
<td>I would find the characters in the virtual reality therapy unsettling. (n=6)</td>
<td>1 (17)</td>
<td>5 (83)</td>
</tr>
<tr>
<td>45</td>
<td>I am skeptical about the effectiveness of virtual reality therapy. (n=13)</td>
<td>6 (46)</td>
<td>7 (54)</td>
</tr>
<tr>
<td>44</td>
<td>I think that the virtual reality therapy would make me feel present enough to be effective. (n=7)</td>
<td>3 (43)</td>
<td>4 (57)</td>
</tr>
</tbody>
</table>

**Factor 3: preference for VRT**  
(54, 39, 53, and 41)

<table>
<thead>
<tr>
<th>Item</th>
<th>Description</th>
<th>Strongly agree</th>
<th>Strongly disagree</th>
</tr>
</thead>
<tbody>
<tr>
<td>39</td>
<td>I think virtual reality therapy would be better than face-to-face therapy. (n=29)</td>
<td>2 (7)</td>
<td>27 (93)</td>
</tr>
<tr>
<td>40</td>
<td>I would trust a virtual coach the same amount as a real therapist. (n=20)</td>
<td>4 (20)</td>
<td>16 (80)</td>
</tr>
</tbody>
</table>

**Factor 4: cost-effectiveness**  
(46)

<table>
<thead>
<tr>
<th>Item</th>
<th>Description</th>
<th>Strongly agree</th>
<th>Strongly disagree</th>
</tr>
</thead>
<tbody>
<tr>
<td>46</td>
<td>I think virtual reality therapy will be worth the cost. (n=9)</td>
<td>6 (66)</td>
<td>3 (34)</td>
</tr>
</tbody>
</table>

**Nonfactor answers**  
(45, 48, 44, and 51)

<table>
<thead>
<tr>
<th>Item</th>
<th>Description</th>
<th>Strongly agree</th>
<th>Strongly disagree</th>
</tr>
</thead>
<tbody>
<tr>
<td>47</td>
<td>I think I would be able to use the virtual reality equipment easily. (n=26)</td>
<td>24 (92)</td>
<td>2 (8)</td>
</tr>
<tr>
<td>49</td>
<td>I think that virtual reality equipment could spread diseases. (n=19)</td>
<td>0 (0)</td>
<td>19 (100)</td>
</tr>
<tr>
<td>48</td>
<td>If my skills with technology were poor, I would feel confident using virtual reality therapy if the health care professional accompanying me was trained to a high standard. (n=19)</td>
<td>17 (89)</td>
<td>2 (11)</td>
</tr>
</tbody>
</table>

**Factor 1: Attitude Toward VRT**  
Individuals who scored highly on this factor had a positive attitude toward VRTs and VRTs delivered by a virtual coach, whereas those who scored low had a negative attitude. Within the items where strong agreement or disagreement elicited a text response (items 41, 51, 50, 52, 53, 54, and 42), those with positive attitudes highlighted the value of having a choice in mental health therapies. They emphasized the need to be willing to try different treatments to find the one that worked, reflecting on how more automated and digital options can help to increase access. Those with more negative attitudes indicated that it would be a type of therapy that they would not choose.

**Factor 2: Expectation of Presence**  
Individuals who scored highly on this factor felt that VR would not be real, that is, low presence. Individuals who scored low felt that VR was immersive. Within the items where strong agreement or disagreement elicited a text response (items 43, 45, and 44), the respondents indicated several factors that affected their expectation that VR would be “real enough.” Previous experiences appeared to be linked to the expectations of presence. People who enjoyed their experiences had higher expectations of presence. Those with lower expectations felt that VRT would be too much like a game, whereas others indicated that experiencing cybersickness meant they had not felt present.

**Factor 3: Preference for VRT**  
Individuals who scored highly on this factor showed a preference for VRT, whereas those who scored low showed a preference for face-to-face therapy. Within the items where strong agreement or disagreement elicited a text response (items 39 and 40), there was a strong sense that those who preferred face-to-face therapy would feel the loss of human interaction most and feel that a real person was needed to build a relationship and trust. Those who were more in favor of VRT
and the virtual coach felt that it would be more convenient and potentially enable more disclosures related to their mental health.

**Factor 4: Cost-Effectiveness**

Individuals who scored high on this factor felt that VRT was cost-effective, whereas those who scored low did not. Within the item where strong agreement or disagreement elicited a text response (item 46), those who felt it was cost-effective highlighted the decreasing costs of equipment and the benefits this could bring to mental health services. For those who felt that VR was still too expensive, there was also recognition of the difficulties that services might have in adopting VR.

**Discussion**

**Overview**

This study aimed to develop a new instrument for assessing the public perception of VRT delivered by a virtual coach. We received 295 responses. We found that a 4-factor solution was the best fit for the AVRT Scale, with all subscales having excellent internal consistency. The 4 factors were (1) attitudes toward VRT, (2) expectation of presence, (3) preference for VRT, and (4) cost-effectiveness. We found that being more familiar with VR was correlated with more positive attitudes toward VRT (factor 1), a higher expectation of presence (factor 2), preference for VRT over face-to-face therapy (factor 3), and belief that VRT is cost-effective (factor 4). Familiarity with mental health was not associated with any factor. The qualitative data supported the quantitative findings, with many respondents stating that their previous experience with VR may have affected their perception of VRT. Respondents identified their homes and spaces that felt safe and quiet as the best locations for delivering VRT. The virtual coach was a salient concept throughout the qualitative responses, with participants wanting to better understand it and the relationships that could be facilitated.

**Principal Findings**

Previous literature has indicated a correlation between VR familiarity and more positive attitudes toward VRTs [14]; however, this is the first study to demonstrate this through a public survey. Although we do not know the direction of this association, the qualitative findings suggest that as the reach of VR headsets increases, VRTs will likely be viewed more positively. The perceptions of potential patients are important in determining the efficacy of VRT, as positive expectations of any psychological therapy are associated with better treatment outcomes [35,36]. Therefore, an increase in the popularity of VR kits may indirectly improve the efficacy of VRTs.

This study is the first to explore people’s perceptions of VRTs guided by a virtual coach. Although most participants had no personal or professional experience of mental health therapy, many mentioned aspects relating to the virtual coach that draw parallels with “therapeutic alliance”; in psychotherapy, this denotes the importance of the relationship between the therapist and service user. In psychotherapy research, a strong therapeutic alliance is associated with better treatment outcomes [37]. The concept of therapeutic alliance has been studied more broadly in relation to VR-assisted therapies [30] and digital mental health [38]. The effects are similar but may be predictive of treatment outcomes to a lesser extent and more so predict engagement.

Understanding whether it is possible to foster a “therapeutic alliance” with a virtual coach and, if so, the nature of this relationship is something that the public is concerned with and therefore requires further investigation. Our findings provide initial insights into how therapeutic alliances may operate in VRT with a virtual coach. Respondents who showed a preference for VRTs indicated that the presence of a virtual coach would aid disclosure. This may be because of the anonymity that this form of communication offers [39]. Furthermore, it is notable that many were curious about the level of automation and formulation offered by the virtual coach. Qualitative findings from a trial of VRT guided by a virtual coach found that the presence of a member of staff helped to reinforce learning, which may suggest that certain elements of therapy require a certain level of formulation [40]. Our data suggest that the public view personalization as an important component of therapy and that VRTs can be improved by offering a certain level of formulation.

Another novelty of this study is the exploration of presence in relation to VRTs. Previous research has suggested that increasing presence can increase the effectiveness of VRT [41]. Newer VR-enabled headsets provide a greater level of presence as the quality of graphics and functionality, such as interactivity and sensitivity of sensors (eg, eye tracking), have improved significantly. Therefore, we sought to explore the importance of this sense of presence in the general population. Our findings indicate that those with a greater expectation of presence are more positive and more likely to show a preference for VRTs. Notably, our findings indicate that even those who are familiar with VR share concerns regarding the lack of presence and immersion in VRTs. This suggests that developers and researchers must continue to develop and update their intervention designs to ensure that VRTs do not become stagnant and continue to elicit a sense of presence.

Most of those who viewed VRTs positively emphasized the need for a choice to help increase access to mental health treatment. More automated VRTs have been designed considering the pressure to deliver psychological therapies in mental health services and the lack of resources to meet this need [10,11]. Our findings indicate that the public is aware of this and views VRTs guided by a virtual coach as an acceptable solution. Our respondents also indicated that VRT guided by a virtual coach would be suitable for delivery at home, further alleviating resource pressures. However, this was not the case for all participants, with a notable proportion wanting to access VRTs in a location that was safe, familiar, free of distractions, and large enough to use the VRT. The flexibility of location in delivering VRT is an important consideration for increasing access and meeting the needs of service users, especially when considering the strong links between poor mental health and housing quality [42]. The delivery location for VRT should be considered on a case-by-case basis.

A further consideration for the implementation of VRT found in our study is the importance of information. Our qualitative
findings indicate that people were keen to better understand what was involved in VRT and the virtual coach. Preintervention expectations are key in managing service users’ expectations while also fostering hopefulness, which in turn improves engagement [43]. This information can also help to allay any concerns service users might have about VRT and help developers to understand their needs to improve the design and implementation of VRTs. For example, a small number of participants opposed the use of VRTs, which when expanded in our qualitative data collection, indicated certain ethical and moral concerns about its use in mental health care. All these views were valid, but a few may be rooted in misconceptions about VR or expectations about how it will be implemented. Therefore, potentially increasing the public’s awareness and understanding of VR and VRT may help to appease them and improve how it is deployed.

We found mixed findings regarding the impact of cybersickness on willingness to engage in VRT. For some participants, cybersickness would dissuade people from using VRT. However, this finding was not ubiquitous, with some saying that they were still interested in trying VRT even if they had experienced cybersickness. This contradicts previous research, suggesting cybersickness is a considerable barrier [17]. As technology progresses, cybersickness might become less important. We also included a question on hygiene, as our questionnaire was shared during the COVID-19 pandemic. However, this was excluded, suggesting that it was not a significant concern for the public.

Limitations

The questionnaire has only been validated using the EFA. Further validation is required before we can confidently recommend its widespread use. Specifically, we must confirm the factor structure in a new sample using conﬁrmatory factor analysis and assess its concurrent and discriminant validity. If we are able to replicate the strong psychometric properties found in this study, this questionnaire can be used to understand attitudes toward VRTs delivered by virtual coaches. The scale will also need to be adapted to contexts outside the United Kingdom, for example, by amending items and further validation.

Most of our respondents were female and had no previous experience with mental health conditions or therapy. However, men and those with more experience with mental health conditions or therapy may have different perceptions of VRTs. A recent review found that gender differences might affect the use and acceptability of VR, specifically that women are more susceptible to cybersickness and therefore may be less willing to use VR [44]. On the basis of this, it may be assumed that if we were to conduct a survey with male participants, the attitudes toward VRT guided by a virtual coach could be even more positive. We do not have any available evidence to indicate whether those who are living with or have lived with mental health conditions are likely to be more or less accepting of VRTs. Purposive sampling should be used in future studies to ensure that the views of these groups are included in future validation studies.

Furthermore, we sought text responses for strong agreement or disagreement with certain items. Notably, those with more neutral views may have offered additional insights, but we weighed this against the additional burden on respondents. This may also have led some participants to neutralize their views to avoid triggering a free-text question. However, there were no instructions in the questionnaire that responding differently removed the free-text responses. In addition, the range of scores indicated that this did not deter the participants from giving extreme answers. The qualitative data we captured were sufficient for our analysis.

Finally, the analysis of the relationship between familiarity and attitudes toward VR and VRT was correlational. We could not make any claim regarding the direction or causal nature of these associations. For example, those with more positive attitudes and a better understanding are more likely to become familiar with VR through continued use. However, our qualitative findings indicate that negative experiences with VR do not factor in a willingness to use VRT.

Recommendations

Future research should further validate this questionnaire. Once this has been accomplished, the questionnaire could be used to investigate the factors that improve or worsen attitudes toward VRT and VRT guided by a virtual coach. For example, asking questions such as whether trying VR improves perceptions or whether increasing sales of domestic VR kits is associated with improved attitudes. It is also important to explore how these attitudes translate into behavior, that is, whether positive attitudes predict patient preferences and engagement with VRTs. The impact of the level of automation versus the formulation of the virtual coach on attitudes should also be explored, as this was a salient concept within the qualitative data. The AVRT Scale could be adapted and applied to other areas where VR is used to deliver interventions, such as behavior change interventions, or as a training tool. The questionnaire can also be used alongside treatment development, evaluation, and implementation to explore the barriers and facilitators specific to VRT and VRT guided by a virtual coach or the perceptions of certain populations to aid the translation of research into practice [45]. In the long term, any research that considers barriers to the uptake, engagement, and adoption of VRT has the potential to alleviate the demand for trained therapists in clinical settings, thus improving access to psychological therapies.

Acknowledgments

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. ADGB was funded by the National Institute for Health and Care Research MindTech MedTech Co-operative. The views expressed are those of the authors and not necessarily those of the National Health Service, the National Institute for Health and Care Research,
or the Department of Health. These organizations had no role in the study design, including collection, management, analysis, and interpretation of data, writing of the report, or the decision to submit the report for publication.

ADGB would like to acknowledge the support of the National Institute for Health and Care Research Nottingham Biomedical Research Centre.

Data Availability
All relevant data have been included in this publication. Researchers who would like to access the scale for further validation or adaptation may contact the corresponding author with a methodologically sound proposal.

Conflicts of Interest
None declared.

Multimedia Appendix 1
Description of virtual reality therapy and the virtual coach.
[PDF File (Adobe PDF File), 18 KB - mental_v11i1e48537_app1.pdf ]

Multimedia Appendix 2
Factor structure and loadings.
[PDF File (Adobe PDF File), 172 KB - mental_v11i1e48537_app2.pdf ]

References


21. JISC: online surveys. JISC. 2022. URL: https://www.jisc.ac.uk [accessed 2023-12-14]


Abbreviations

AVRT: Attitudes Towards Virtual Reality Therapy
EFA: exploratory factor analysis
VR: virtual reality
VRT: virtual reality therapy

©Aislinn D Gomez Bergin, Aoife M Allison, Cassie M Hazell. Originally published in JMIR Mental Health (https://mental.jmir.org/). 12.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Mental Health, is properly cited. The complete bibliographic information, a link to the original publication on https://mental.jmir.org/, as well as this copyright and license information must be included.
Personalized Virtual Reality Compared With Guided Imagery for Enhancing the Impact of Progressive Muscle Relaxation Training: Pilot Randomized Controlled Trial

Susanna Pardini1,2,3, MSc, PsyD; Silvia Gabrielli2, PhD; Silvia Olivetto1, MSc, PsyD; Francesca Fusina1,4, PhD, PsyD; Marco Dianti5, MSc; Stefano Forti5, MSc; Cristina Lancini1, MSc, ClinPsy; Caterina Novara1, PhD

1Department of General Psychology, University of Padova, Padova, Italy
2Digital Health Research, Centre for Digital Health and Wellbeing, Fondazione Bruno Kessler, Trento, Italy
3Human Inspired Technology Research Centre (HIT), University of Padova, Padova, Italy
4Padova Neuroscience Center, University of Padova, Padova, Italy
5Centre for Digital Health and Wellbeing, Fondazione Bruno Kessler, Trento, Italy

Corresponding Author:
Susanna Pardini, MSc, PsyD
Department of General Psychology
University of Padova
8 Venezia St
Padova
Italy
Phone: 39 3335944315
Email: susanna.pardini@phd.unipd.it

Abstract

Background: Empirical evidence has shown that virtual reality (VR) scenarios can increase the effects of relaxation techniques, reducing anxiety by enabling people to experience emotional conditions in more vivid settings.

Objective: This pilot randomized controlled study aims to investigate whether the progressive muscle relaxation technique (PMRT) associated with a personalized scenario in VR promotes psychological well-being and facilitates the recall of relaxing images more than the standard complementary intervention that involves the integration of PMRT and guided imagery (GI).

Methods: On the basis of a longitudinal, between-subject design, 72 university students were randomly exposed to one of two experimental conditions: (1) standard complementary procedure (PMRT and GI exposure) and (2) experimental procedure (PMRT and personalized VR exposure). Individuals were assessed by a therapist before and after 7 training sessions based on measures investigating anxiety, depression, quality of life, coping strategies, sense of presence, engagement, and side effects related to VR exposure. Heart rate data were also collected.

Results: Differences in changes between the 2 groups after the in vivo PMRT session conducted by the psychotherapist (T1) were statistically significant for state anxiety ($F_{1,67}=30.56; P<.001$) and heart rate ($F_{1,67}=4.87; P=.01$). Individuals in the VR group obtained lower scores both before ($t_{67}=-2.63; P=.01; \text{Cohen } d=0.91$) and after ($t_{67}=-7.23; P<.001; \text{Cohen } d=2.45$) the relaxation session when it was self-administered by participants (T2). A significant reduction in perceived state anxiety at T1 and T2 was observed for both groups ($P<.001$). After the VR experience, individuals reported feeling higher engagement in the experience than what was mentioned by participants in the GI group ($F_{1,67}=2.85; P=.03; \eta^2_p=0.15$), and they experienced the environment as more realistic ($F_{1,67}=4.38; P=.003; \eta^2_p=0.21$). No differences between groups regarding sense of presence were found ($F_{1,67}=1.99; P=.11; \eta^2_p=0.11$). Individuals exposed before to the VR scenario (T1) referred to perceiving the scenario recalled in-imagination at T2 as more realistic than what those in the GI group experienced ($F_{1,67}=3.21; P=.02; \eta^2_p=0.12$). The VR group had lower trait anxiety levels than the GI group after the relaxation session during session 7 (T2; $t_{67}=-2.43; P=.02$).

Conclusions: Personalized relaxing VR scenarios can contribute to improving relaxation and decreasing anxiety when integrated with PMRT as a complementary relaxation method.

Trial Registration: ClinicalTrials.gov NCT05478941; https://classic.clinicaltrials.gov/ct2/show/NCT05478941
Introduction

Background

In working adults and university students, the prevalence of depression, anxiety, and stress increases, predisposing them to physical diseases and general repercussions on well-being [1,2]. The situation has worsened owing to the COVID-19 outbreak, which affected individuals’ general well-being worldwide [3,4]. Different types of standardized relaxation interventions (eg, mindfulness-based stress reduction, progressive relaxation techniques by Jacobson, autogenic training by Schultz, abdominal relaxations, and visualizations) before, during, and after the COVID-19 pandemic have been shown to reduce anxiety and stress symptomatology in university students [4,5], with a more significant effect when integrated with other complementary techniques such as guided imagery (GI) [6,7]. GI is useful in creating mental imagery and refocusing attention on pleasant and relaxing imagined visual, auditory, tactile, or olfactory sensations, resulting in specific psychological and physiological responses such as relaxation and reduction of the autonomic nervous system responses [6,8,9].

The integration of the progressive muscle relaxation technique (PMRT) and GI can promote a higher sense of relaxation during training sessions, allowing for the application of the complementary techniques to cope with the stress experienced in daily activities [6,10-14].

Virtual reality (VR) is a usable, engaging, and user-friendly technology that promotes full immersion in the virtual context and facilitates the control of disturbing external stimuli [15,16].

Various studies have demonstrated that being immersed in virtual natural environments through a head-mounted display (HMD) facilitates the reduction of anxiety and stress symptoms in college students [17] as well as in association with different types of relaxation training (eg, body scan and muscle relaxation training) [18-21].

Owing to the positive impact of GI with PMRT in promoting relaxation, it may be hypothesized that exposure to a more vivid, closer-to-reality VR experience could be a helpful strategy for improving the relaxation learning promoted by the PMRT and decreasing stress and anxiety symptomatology.

GI offers an undefined number of possibilities for personalizing the imagined scenarios in a safe context. The personalization of content in VR should be helpful in recreating situations close to users’ needs to promote relaxation and can override the limits of GI. The personalization of VR environments is a crucial element to consider as it allows for relaxation and the perception of safety in the virtual context [18,22] owing to the possibility of offering a more realistic emotional experience, reflecting the users’ needs [23].

Furthermore, the first part of the PMRT, named “active PMRT,” involves a series of active sessions in which people learn how to tense and release the different muscle groups from the bottom to the top of the body to recognize the subjective state of muscle relaxation and relax the muscle areas of the body, reducing tension interfering with skeletal muscle activity.

These previous active PMRT phases are essential and are typically administered face-to-face by a psychotherapist or health operator trained to conduct the PMRT. By considering the effective results of web-based interventions in reducing stress among college students [24] and the relevant advantages that web-based therapy can offer (eg, saving costs related to attending psychotherapy sessions, allowing for partial self-management of the relaxation training), we assumed that assessing the efficacy of PMRT in alternative settings may facilitate the administration of treatment when the implementation of the standard procedure is not possible.

In light of the evidence described, we hypothesized that VR is more effective than in-imagination exposure in allowing for relaxation and decreasing state anxiety because of a more realistic sensory experience, facilitating visualization.

Objectives

For this reason, our primary aim was to deploy the active PMRT sessions remotely via Zoom (Zoom Video Communications) and the last complementary session in the therapist’s presence, exposing people to a passive progressive relaxation session with GI or VR.

More in detail, we investigated whether PMRT associated with a personalized relaxing scenario in VR can be more effective in reducing state anxiety, tension or activation, and heart rate.

The secondary purposes of this research pilot were to (1) understand whether VR promotes a better sense of presence and engagement in the scenario compared with GI after session 6 (T1) and whether it helps recall the image and be immersed in the relaxing scenario in session 7 (T2) and (2) investigate whether exposure to the VR scenario promotes a better perception of psychological well-being, stress, and trait anxiety symptoms.

Methods

Overview

This pilot study was a randomized, parallel-assignment, open-label, controlled, single-center trial based on a longitudinal, between-subject design conducted at the University of Padova (Italy) in collaboration with the Center for Digital Health and...
Wellbeing—Fondazione Bruno Kessler (Italy). This study is part of a larger research protocol published by Pardini et al [25] (International Registered Report Identifier: RR2-10.2196/44183). The study’s recruitment phase ended in February 2023.

Participants

Before study enrollment, during the T0 face-to-face assessment phase at the university, the participants signed a written informed consent form based on a paper-and-pencil form agreeing to participate in all the study sessions. They were informed that (1) their data would be confidential, (2) they could omit any information they did not wish to provide, and (3) they could withdraw from the study without explanation.

Study participants were recruited in Northeast Italy via social networking websites (on the web) and during university lectures (offline). Those individuals interested in participating were asked to attend a face-to-face assessment with the investigators to participate in the first evaluation phase (T0) to comply with the inclusion and exclusion criteria, provide written informed consent, and undergo a baseline evaluation.

Eligible participants were adults from the general population (aged ≥18 years) and native Italian speakers, owned a PC, and were able to use a PC and a smartphone.

Participants were excluded from the study if they had been diagnosed with a severe mental disorder or medical conditions (eg, neuromuscular disorders) or were undergoing current psychotherapeutic treatment.

Eligible participants were randomly allocated to one of the experimental conditions based on a simple blinded randomization via an Excel (Microsoft Corp) file using the “RAND” function. In total, 2 experimenters trained in cognitive and behavioral therapy conducted the relaxation sessions in a balanced way, each administering the sessions to 50% of the sample of each group to control for possible biases related to the therapist’s personality and competencies. During the allocation to one of the experimental conditions, participants were unaware of whether they had to undergo the trial in the VR or GI condition. Participants were then informed about the type of procedure. The random allocation sequence and participant enrollment were conducted by 2 experimenters trained in cognitive and behavioral therapy.

Intervention

Hardware and Software Equipment

A commercially available VR headset (Meta Quest 2; Reality Labs) with an Alienware m15 Ryzen Edition R5” workstation and a link cable were used. The virtual scenarios were developed using the Unity framework (Unity Technologies) and the C# programming language (source of assets: Freesound [26], Unity, Poly Haven, and HDRIs). The code was versioned via GitLab. More detailed information about the virtual environment design is provided in the study by Pardini et al [18].

Procedure

For each group, the intervention implied previous learning of the adapted abbreviated PMRT [27] in 4 web-based active PMRT sessions. Each session required approximately 25 minutes (Figure 1). To check whether participants completed the active PMRT sessions, the therapists conducting the experimental procedure could check whether participants completed the assessment phases and the active PMRT sessions accessing Moodle (Moodle HQ). Therapists sent a reminder to participants a day before each session in the form of a direct message from the Moodle platform that was received as a personal email on the webmail university platform.
The first session focused on tensing and releasing the hands, forearms, arms, neck, shoulders, and back muscles; the second session focused on the facial muscles; the third session focused on teaching diaphragmatic breathing; and the fourth session focused on the abdomen, buttocks, and lower limbs (Figure 1).

The web-based sessions took place on the Zoom platform and were conducted by a cognitive behavioral therapist. Each Zoom session consisted of (1) filling out the State-Trait Anxiety Inventory—Form Y1 (STAI-Y1; state anxiety form), (2) sharing general standardized instructions for relaxation and the PMRT background with the participant, and (3) active PMRT session focused on a particular body section.

The fifth VR session was conducted at the VR laboratory at the University of Padova. All participants were asked to wear the Xiaomi 2 smartwatch to detect their heart rate frequency before, during, and after the relaxation session. A total of 5 measurements (1 per minute) were taken before the relaxation experience, 12 were taken during the entire exposure, and 5 were taken after the experience in the virtual context. Both the PMRT with GI and the PMRT with VR sessions were approximately 12 minutes long. This duration was also established based on previous studies’ outcomes and to avoid potential cybersickness symptoms [28,29].

Specifically, the compared groups’ conditions were characterized as follows: (1) the **PMRT and GI condition** consisted of the deployment of a standard behavioral intervention based on 4 individual active PMRT sessions via Zoom (sessions 2-5; T1-T4), an in vivo PMRT relaxation session with GI conducted by the psychotherapist (session 6; T1) a week after the baseline assessment (session 1; T0), and a follow-up phase (session 7; T2) after 2 weeks consisting of recovering the in-imagination relaxing scenario and a PMRT session (Figure 2); and (2) the **PMRT and VR condition** consisted of 4 active PMRT sessions administered via Zoom (sessions 2-5; T1-T4), a passive PMRT session integrated with the exposure to personalized VR scenarios deployed using the Meta Quest 2 HMD (session 6; T1), and a follow-up phase consisting of the same activities as the PMRT and GI condition (session 7; T2; Figure 3).
Figure 2. Progressive muscle relaxation technique (PMRT) and guided imagery (GI) exposure.

Figure 3. Progressive muscle relaxation technique (PMRT) and personalized virtual reality (VR) exposure.
The assessment phases were completed via the Moodle open-source learning platform. For this purpose, students could use their institutional accounts to access the platform. The administrator of the experimental procedure then created an ID profile for each participant. To access the Moodle platform, participants did not have to pay.

The assessment phases took place at baseline (T0); before and after each of the 4 active PMRT sessions; before and after the passive PMRT session administered with VR or GI (T1); and a week later, when participants were asked to lead the relaxation session autonomously (T2).

The T0 (baseline) assessment phase lasted approximately 40 minutes; it was the same for all participants and was administered at the VR laboratory at the University of Padova. It involved the administration of the following measures: (1) a demographic schedule [30]; (2) a series of self-report questionnaires investigating depression, anxiety, stress, quality of life, and distress coping strategies (State-Trait Anxiety Inventory–Form Y [STAI-Y]; Depression, Anxiety, and Stress Scale–21 [DASS-21]; Psychological General Well-Being Index [PGWBI]; and Coping Orientation to the Problems Experienced–Nuova Versione Italiana [COPE-NVI]); and (3) resting heart rate detection using the Xiaomi Mi Band 2.

Before and after each relaxation session (T1-T4), in approximately 20 minutes, the personal level of tension and relaxation was assessed using a visual analog scale (VAS) from 0 (no tension) to 10 (extreme tension level). The state anxiety level was evaluated based on the STAI-Y1. The 4 relaxation sessions were administered 2 to 3 days apart for all 3 groups. The assessment phase was web-based administered through Moodle, an e-learning platform used for data collection.

The T1 phase (day 7) took approximately 60 minutes and took place at the VR laboratory at the University of Padova. Before and after the relaxation session, states of tension and relaxation and anxiety were assessed using a VAS from 0 (no tension) to 10 (extreme tension level). The state anxiety level was evaluated based on the STAI-Y1. Participants were then exposed to a PMRT session merged with a VR or GI procedure. Before the in-imagination or VR experience, all the participants filled out the Vividness of Visual Imagery Questionnaire and the Test of Visual Imagery Control. After the PMRT session, users filled out a series of self-report questionnaires investigating depression, anxiety, and stress (STAI-Y and DASS-21), and only the VR group filled out the Virtual Reality Symptom Questionnaire (VRSQ) to monitor VR-related side effects (eg, sickness) and the International Test Commission–Sense of Presence Inventory (ITC-SOPI) to assess the sense of presence at the end of the T1 phase. The Xiaomi Mi Band 2 was used during the entire T1 phase to detect resting heart rate activity. The assessment phase was administered through the Moodle e-learning platform.

The T2 phase was deployed for approximately 45 minutes at the VR laboratory at the University of Padova. Before and after the relaxation session, states of tension and relaxation and anxiety were assessed using a VAS from 0 (no tension) to 10 (extreme tension level). The state anxiety level was evaluated based on the STAI-Y1. All users were exposed to a self-GI experience in which those who were part of the VR group were asked to recall the personalized VR scenario experienced during the T1 phase (day 7). The GI group retrieved instead the image that participants had used in association with the PMRT during the T1 phase (day 7). After the session, participants filled out a series of self-report questionnaires investigating depression, anxiety, stress, and quality of life (STAI-Y, DASS-21, and PGWBI) and an ad hoc version based on the ITC-SOPI to assess the sense of presence experienced during the imagery session. This assessment phase was administered based on the Moodle e-learning platform. The Mi Band 2 was used during the entire T1 phase administration (day 7) to detect resting heart rate activity.

Sample Size Estimation

To investigate whether the parameters based on our sample size could be acceptable, a formal sample size calculation was conducted using the G*Power (version 3.1) software [31]. As a statistical test, the repeated-measure ANOVA between the factors was considered. The effect size was 0.28, the Cronbach α was .05, and the power (1 − β error probability) was 0.80. We had 2 separate groups and 3 measurements. On the basis of these parameters, it was estimated that at least 35 participants should be recruited for each group. The enrollment process is summarized in Figure 4.
Figure 4. CONSORT flow diagram. Progressive muscle relaxation technique (PMRT) and guided imagery (GI) condition: “As usual” intervention; PMRT and virtual reality (VR) condition: “Experimental” intervention.

Statistical Analysis
Quantitative statistical analyses were conducted using the SPSS (version 29.0; IBM Corp) software [32]. Frequencies, means, and SDs were measured to explore the sociodemographic features.

To examine the between- and within-subject differences, the 2 groups were evaluated using a repeated-measure multivariate analysis of covariance (MANCOVA) for mixed designs, multivariate MANCOVA, and $t$ tests. To compare the differences between and within groups, pairwise comparisons with Bonferroni CI adjustment were calculated. Multiple linear regression analyses were conducted to investigate whether the sense of presence, engagement, and perception that the scenario was realistic could be predictive of the state anxiety level experienced after the VR exposure.

Outcome Measures
The measures were administered before and after each relaxation session. The heart rate frequency was recorded based on the Xiaomi Mi Band 2 before the session to obtain baseline data during and after the trials.

A sociodemographic schedule was filled out to obtain information about gender; age; mother tongue; marital status; years of education; occupation; psychological, medical, and neuromuscular problems; use of drugs; and whether participants had relaxation training experience or had used VR devices in the past.

The STAI-Y [33-35] is a self-report questionnaire that allows for the investigation of state and trait anxiety using 40 items on a 4-point Likert scale. Items are grouped into 2 scales focused on how participants generally feel (trait anxiety) or what they experience at particular times (state anxiety). The reliability and validity of the STAI-Y are good in the Italian sample. This study highlighted a moderate to acceptable internal consistency for the state and trait anxiety scales administered ($0.70 < \text{Cronbach } \alpha < 0.84$).

The VAS for the tension and relaxation level is a measure of tension and relaxation intensity administered before and after each VR experience. Participants had to express how tense and activated they felt (0=not at all; 10=completely). Lower scores indicated higher relaxation levels.
The DASS-21 [36-38] is a self-report questionnaire based on 21 items that provides information about anxiety, depression, and stress symptomatology on a 4-point Likert scale from 0 to 3. Internal consistency and convergent, divergent, and criterion-oriented validity are adequate in the original and Italian versions. On the basis of our sample, a moderate to acceptable internal consistency emerged for the total and the 3 subscales (.77<\text{Cronbach } \alpha<.87).

The PGWBI [39-41] is a self-report questionnaire consisting of 22 items that provides a general subjective assessment of psychological well-being. It comprises 6 subscales: anxiety, depression, positivity and well-being, self-control, general health, and vitality. The scores for all subscales can be summarized to provide a summary score, which reaches a maximum of 110 points representing the best achievable “well-being.” The tool’s psychometric properties are good for the original version and Italian validation. Considering our sample, an acceptable internal consistency emerged for the total score (Cronbach \( \alpha=.72 \)).

The COPE-NVI [42] is a 60-item self-report questionnaire on a 5-point Likert scale that investigates how often people use certain coping strategies with stressful or difficult events. Items are grouped into 5 subscales referring to different coping strategies: social support, avoidance strategies, positive attitude, problem-solving, and turning to religion. This tool is psychometrically valid to measure coping styles in the Italian context. Considering our sample, an acceptable internal consistency emerged for the total and the 6 subscales (.72<\text{Cronbach } \alpha<.89).

The Vividness of Visual Imagery Questionnaire [43-46] comprises 16 items on a 5-point Likert scale and investigates individual differences regarding the ability to imagine visual contexts vividly. The participant is asked to generate 4 mental images and evaluate their vividness. On the basis of our sample, a good internal consistency was observed (Cronbach \( \alpha=.89 \)).

The Test of Visual Imagery Control [45-47] is a measure that evaluates individual differences in the ability to control and modify mental images intentionally. For example, participants are asked to visualize a car and then transform the image according to 10 different descriptions. Responses are recorded. On the basis of our sample, a good internal consistency was found (Cronbach \( \alpha=.86 \)).

The ITC-SOPI [48,49] is a questionnaire consisting of 42 items on a 5-point Likert scale that allows for the investigation of the sense of presence experienced in a VR context. It comprises 4 subscales investigating the sense of physical space, level of engagement experience in the virtual context, ecological validity, and negative effects of exposure. On the basis of our sample, a good internal consistency was found for all subscales (.67<\text{Cronbach } \alpha<.78).

The VRSQ [50,51] assesses the general and eye-related physical symptoms of exposure to a VR environment. The score assigned to each item ranges from 0 to 6, with a maximum total score of 84 (48 for general symptoms and 36 for eye symptoms). Higher scores represent worse symptoms, with 0 corresponding to no adverse effects and 84 to serious adverse effects.

**Ethical Considerations**

The proposed study protocol was approved by the institutional review board of the Interdepartmental Ethical Committee of Psychology (17 Area) of the University of Padova (Italy; approval 4701; April 29, 2022). The study complied with the relevant ethical regulations of the Declaration of Helsinki (Italian law 196/2003; European Union General Data Protection Regulation 679/2016).

**Results**

**Sociodemographic Features and Comparisons**

The sociodemographic features of the study sample are outlined in Table 1.
A multivariate ANOVA was conducted to investigate whether the groups differed in sociodemographic and psychological characteristics before the relaxation training. A difference between the groups was found only for age, years of education, and marital status (Table 1).

To control for the possible impact that baseline differences could have on the research outcomes, the effects of all these variables were controlled for in the subsequent analyses. The psychological problems referred to by participants were related to problematic relationships with parents, low self-esteem, and relational problems. Moreover, for individuals in the VR group, the VRSQ was administered to assess the possible collateral effects of VR exposure–related nausea. On average, individuals showed light susceptibility to general and eye-related motion.

### Table 1. Demographic features and comparisons considering psychological constructs.

<table>
<thead>
<tr>
<th></th>
<th>Virtual reality group (n=36)</th>
<th>Guided imagery group (n=36)</th>
<th>F test (df)</th>
<th>Chi-square (df)</th>
<th>η²p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender (women), n (%)</td>
<td>28 (78)</td>
<td>27 (75)</td>
<td>N/A³</td>
<td>0.8 (2)</td>
<td>N/A</td>
</tr>
<tr>
<td>Age (years), mean (SD)</td>
<td>23.83 (6.10)</td>
<td>30.42 (8.36)</td>
<td>14.56 (1, 69)</td>
<td>N/A</td>
<td>0.17</td>
</tr>
<tr>
<td>Years of education, mean (SD)</td>
<td>16.81 (1.31)</td>
<td>18.25 (2.60)</td>
<td>8.87 (1, 69)</td>
<td>N/A</td>
<td>0.11</td>
</tr>
<tr>
<td>Marital status, n (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Single</td>
<td>18 (50)</td>
<td>16 (44)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Engaged, noncohabiting</td>
<td>17 (47)</td>
<td>10 (28)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Married and cohabiting</td>
<td>1 (3)</td>
<td>10 (28)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medication (yes), n (%)</td>
<td>8 (22)</td>
<td>5 (14)</td>
<td></td>
<td>0.5 (2)</td>
<td>N/A</td>
</tr>
<tr>
<td>Psychological problems (yes), n (%)</td>
<td>15 (42)</td>
<td>12 (33)</td>
<td></td>
<td>0.5 (2)</td>
<td>N/A</td>
</tr>
<tr>
<td>Medical problems (yes), n (%)</td>
<td>6 (17)</td>
<td>10 (28)</td>
<td></td>
<td>1.3 (2)</td>
<td>N/A</td>
</tr>
<tr>
<td>Experienced relaxation protocol in the past (yes), n (%)</td>
<td>8 (22)</td>
<td>3 (8)</td>
<td>N/A</td>
<td>2.7 (2)</td>
<td>N/A</td>
</tr>
<tr>
<td>Experience with VR⁵ in the past (yes), n (%)</td>
<td>10 (28)</td>
<td>8 (22)</td>
<td>N/A</td>
<td>0.3 (2)</td>
<td>N/A</td>
</tr>
<tr>
<td>STAI-Y2⁶ (total), mean (SD)</td>
<td>48.06 (3.78)</td>
<td>47.56 (3.56)</td>
<td>0.96 (1, 69)</td>
<td>N/A</td>
<td>0.05</td>
</tr>
<tr>
<td>DASS-21⁷ (anxiety), mean (SD)</td>
<td>3.61 (1.87)</td>
<td>3.69 (1.95)</td>
<td>0.03 (1, 69)</td>
<td>N/A</td>
<td>0.001</td>
</tr>
<tr>
<td>DASS-21 (depression), mean (SD)</td>
<td>4.46 (4.09)</td>
<td>5.47 (3.70)</td>
<td>1.18 (1, 69)</td>
<td>N/A</td>
<td>0.02</td>
</tr>
<tr>
<td>DASS-21 (stress), mean (SD)</td>
<td>8.56 (3.75)</td>
<td>7.69 (3.00)</td>
<td>1.16 (1, 69)</td>
<td>N/A</td>
<td>0.02</td>
</tr>
<tr>
<td>COPE-NVI⁸ (social support), mean (SD)</td>
<td>34.92 (6.36)</td>
<td>32.11 (6.47)</td>
<td>3.44 (1, 69)</td>
<td>N/A</td>
<td>0.05</td>
</tr>
<tr>
<td>COPE-NVI (avoidance), mean (SD)</td>
<td>22.83 (4.91)</td>
<td>23.00 (4.04)</td>
<td>0.03 (1, 69)</td>
<td>N/A</td>
<td>0.001</td>
</tr>
<tr>
<td>COPE-NVI (positive attitude), mean (SD)</td>
<td>30.78 (5.03)</td>
<td>30.72 (5.26)</td>
<td>0.002 (1, 69)</td>
<td>N/A</td>
<td>0.001</td>
</tr>
<tr>
<td>COPE-NVI (problem-solving orientation), mean (SD)</td>
<td>31.97 (5.03)</td>
<td>31.69 (5.27)</td>
<td>0.05 (1, 69)</td>
<td>N/A</td>
<td>0.001</td>
</tr>
<tr>
<td>COPE-NVI (transcendental orientation), mean (SD)</td>
<td>17.44 (3.49)</td>
<td>17.31 (3.47)</td>
<td>0.03 (1, 69)</td>
<td>N/A</td>
<td>0.001</td>
</tr>
<tr>
<td>PGWBI-22⁹ (total), mean (SD)</td>
<td>60.86 (6.97)</td>
<td>63.52 (9.48)</td>
<td>1.85 (1, 69)</td>
<td>N/A</td>
<td>0.03</td>
</tr>
<tr>
<td>VVIQ¹⁰ (total), mean (SD)</td>
<td>58.53 (9.37)</td>
<td>58.39 (11.21)</td>
<td>0.003 (1, 69)</td>
<td>N/A</td>
<td>0.001</td>
</tr>
<tr>
<td>TVIC¹¹ (total), mean (SD)</td>
<td>41.56 (5.58)</td>
<td>43.44 (5.32)</td>
<td>2.16 (1, 69)</td>
<td>N/A</td>
<td>0.03</td>
</tr>
</tbody>
</table>

aN/A: not applicable.
bP<.001.
cP<.01.
dP<.05.
eVR: virtual reality.
fSTAI-Y2: State-Trait Anxiety Inventory–Form Y2.
gDASS-21: Depression, Anxiety, and Stress Scale–21.
hCOPE-NVI: Coping Orientation to the Problems Experienced–Nuova Versione Italiana.
jVVIQ: Vividness of Visual Imagery Questionnaire.
kTVIC: Test of Visual Imagery Control.
sickness levels (VRSQ general score: mean 1.39, SD 1.32, range 0-4; VRSQ eye symptom score: mean 1.56, SD 1.32, range 0-5).

**Objective 1: Investigate Whether PMRT Associated With a Personalized Relaxing Scenario in VR Can Be More Effective in Reducing State Anxiety, Tension and Activation, and Heart Rate Frequency**

To investigate the differences between the “Virtual Reality” and “Guided Imagery” groups before and after the complementary relaxation experience in session 6 (T1) and session 7 (T2; Figures 2 and 3), we applied a repeated-measure MANCOVA. No significant within-subject main effect emerged ($F_{2,67}=1.01; P=.37$). A significant between-subject main effect was found ($F_{1,67}=30.56; P<.001$), and the VR group showed a greater change after the VR exposure than the GI group did in state anxiety (STAI-Y1: $F_{1,67}=10.27; P<.001$). After the relaxation experience at T1 ($t_{67}=-7.82; P<.001$), the VR group displayed lower state anxiety levels than the GI group both before ($t_{67}=-2.63; P=.01; Cohen d=0.91$) and after ($t_{67}=-7.23; P<.001; Cohen d=2.45$) the relaxation session during session 7 (T2; Table 2). Pairwise comparisons between and within the groups were performed before and after the relaxation experience at T1 and T2 (see also Multimedia Appendix 1).

After the relaxation sessions at T1 and T2, the VR and GI groups experienced a significant decrease in state anxiety ($P<.001$). Pairwise comparisons within groups highlighted a significant reduction in perceived state anxiety at T1 and T2 for both the VR and GI groups (Table 3).

**Table 2.** Descriptive analysis for group differences.

<table>
<thead>
<tr>
<th>Dependent variable, time, and group type</th>
<th>Values, mean (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>STAI-Y1</strong> (T1)</td>
<td></td>
</tr>
<tr>
<td>Before $^b$</td>
<td></td>
</tr>
<tr>
<td>Virtual reality (n=36)</td>
<td>48.06 (3.52)</td>
</tr>
<tr>
<td>Guided imagery (n=36)</td>
<td>50.16 (5.16)</td>
</tr>
<tr>
<td>After $^c$</td>
<td></td>
</tr>
<tr>
<td>Virtual reality (n=36)</td>
<td>31.17 (4.54)</td>
</tr>
<tr>
<td>Guided imagery (n=36)</td>
<td>40.00 (4.28)</td>
</tr>
<tr>
<td><strong>STAI-Y1</strong> (T2)</td>
<td></td>
</tr>
<tr>
<td>Before</td>
<td></td>
</tr>
<tr>
<td>Virtual reality (n=36)</td>
<td>48.89 (3.04)</td>
</tr>
<tr>
<td>Guided imagery (n=36)</td>
<td>51.78 (3.99)</td>
</tr>
<tr>
<td>After</td>
<td></td>
</tr>
<tr>
<td>Virtual reality (n=36)</td>
<td>32.39 (4.33)</td>
</tr>
<tr>
<td>Guided imagery (n=36)</td>
<td>40.75 (4.74)</td>
</tr>
</tbody>
</table>

$^a$STAI-Y1: State-Trait Anxiety Inventory–Form Y1.

$^b$Refers to the assessment filled out before the relaxation experience at T1 and T2.

$^c$Refers to the assessment filled out after the relaxation experience at T1 and T2.
### Table 3. Within-group pairwise comparisons (Bonferroni CI adjustment).

<table>
<thead>
<tr>
<th>Dependent variable, group type, and time</th>
<th>Mean difference (before – after)</th>
<th>SE</th>
<th>t test (df)</th>
<th>Cohen d</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>STAI-Y1</strong> <em>(T1)</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Virtual reality <em>(n=36)</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Before* and after</td>
<td>17.13d</td>
<td>1.07</td>
<td>16.01 (67)</td>
<td>4.15</td>
</tr>
<tr>
<td>Guided imagery <em>(n=36)</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Before and after</td>
<td>9.93d</td>
<td>1.07</td>
<td>9.30 (67)</td>
<td>2.14</td>
</tr>
<tr>
<td><strong>STAI-Y1</strong> <em>(T2)</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Virtual reality <em>(n=36)</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Before and after</td>
<td>16.58d</td>
<td>1.10</td>
<td>16.40 (67)</td>
<td>4.41</td>
</tr>
<tr>
<td>Guided imagery <em>(n=36)</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Before and after</td>
<td>10.69d</td>
<td>1.10</td>
<td>10.33 (67)</td>
<td>2.52</td>
</tr>
</tbody>
</table>

*a*STAI-Y1: State-Trait Anxiety Inventory–Form Y1.

*b*Refers to the assessment filled out before the relaxation experience at T1 and T2.

*c*Refers to the assessment filled out after the relaxation experience at T1 and T2.

*d*P<.001.

Regarding the outcomes that emerged from the VAS, no significant differences were observed before the relaxation session at T1 (VR group: mean 3.71, SD 0.37; GI group: mean 4.13, SD 0.37; t<sub>67</sub>=-0.76; P=.44; η<sup>p</sup><sup>2</sup>=0.009), but the groups differed after the T1 session, with higher levels of tension experienced by participants in the GI group (VR group: mean 1.05, SD 0.30; GI group: mean 2.06, SD 0.30; t<sub>67</sub>=−2.23; P=.03; η<sup>p</sup><sup>2</sup>=0.069). The same pattern was observed at T2, with no differences between the groups before T2 (VR group: mean 3.52, SD 0.35; GI group: mean 4.17, SD 0.35; t<sub>67</sub>=−1.24; P=.22; η<sup>p</sup><sup>2</sup>=0.022) and higher scores in tension obtained by the individuals in the GI group after the relaxation session (VR group: mean 2.06, SD 1.71; GI group: mean 2.58, SD 1.80; t<sub>67</sub>=−2.06; P=.04; η<sup>p</sup><sup>2</sup>=0.060). Moreover, the levels of tension decreased in each group after the relaxation session at T1 (VR group: t<sub>67</sub>=5.39, P<.001, and Cohen d=5.05; GI group: t<sub>67</sub>=4.45, P<.001, and Cohen d=7.20) and T2 (VR group: t<sub>67</sub>=8.55, P<.001, Cohen d=7.90; GI group: t<sub>67</sub>=6.68, P<.001, and Cohen d=6.15).

Heart rate was assessed 3 times (before, during, and after the relaxation exposure). The Mauchly test of sphericity was adequate for the heart rate recorded in both sessions 6 and 7 (P=.09). Heart rate frequency was found to have a significant main effect between groups (F<sub>1,67</sub>=4.87; P=.01) but to be not significant within groups (F<sub>2,134</sub>=0.51; P=.73). An interaction between the “time of assessment” and “group” factors for heart rate (F<sub>2,134</sub>=2.22; P=.01) was observed. Specifically, individuals in the VR group recorded lower levels of heart rate during (mean 68.26, SD 5.49) and after (mean 69.11, SD 4.98) the VR exposure at T1 than those of individuals in the GI condition (heart rate during: mean 72.86, SD 5.23; heart rate after: mean 72.22, SD 4.73; t<sub>67</sub>=−2.79; P<.01). No significant outcomes were observed at T2.

**Objective 2: Understand Whether VR Promotes a Better Sense of Presence and Engagement in the Scenario Compared With GI After Session 6 (T1) and Whether it Helps Recall the Image and Be Immersed in the Relaxing Scenario in Session 7 (T2)**

Individuals in the VR group reported feeling higher engagement after the experience than participants in the GI group (ITC-SOPI–Engagement: F<sub>1,67</sub>=2.85; P=.03; η<sup>p</sup><sup>2</sup>=0.15). Moreover, the VR group participants reported experiencing a more realistic environment than that experienced by individuals who used imagination to create the scenario (ITC-SOPI–Ecological Validity: F<sub>1,67</sub>=4.38; P=.003; η<sup>p</sup><sup>2</sup>=0.21). No differences between the groups regarding sense of presence were found (ITC-SOPI–Sense of Presence: F<sub>1,67</sub>=1.99; P=.11; η<sup>p</sup><sup>2</sup>=0.11; Table 4).
To investigate whether VR facilitated image recall during session 7 more than the GI technique did, the sense of presence and engagement experiences were assessed after the relaxation experience in session 7 (T2). Individuals previously exposed to the personalized VR scenario referred to perceiving the same scenario recalled in T2 as more realistic than did individuals that, in the previous session (T1), were exposed to an imagined scenario (ITC-SOPI–Ecological Validity: $F_{1,67} = 3.21; P = .02; \eta_p^2 = 0.12$). No differences emerged between the groups at session 7 (T2) for sense of presence (ITC-SOPI–Sense of Presence: $F_{1,67} = 0.76; P = .55; \eta_p^2 = 0.04$) and engagement (ITC-SOPI–Engagement: $F_{1,67} = 2.30; P = .07; \eta_p^2 = 0.12$; Table 4).

Objective 3: Investigate Group Differences Regarding Trait Anxiety, Depressive Symptoms, Stress, Coping, and Well-Being

To investigate the differences in trait anxiety, depressive symptoms, and stress between the VR and GI groups at baseline (T0), after session 6 (T1), and after session 7 (T2), we applied repeated-measure MANCOVA. The Mauchly test of sphericity was adequate only for the State-Trait Anxiety Inventory–Form Y2 ($P > .05$). The Huynh-Feldt correction was adopted for the DASS-21 anxiety subscale. The Greenhouse-Geisser correction was adopted for the DASS-21 depression and stress subscales. A significant interaction between the variables “time of assessment” and “group” factors ($F_{1,140} = 3.62; P = .02$) emerged.

The “Virtual reality” group had lower trait anxiety levels than the “Guided Imagery” group after the relaxation session during session 7 (T2; $t_{67} = -2.43; P = .02$; Tables 5-7). No significant interactions between “time of assessment” and “group” were found for stress ($F_{1.16,114} = 3.15; P > .05$), anxiety ($F_{1.89,132.58} = 1.11; P > .05$), and depressive symptoms ($F_{1.16,123.23} = 1.09; P > .05$) assessed using the DASS-21.

Table 4. Multivariate analysis of covariance.

<table>
<thead>
<tr>
<th>Dependent variable and group type</th>
<th>Values, mean (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ITC-SOPI–Sense of Presence (T1)</strong></td>
<td></td>
</tr>
<tr>
<td>Virtual reality (n=36)</td>
<td>60.06 (8.64)</td>
</tr>
<tr>
<td>Guided imagery (n=36)</td>
<td>57.00 (8.99)</td>
</tr>
<tr>
<td><strong>ITC-SOPI–Engagement (T1)</strong></td>
<td></td>
</tr>
<tr>
<td>Virtual reality (n=36)</td>
<td>46.61 (4.66)</td>
</tr>
<tr>
<td>Guided imagery (n=36)</td>
<td>42.42 (5.71)</td>
</tr>
<tr>
<td><strong>ITC-SOPI–Ecological Validity (T1)</strong></td>
<td></td>
</tr>
<tr>
<td>Virtual reality (n=36)</td>
<td>19.92 (2.12)</td>
</tr>
<tr>
<td>Guided imagery (n=36)</td>
<td>17.67 (2.75)</td>
</tr>
<tr>
<td><strong>ITC-SOPI–Sense of Presence (T2)</strong></td>
<td></td>
</tr>
<tr>
<td>Virtual reality (n=36)</td>
<td>52.81 (12.56)</td>
</tr>
<tr>
<td>Guided imagery (n=36)</td>
<td>52.69 (11.55)</td>
</tr>
<tr>
<td><strong>ITC-SOPI–Engagement (T2)</strong></td>
<td></td>
</tr>
<tr>
<td>Virtual reality (n=36)</td>
<td>40.40 (6.88)</td>
</tr>
<tr>
<td>Guided imagery (n=36)</td>
<td>40.36 (6.16)</td>
</tr>
<tr>
<td><strong>ITC-SOPI–Ecological Validity (T2)</strong></td>
<td></td>
</tr>
<tr>
<td>Virtual reality (n=36)</td>
<td>17.44 (3.85)</td>
</tr>
<tr>
<td>Guided imagery (n=36)</td>
<td>14.17 (4.06)</td>
</tr>
</tbody>
</table>

*aITC-SOPI: International Test Commission–Sense of Presence Inventory.
### Table 5. Descriptive analysis of the State-Trait Anxiety Inventory–Form Y2 (STAI-Y2) at T1 and T2.

<table>
<thead>
<tr>
<th>Dependent variable, time, and group type</th>
<th>Mean (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>STAI-Y2</strong></td>
<td></td>
</tr>
<tr>
<td>T1</td>
<td></td>
</tr>
<tr>
<td>Virtual reality (n=36)</td>
<td>42.75 (3.07)</td>
</tr>
<tr>
<td>Guided imagery (n=36)</td>
<td>44.11 (4.58)</td>
</tr>
<tr>
<td>T2</td>
<td></td>
</tr>
<tr>
<td>Virtual reality (n=36)</td>
<td>44.25 (4.07)</td>
</tr>
<tr>
<td>Guided imagery (n=36)</td>
<td>47.36 (4.29)</td>
</tr>
</tbody>
</table>

### Table 6. Between-group pairwise comparisons for the State-Trait Anxiety Inventory–Form Y2 (STAI-Y2; Bonferroni CI adjustment).

<table>
<thead>
<tr>
<th>Group type</th>
<th>Mean difference (before – after)</th>
<th>SE</th>
<th>t test (df)</th>
<th>Cohen d</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>STAI-Y2 (T0)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Virtual reality (n=36)</td>
<td>0.78</td>
<td>0.98</td>
<td>0.80 (67)</td>
<td>0.14</td>
</tr>
<tr>
<td>Guided imagery (n=36)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>STAI-Y2 (T1)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Virtual reality (n=36)</td>
<td>−0.17</td>
<td>0.99</td>
<td>−2.18 (67)</td>
<td>−0.35</td>
</tr>
<tr>
<td>Guided imagery (n=36)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>STAI-Y2 (T2)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Virtual reality (n=36)</td>
<td>−2.76a</td>
<td>1.13</td>
<td>−2.44 (67)</td>
<td>−0.74</td>
</tr>
<tr>
<td>Guided imagery (n=36)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*P<.05.

### Table 7. Within-group pairwise comparisons for the State-Trait Anxiety Inventory–Form Y2 (STAI-Y2; Bonferroni CI adjustment).

<table>
<thead>
<tr>
<th>Group type and time</th>
<th>Mean difference (before – after)</th>
<th>SE</th>
<th>t test (df)</th>
<th>Cohen d</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Virtual reality (n=36)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T0-T1</td>
<td>4.83 a,b</td>
<td>0.94</td>
<td>5.14 (67)</td>
<td>1.54</td>
</tr>
<tr>
<td>T0-T2</td>
<td>3.77 a</td>
<td>0.84</td>
<td>4.49 (67)</td>
<td>0.97</td>
</tr>
<tr>
<td>T1-T2</td>
<td>−1.06</td>
<td>0.96</td>
<td>−1.10 (67)</td>
<td>−0.42</td>
</tr>
<tr>
<td><strong>Guided imagery (n=36)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T0-T1</td>
<td>3.90 a</td>
<td>0.95</td>
<td>4.41 (67)</td>
<td>0.84</td>
</tr>
<tr>
<td>T0-T2</td>
<td>0.23</td>
<td>0.84</td>
<td>0.27 (67)</td>
<td>0.05</td>
</tr>
<tr>
<td>T1-T2</td>
<td>−3.67</td>
<td>0.93</td>
<td>−3.95 (67)</td>
<td>−0.73</td>
</tr>
</tbody>
</table>

*aP<.001.

bItalicization indicates that the P value is statistically significant.

Differences within and between groups in coping strategies and psychological well-being were investigated at baseline (T0) and after session 7 (T2). No significant differences were found between the “time of assessment” and “group” factors for both the PGWBI questionnaire ($F_{1,70}=0.63; P>.05$) and all the COPE-NVI subscales (COPE-NVI–Social Support: $F_{1,70}=0.95$ and $P>.05$; COPE-NVI–Avoidance Strategies: $F_{1,70}=3.01$ and $P>.05$; COPE-NVI–Positive Attitude: $F_{1,70}=3.59$ and $P>.05$; COPE-NVI–Problem-Solving: $F_{1,70}=0.28$ and $P>.05$; COPE-NVI–Turning to Religion: $F_{1,70}=3.71$ and $P>.05$).

### Discussion

#### Principal Findings

Research on the use of personalized VR scenarios for promoting relaxation is growing, although few conclusions on their effectiveness have been reached. We know even less about the usefulness of integrating standardized and evidence-based relaxation techniques (eg, PMRT) with new technologies that could promote the independent use by patients, providing a more economical solution to treatment costs. As stated...
previously, the possibility of personalizing audio and visual stimuli in a VR environment is a promising approach for meeting users’ preferences and needs [22], and it has shown encouraging results in terms of both feasibility and potential impact on psychological well-being [18].

The general aim of this pilot study was to evaluate the impact of a novel complementary relaxation training composed of PMRT and the exposure to a personalized relaxing scenario in VR aimed at reducing anxiety and promoting relaxation in a sample of university students.

First, we were interested in investigating the differences in state anxiety between and within groups in the T1 session, where participants were directly exposed to a virtual or imaginative relaxation session. Although a reduction in state anxiety was observed in both the VR and GI groups, our results highlighted that being in a personalized virtual scenario promoted a more significant reduction in self-perceived state anxiety and tension than being in the imaginative condition. Our results are consistent with those of other studies that showed that VR could be a useful tool in reducing stress and promoting relaxation from an assessment based on self-reported measures [52]. Although the target user sample in the study by Hoag et al [53] comprised children and young adults with acute and chronic illnesses, our results align with their results, showing a significant decline in state anxiety from before to after VR exposure. Our data support the role of VR in facilitating a powerful distraction effect on users, reducing their focus on thoughts and external events that would elicit anxiety responses [54,55].

In addition, our data highlighted how engagement with VR scenarios contributes to reducing perceived anxiety after the VR experience. Indeed, it seems that exposure to the VR scenario instead of the imaginative one facilitates the reduction in anxiety levels after the relaxation session. Even if our results need to be further investigated with larger user samples, they are aligned with those of previous studies that underlined the potential benefit of using pleasant, relaxing, and immersive VR scenarios for facilitating relaxation and engagement in individuals from the general population [18,56,57]. A point of strength of this experimental design is the assessment of the impact that the ability to control and vividly picture mental stimuli in a VR environment is a promising approach for meeting users’ preferences and needs [22], and it has shown encouraging results in terms of both feasibility and potential impact on psychological well-being [18].

Moreover, the VR group participants perceived the recalled scenario at T2 as more realistic than individuals in the GI group did. In addition, participants in the VR group obtained additional benefits in terms of relaxation and state anxiety reduction related to a more realistic sensorial experience that played a substantial role in facilitating the visualization of the scenario and enabled users to focus their attention on the relaxation activities. Consistent with other studies [55-57], this core outcome confirms the role of immersive VR in promoting relaxation through visualization, engagement, and immersion processes. The prominent impact of being exposed to realistic scenarios in VR on the enhanced visualization at T2 is a key contribution of our study, shedding light on the impact that exposure to VR may also have in more ecological everyday settings, such as when people are not wearing the HMD but have the chance to transfer the relaxation skills learned in VR to real-world situations. Our data are consistent with those of studies that show that a graphical representation of a scenario is more effective in the retention and recall processes than an imaginative representation [58,59].

Even if the aim was to compare VR scenarios with a scene on a PC, our outcome is in line with that of the study by Krokos et al [60], which highlighted the prominent impact of VR scenarios on memory recall ability. The hypotheses of this study and that by Krokos et al [60] are anchored on classic studies in cognitive psychology based on the method of loci [61] and the context-dependent memory theory [62]. These theories imply the essential role of learning and mnemonic processes in creating an association between the mnemonic content and a mental frame of scenarios and then recalling contents by mentally visualizing the scenarios in which the learning and memorization processes took place [60]. As that presence, immersion, and engagement in the VR scenarios imply sensorimotor contingencies similar to those in the real world [63] and the way we create and recall mental constructs is influenced by perception and action in the environment [64,65], our data coherently confirm the potential of immersive virtual environments in enhancing learning and recall for the intervention of vestibular and sensorimotor inputs [66].

Regarding our collected data on heart rate activity, our findings showed a more significant decrease in heart rate frequency in the VR group than in the GI group but only in the session in which participants were directly exposed to 1 of the 2 experimental conditions (at T1) and not when individuals had to self-administer the relaxation session (at T2). We derive from the differences found at T1 that were absent in T2 that the experience in VR was stronger and more engaging than in the condition in which individuals had to recall the immersive image without HMD support. Our data supporting previous studies’ outcomes highlighted the impact of VR relaxing scenarios in maintaining lower levels of heart rate frequency than normal [67], but additional research is needed to deeply investigate whether and how naturalistic and relaxing VR scenarios induce relaxation and stress reduction by providing feedback on changes, for example, in heart rate frequency and variability, respiration rate, or skin conductance.

An interesting result is also related to differences over time and between groups in trait anxiety scores. Indeed, our data showed
that, in the VR group, the decrease in trait anxiety scores found at T1 was maintained over time (at T2). Nevertheless, the same was not found for the GI group, in which a decrease in anxiety scores was observed at T1, but it significantly increased again at T2, returning to the baseline values (at T0). The 2 groups did not differ at T1, and both showed positive effects of the relaxation sessions. However, the most interesting fact is that, when anxiety was reevaluated after a week and we asked participants to respond considering how they had generally felt during the previous days, people who had used VR claimed to have a lower level of anxiety than that of participants in the other group. These data also need to be further investigated with more comprehensive samples of participants but support the claim that VR plays a crucial role in amplifying the effectiveness of already validated interventions, maintaining their effect over time.

No differences over time were highlighted in coping strategies and psychological well-being, and this could be because our sample was composed of individuals without a clinical diagnosis and high levels of distress at baseline, or it could even be traced back to the fact that these types of psychological constructs require more sessions and time to highlight an effective change.

**Strengths and Limitations**

This contribution adds new knowledge on the importance of customizing and personalizing digital interventions according to the users’ perspectives, needs, and preferences [22]. Another point of strength of this experimental design is the assessment of the impact that the ability to control and vividly picture mental images may have. The outcomes were gathered from a well-designed pilot randomized controlled trial involving 2 selected groups of 72 university students whose sociodemographic and psychological characteristics were controlled for to balance their effect on the variables investigated. The experimental procedure adopted supports the reliability and validity of the results and conclusions presented in this paper. However, this study has some limitations that affect the generalizability of the findings. One limitation is that the results cannot be generalized to the entire nonclinical population as our sample consisted of students from a single university. Moreover, the fact that our sample comprised mainly young female participants may be considered as selection bias as the recruited sample may well be more receptive and motivated to participate in the comparison with people with other sociodemographic features [68]. In general, to further generalize and validate our results’ effectiveness, the involvement of participants belonging to clinical and nonclinical populations, stratified according to different sociodemographic characteristics, should be considered. The use of VR can be helpful in overcoming several barriers that standard relaxation procedures present as it is less costly; promotes the availability of relaxing content that could be difficult to generate in the imagination; and is able to promote the sense of presence, immersion, and engagement closer to what can be obtained in a real-world situation but in a safer context. This is also a reason for considering its further investigation based on clinical and nonclinical samples in future studies. The preliminary results of our study highlight the potential of VR in reducing the number of psychotherapy sessions and their cost as it allows for partial self-management of the treatment.

Furthermore, in the case of patients with medical problems, the integration of VR could facilitate the administration of the relaxation intervention during specific invasive treatments (such as chemotherapy). If the merged administration of a customized VR relaxation scenario and PMRT is effective in obtaining a better subjective perception of relaxation than the standard procedure, it could allow the users to relax with greater autonomy. For this reason, assessing the efficacy of the PMRT and GI in alternative ways could extend treatment administration, especially in situations in which the standard procedure is more challenging. Future studies should consider structuring the relaxation protocol with more VR sessions to better understand the impact of a virtual scenario on relaxation.

Another aspect to be considered in further experimental designs is the introduction of a control group that receives the training session with the physical presence of a therapist. Considering the impact of customization on anxiety and on engaging users in the virtual scenarios, another important aspect that needs to be considered is the opportunity to introduce customized stimuli based on the personal life experience of each participant in the virtual environment. As an example, introducing olfactory stimuli could be another essential aspect to enhance the sense of realism, immersion, and presence in the virtual scenarios [69,70]. A further limitation of this study is the possible exclusion of people who are not familiar with social media. Another limitation regards the limited use of objective data (eg, psychophysiological outcomes) for measuring anxiety and relaxation and the limited range of customization settings offered to users to adapt the VR environment to their preferences. The limitations of this study can be overcome through further investigation.

**Conclusions**

The main objective of psychological interventions is to offer people the opportunity to learn strategies to manage their daily lives independently and effectively. The opportunity to deploy VR for promoting self-management of state anxiety also in real-world situations constitutes an interesting line of investigation to be further explored in future studies involving larger nonclinical and clinical populations (eg, patients with chronic pain) to validate and standardize relaxation protocols integrated with new VR tools. This study has shown that personalized VR scenarios can be effective in improving relaxation and decreasing anxiety when integrated with the PMRT as a complementary relaxation method, thereby highlighting the need for further investigation.

**Acknowledgments**

This research project was funded by the Fondazione Valorizzazione Ricerca Trentina with a grant for the project “ViRLab_Salute,” second call of the Next Generation 2021 grant. The authors wish to thank the students from the University of Padova for participating...
in the study and Dr Lucia Ronconi (“Statistical Consulting Service,” University of Padova) for her contribution in revising the analysis.

Data Availability
The data set generated and analyzed during this study is available from the corresponding author upon reasonable request.

Authors’ Contributions
SP wrote the manuscript, designed the study protocol, ran the intervention trial, conducted data analyses, and contributed to manuscript editing and revision. SG contributed to the design of the study protocol, revised the analysis and the original manuscript, and contributed to manuscript editing and revision. SO contributed to the curation and formal analysis, wrote the manuscript, and contributed to manuscript editing and revision. FF contributed to the curation and formal analysis and to manuscript editing and revision. MD worked on the hardware and software components. SF contributed to the design of the study protocol. CL contributed to manuscript editing and revision. CN contributed to the design of the study protocol, revised the analysis and the original manuscript, and contributed to manuscript editing and revision. All the authors have read and agreed to the published version of the manuscript.

Conflicts of Interest
None declared.

Multimedia Appendix 1
Between-group pairwise comparisons (Bonferroni CI adjustment).
[DOCX File, 15 KB - mental_v11i1e48649_app1.docx ]

Multimedia Appendix 2
CONSORT-EHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth) checklist (V 1.6.1).
[PDF File (Adobe PDF File), 750 KB - mental_v11i1e48649_app2.pdf ]

References


26. Freesound. URL: https://freesound.org/ [accessed 2020-01-23]


Abbreviations

COPE-NVI: Coping Orientation to the Problems Experienced–Nuova Versione Italiana
DASS-21: Depression, Anxiety, and Stress Scale–21
GI: guided imagery
HMD: head-mounted display
ITC-SOPI: International Test Commission–Sense of Presence Inventory
MANCOVA: multivariate analysis of covariance
PGWBI: Psychological General Well-Being Index
PMRT: progressive muscle relaxation technique
STAI-Y: State-Trait Anxiety Inventory–Form Y
STAI-Y1: State-Trait Anxiety Inventory–Form Y1
VAS: visual analog scale
VR: virtual reality
VRSQ: Virtual Reality Symptom Questionnaire

https://mental.jmir.org/2024/1/e48649
JMIR MENTAL HEALTH 2024 | vol. 11 | e48649 | p.383
(page number not for citation purposes)
Review

Effectiveness and User Experience of Virtual Reality for Social Anxiety Disorder: Systematic Review

Simon Shahid1, BPsych, MPsych; Joshua Kelson1, PhD; Anthony Saliba1, PhD
Faculty of Business, Justice, and Behavioural Sciences, Charles Sturt University, Bathurst, Australia

Corresponding Author:
Joshua Kelson, PhD
Faculty of Business, Justice, and Behavioural Sciences
Charles Sturt University
Panorama Ave
Bathurst, 2795
Australia
Phone: 61 2 6338 4570
Email: jkelson@csu.edu.au

Abstract

Background: Social anxiety disorder (SAD) is a debilitating psychiatric disorder that affects occupational and social functioning. Virtual reality (VR) therapies can provide effective treatment for people with SAD. However, with rapid innovations in immersive VR technology, more contemporary research is required to examine the effectiveness and concomitant user experience outcomes (ie, safety, usability, acceptability, and attrition) of emerging VR interventions for SAD.

Objective: The aim of this systematic review was to examine the effectiveness and user experience of contemporary VR interventions among people with SAD.

Methods: The Cochrane Library, Emcare, PsycINFO, PubMed, ScienceDirect, Scopus, and Web of Science databases were searched between January 1, 2012, and April 26, 2022. Deduplicated search results were screened based on title and abstract information. Full-text examination was conducted on 71 articles. Studies of all designs and comparator groups were included if they appraised the effectiveness and user experience outcomes of any immersive VR intervention among people with SAD. A standardized coding sheet was used to extract data on key participant, intervention, comparator, outcome, and study design items.

Results: The findings were tabulated and discussed using a narrative synthesis. A total of 18 studies met the inclusion criteria.

Conclusions: The findings showed that VR exposure therapy–based interventions can generally provide effective, safe, usable, and acceptable treatments for adults with SAD. The average attrition rate from VR treatment was low (11.36%) despite some reported user experience difficulties, including potential simulator sickness, exposure-based emotional distress, and problems with managing treatment delivered in a synchronous group setting. This review also revealed several research gaps, including a lack of VR treatment studies on children and adolescents with SAD as well as a paucity of standardized assessments of VR user experience interactions. More studies are required to address these issues.

Trial Registration: PROSPERO CRD42022353891; https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=353891

(JMIR Ment Health 2024;11:e48916) doi:10.2196/48916

KEYWORDS
social anxiety disorder; social phobia; virtual reality; VR; VR exposure therapy; effectiveness; user experience; safety; usability; acceptability; anxiety; phobia; exposure; systematic; review methods; review methodology; social; psychiatric; mental health; mobile phone

Introduction

Background

Social anxiety disorder (SAD; also known as social phobia) is a psychiatric disorder that is distinguished by a fear of humiliation or negative evaluation by others [1]. Current guidelines highlight the use of cognitive behavioral therapy to treat SAD [2]. This can involve activities such as psychoeducation, relaxation, distraction, cognitive restructuring, exposure therapy, and relapse prevention [2]. Despite the efficacy of cognitive behavioral therapy for SAD treatment [3,4], many who are diagnosed do not go on to seek help [5]. This is partly due to the in vivo (real-life) nature of the exposure
therapy used to desensitize and habituate patients to feared situations. It takes significant time, effort, and resources to accurately recreate scenarios that will incite an appropriate level of fear response in social settings [6]. For example, to conduct in vivo exposure therapy with an individual with a fear of public speaking, a therapist would need to gather an audience in an appropriate context (ie, ensuring confidentiality and nonjudgment). Furthermore, social anxiety–provoking environments can be unpredictable, providing therapists with little control and a higher chance that a patient is embarrassed, leading to higher attrition rates [7]. To potentially overcome these issues with delivering in vivo exposure therapy, some researchers have examined the use of virtual reality (VR) technology [7].

**VR Technology**

VR technology provides a digital modality to deliver psychological interventions [7,8]. It involves the use of computer hardware and software technology (eg, stereoscopic displays of digital environments) to simulate real-world experiences [7]. For instance, one may enter a virtual environment that mimics a physical environment and could adopt a virtual avatar to interact with this virtual environment [9]. VR was first formulated in the 1960s, with the first commercial device developed in the 1980s [10]. As technology has developed, the quality of images has improved, and costs have been reduced. VR systems can be divided into 2 categories: immersive and nonimmersive systems [11]. Immersive systems, such as head-mounted displays (HMDs) or cave automatic virtual environments (CAVEs), provide users with a realistic experience of VR environments, whereas nonimmersive systems, such as computer monitors, result in users not feeling as present [12]. Presence in VR refers to the extent of an individual’s perception of being in a particular environment [13]. For VR therapy to be effective, an individual must feel present and immersed in the digital environment [14]. A CAVE system consists of an empty room with multiple screens arranged in a cubelike formation with users wearing stereoscopic glasses and interacting with virtual objects projected onto the screens [15]. Although CAVE systems have the potential to be more immersive than nonimmersive systems, they are expensive and complex to set up, require frequent physical and digital adjustments, and require dedicated personnel [16]. Conversely, a recent systematic review found that current HMDs offer a more immersive experience than CAVEs and are significantly more user-friendly in cost and setup, with a “plug-n-play” setup solution [15].

When integrated with therapy, VR technology can help address factors that influence the success of exposure-based treatments. For instance, VR allows for the creation of controlled digital environments, which enables therapists to predictably customize exposure scenarios to the specific needs and fears of individual clients [7]. VR can also improve accessibility to exposure therapy for individuals who find it logistically challenging or emotionally overwhelming to engage in real-world scenarios [17]. The immersive nature of VR helps bridge the gap between simulated experiences and real-life situations, fostering a sense of presence and engagement that can potentially enhance treatment adherence and effectiveness [17].

**Effectiveness of VR for SAD**

Virtual environments and avatars can be used to simulate socially distressing situations for SAD treatment. For example, a study immersed participants with SAD into a computer-generated classroom where they were asked to speak publicly on a topic while a therapist controlled the virtual audience’s reactions according to the stage of therapy [18]. VR environments have also been shown to provide acceptable levels of presence and immersion that are necessary for exposure therapy in youth with social anxiety [19]. Several systematic reviews and meta-analyses have demonstrated the effectiveness of VR exposure therapy (VRET) in the treatment of SAD [6,20-24]. Indeed, researchers have established a large effect size for VRET versus waitlist (g=0.90), a medium to large effect size for VRET versus psychological placebo conditions (g=0.78) [21], a large overall effect size for VRET (g=0.82) [22], and a medium to large effect size for VRET at the 12-month follow-up (g=0.74) [6]. This consistent pattern of symptom reduction can be observed across various contexts, such as participant countries (eg, the United States, France, Israel, and South Korea) and treatment settings (eg, universities, hospitals, and clinics) [6,20-24]. However, although existing reviews have explored VR-based therapy from an effectiveness standpoint (eg, reduction in anxiety symptoms), there are gaps in the literature on evaluating the VR user experience for people with SAD on key concomitant outcomes of safety, usability, acceptability, and attrition in different contexts.

**User Experience of VR for SAD**

**Safety**

Studies using VR for workplace training, physical rehabilitation, psychological therapy, and other settings highlight a significant safety issue: simulator sickness [25]. Simulator sickness (otherwise known as VR sickness or cybersickness) [26-28] is characterized by general discomfort, headache, eyestrain, nausea, difficulty concentrating, fatigue, blurred vision, dizziness, and vertigo. On the basis of postural instability theory, simulator sickness is arguably due to VR technology inducing sensory differences in the visual and vestibular systems, which coordinate balance and movement [28-30]. The human body may interpret these disparities as possibly deadly causes (ie, consuming poison) and seek to purge as a result [25]. Consequently, simulator sickness can have a negative impact on participants during VR use and for hours following use [31]. Other aspects of safety include physical injuries from repetitive strain, users colliding with objects in the real world, poor posture, headache discomfort, risk of inducing epileptic seizures, negative mood changes, and infection control [32]. Overall, these issues might put participants at risk of harm or cause them to discontinue using VR. Thus, a comprehensive examination of VR safety for SAD is necessary.

**Usability**

There does not yet appear to be a framework for the evaluation of VR usability in therapy-based applications. Nielsen [33] defines usability as a “quality attribute” that assesses how easy it is to interact with an interface. He highlighted 5 components: learnability (how easy it is for a beginner to use the interface),
efficiency (once the user has learned to use the interface, how quickly they can perform tasks), memorability (re-establishing proficiency after a period of absence), errors (frequency, severity, and recoverability of errors), and satisfaction (level of pleasure from using the interface) [33]. Although this framework is applied to website design, it can also be applied to participants’ perceptions of the usability of VR. Furthermore, the application of VR in real-world settings (eg, in a therapy room) would likely be performed by a clinician rather than a specialized technician. It is important to note that usability issues may arise among clinicians. For example, they may give up on the technology if components fail to load or connect to each other. For this reason, this review included both clinician experiences in administering VR-based therapy and client experiences.

Acceptability
Acceptability is a crucial consideration when evaluating VR interventions [34]. It involves assessing the degree to which the new intervention and its components are received and aligned with the needs of the target population [34]. For example, a study examining VR use in adults with SAD defined acceptability as a participant’s willingness to use a VR program [35]. They measured acceptability by observing rates of attrition and responses to the following question—“Would you recommend this program to others who might have problems similar to yours?”—and inviting further feedback. Participants’ additional feedback was coded into 2 themes: satisfaction (sense of realism, insight, and utility) and perceived effects of the treatment (impact on anxiety). The findings indicated that VR was considered acceptable by participants on all measures [35]. Nevertheless, although there are recent systematic reviews that have addressed the acceptability of VR use for the general population [36], psychosis [37], panic disorder [38], and posttraumatic stress disorder [39,40], a review of the literature on the acceptability of VR in individuals with SAD does not yet exist based on our current knowledge.

Attrition
Attrition, the discontinuation of therapy before treatment completion and resolution of symptoms, can have profound negative effects [41]. These can include the client not fully benefiting from therapy and being discouraged from seeking treatment in the future [41] as well as the effect that this may have on the therapist (eg, loss of revenue, demoralization, and feelings of failure) [42]. A recent meta-analysis of VRET showed significant heterogeneity in attrition rates in the treatment of anxiety disorders, highlighting reasons such as failure to immerse in the virtual environments, simulator sickness, vision complications, and difficulty communicating with a therapist that the participant could not see [43]. A systematic review examining the available literature on rates of attrition of VR-based interventions (both VRET and non-VRET) with participants with SAD does not yet appear to exist based on our current knowledge.

This Study
This study aimed to systematically identify and review available evidence regarding the effectiveness and user experience (ie, safety, usability, acceptability, and rates of attrition) of VR interventions in the treatment of SAD. The following objectives aided in the provision of a comprehensive and up-to-date account of the empirical status of VR therapy for SAD: (1) provide an overview of the existing literature and identify areas in which further research is needed on the treatment of SAD; (2) assess the potential of using VR as a treatment option for SAD, specifically in terms of effectiveness and user experience; and (3) provide guidance and recommendations for future research regarding the use of VR as a treatment option for SAD.

Methods
This systematic review was conducted using the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist [44]. The protocol of this systematic review was prospectively registered in the PROSPERO international database (CRD42022353891).

Eligibility Criteria
For articles to be included in this systematic review, the study participants needed to be people diagnosed with SAD regardless of age. If a study had a mix of people with and without SAD, that study would be included if subgroup analyses were available on the participants with SAD or if they made up the vast majority (ie, ≥ 80%). All studies needed to examine direct participant use of a VR intervention, which includes any system that incorporates immersive VR hardware (ie, HMD or CAVE systems). Only studies that were published after 2012 were included as this marked the introduction of widely available commercial HMD hardware such as the Oculus Rift [9]. Such hardware allowed for the delivery of VR experiences comparable with previously expensive commercial setups at a cheaper cost as well as easier accessibility to researchers [45]. Studies with research design comparators of any kind (eg, comparing VR with other non-VR interventions) were eligible for inclusion. All studies were required to report on VR intervention effectiveness and participant engagement outcomes. This broadly included any standardized or unstandardized measure indicative of usability or acceptability (including attrition rates). Studies of all designs (ie, quantitative, qualitative, and mixed methods) were eligible for inclusion. No studies were excluded based on methodological quality. All the articles needed to be written in English and published in peer-reviewed journals.

Search Strategy
Prominent scientific research databases were searched between January 1, 2012, and April 26, 2022: Cochrane Library, Embase, PsycINFO, PubMed, ScienceDirect, Scopus, and Web of Science. The following keywords were used to search the databases: (“virtual reality” or “VR”) and (“social anxiety” or “social phobia”). The reference lists of eligible articles were also searched.

Article Selection
The search results for all databases were deduplicated, and the remaining article titles and abstracts were scanned. Full-text appraisal was performed on promising articles, and the final study inclusion was agreed upon by the researchers using the
eligibility criteria. Divergent views on inclusion were resolved through discussion and mutual agreement.

Data Extraction

Data from the included studies were extracted by one reviewer (SS) into a standardized coding sheet and then checked by a second reviewer (JK). The data types extracted from eligible papers included the following:

1. Reference source: author surnames, year of publication, and paper title.
2. Sample: country; sample size; and nonidentifiable participant characteristics such as age, sex, and diagnosis.
3. Study design: methodology, comparator trial arms, and measurement points (pretest, midtest, and posttest measurement and follow-up).
4. VR intervention details: intervention program name, purpose of intervention (eg, exposure therapy, cognitive distraction, or relaxation), virtual environment type, hardware (eg, HMD or CAVE system), and treatment length.
5. Effectiveness: standardized measure names, outcomes, and effect sizes.
6. User experience: reported outcomes of intervention safety, usability, acceptability, attrition, and intention-to-treat analyses.

Attrition in this review was defined and measured as the relative number of participants who began using the VR intervention but did not complete measurements during or after the intervention.

Quality Assessment

This systematic review included randomized controlled trial (RCT) and nonrandomized studies. Therefore, the Mixed Methods Appraisal Tool (MMAT) was used to assess the quality of all the included studies [46]. The MMAT was used as it assesses methodological quality across 5 study categories: RCTs, nonrandomized quantitative studies, quantitative descriptive studies, qualitative studies, and mixed methods studies.

Data Analysis

A narrative synthesis approach was used in this systematic review. This involved summarizing and explaining the findings using text as a statistical meta-analysis was not possible because of data heterogeneity across the included studies.

Results

Study Selection

Figure 1 shows that the literature search yielded 683 articles, of which 391 (57.2%) remained after deduplicating citations. Of these 391 records, 18 (4.6%) met the eligibility criteria.
Participant Characteristics

A total of 808 participants were recruited for the VR studies (Table 1). They were largely from South Korea (n=368), followed by the United States (n=163), the Netherlands (n=60), Canada (n=59), Sweden (n=23), Czech Republic (n=10), Denmark (n=9), and Brazil (n=2). The country of origin was missing for some participants (n=114). It is unclear whether participants were unique in 11% (2/18) of studies conducted by research teams with some of the same researchers [47,48]. The sample sizes ranged from 1 to 115 participants, with a median of 48 participants. Participants’ ages ranged from 18 to 65 years. Participants were mainly female, with an average sample proportion of 51.31% (SD 5.36%; range 0%-77.3%). Most participants were diagnosed with SAD. There were 9 participants with a diagnosis of flight phobia and 8 participants with a diagnosis of acrophobia; however, subgroup analyses were available for the participants with SAD in this study [49]. All participant diagnoses were obtained through clinical interviews delivered in person, by phone, or via videoconferencing. In total, 3 therapists were interviewed in addition to the sample of participants in one study [50].
<table>
<thead>
<tr>
<th>Study</th>
<th>Country</th>
<th>Sample size(^a)</th>
<th>Total sample mean age (years; SD)</th>
<th>Total sample age range (years)</th>
<th>Total sample percentage of female participants</th>
<th>Study design</th>
<th>Treatment conditions(^b)</th>
<th>Measurements</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anderson et al [51]</td>
<td>United States</td>
<td>97</td>
<td>39.03 (11.26)</td>
<td>19-69</td>
<td>61.9</td>
<td>RCT(^c)</td>
<td>EGT(^d): 25; VRET(^e): 25; WL(^f): 25</td>
<td>Pre- and posttest measurement and 3- and 12-month FU(^g)</td>
</tr>
<tr>
<td>Arnfred et al [50]</td>
<td>Denmark</td>
<td>9</td>
<td>25.4 (6.54)</td>
<td>19-35</td>
<td>66.7</td>
<td>Interview</td>
<td>VRET: 9</td>
<td>Posttest measurement</td>
</tr>
<tr>
<td>Bouchard et al [52]</td>
<td>Canada</td>
<td>59</td>
<td>34.5 (11.9)</td>
<td>18-65</td>
<td>72.9</td>
<td>RCT</td>
<td>VRET: 17; in vivo: 22; WL: 20</td>
<td>Pre- and posttest measurement and 6-month FU</td>
</tr>
<tr>
<td>Geraets et al [53]</td>
<td>NR(^b)</td>
<td>15</td>
<td>34.9 (12.4)</td>
<td>18-65</td>
<td>53.3</td>
<td>Single group</td>
<td>VRET: 15</td>
<td>Pre- and posttest measurement and 6-month FU</td>
</tr>
<tr>
<td>Hur et al [54]</td>
<td>South Korea</td>
<td>73</td>
<td>NR</td>
<td>NR</td>
<td>42.6</td>
<td>Case controlled</td>
<td>VRET: 25; HC(^i): 22</td>
<td>Pre- and posttest measurement</td>
</tr>
<tr>
<td>Jeong et al [55]</td>
<td>South Korea</td>
<td>115</td>
<td>NR</td>
<td>NR</td>
<td>34.8</td>
<td>Cohort</td>
<td>ET(^j): 52; NT(^k): 43; SE(^l): 20</td>
<td>Pre- and posttest measurement</td>
</tr>
<tr>
<td>Kampmann et al [56]</td>
<td>The Netherlands</td>
<td>60</td>
<td>36.9</td>
<td>18-65</td>
<td>63.3</td>
<td>RCT</td>
<td>VRET: 20; WL: 20; iVET(^m): 20</td>
<td>Pre- and posttest measurement and 3-month FU</td>
</tr>
<tr>
<td>Kim et al [47]</td>
<td>South Korea</td>
<td>54</td>
<td>23</td>
<td>NR</td>
<td>57.7</td>
<td>Controlled clinical trial</td>
<td>VRET: 22; HC: 30</td>
<td>Pre- and posttest measurement</td>
</tr>
<tr>
<td>Kim et al [57]</td>
<td>South Korea</td>
<td>74</td>
<td>NR</td>
<td>19-31</td>
<td>56.9</td>
<td>Longitudinal</td>
<td>VRET: 32; HC: 33</td>
<td>Pre- and posttest measurement</td>
</tr>
<tr>
<td>Kim et al [48]</td>
<td>South Korea</td>
<td>52</td>
<td>NR</td>
<td>19-30</td>
<td>NR</td>
<td>RCT</td>
<td>VRS(^n): 24; WL: 28</td>
<td>Pre- and posttest measurement</td>
</tr>
<tr>
<td>Kovar [58]</td>
<td>Czech Republic</td>
<td>10</td>
<td>34.6 (11.7)</td>
<td>19-51</td>
<td>50</td>
<td>Nonrandomized parallel comparison trial</td>
<td>Psychotherapy: 5; psychotherapy+VRET: 5</td>
<td>Pre- and posttest measurement</td>
</tr>
<tr>
<td>Lindner et al [59]</td>
<td>Sweden</td>
<td>23</td>
<td>40.61 (10.15)</td>
<td>≥18</td>
<td>57</td>
<td>Cohort</td>
<td>VRET: 23</td>
<td>Pre- and posttest measurement</td>
</tr>
<tr>
<td>Moldovan and David [49]</td>
<td>NR</td>
<td>32</td>
<td>NR</td>
<td>≥18</td>
<td>46.9</td>
<td>RCT</td>
<td>VRCBT(^o): 16; WL: 16</td>
<td>Pre- and posttest measurement and FU</td>
</tr>
<tr>
<td>Perandré and Haydu [60]</td>
<td>Brazil</td>
<td>2</td>
<td>23.5 (4.9)</td>
<td>20-27</td>
<td>0</td>
<td>Case study</td>
<td>VRET: 2</td>
<td>Pre- and posttest measurement and 1- and 3-month FU</td>
</tr>
<tr>
<td>Price and Anderson [61]</td>
<td>NR</td>
<td>67</td>
<td>40.31 (11.55)</td>
<td>NR</td>
<td>69</td>
<td>RCT</td>
<td>VRET: 33; EGT: 34</td>
<td>Pre-, mid-, and posttest measurement</td>
</tr>
<tr>
<td>Rubin et al [62]</td>
<td>United States</td>
<td>21</td>
<td>NR</td>
<td>18-65</td>
<td>61.9</td>
<td>RCT</td>
<td>VRET: 10; VRET+AGT(^p): 11</td>
<td>Pre- and posttest measurement and 1-week FU</td>
</tr>
<tr>
<td>Trahan et al [63]</td>
<td>United States</td>
<td>1</td>
<td>36</td>
<td>NR</td>
<td>0</td>
<td>Case study</td>
<td>VRET: 1</td>
<td>Pre- and posttest measurement</td>
</tr>
<tr>
<td>Zainal et al [64]</td>
<td>United States</td>
<td>44</td>
<td>23.3 (9.32)</td>
<td>18-53</td>
<td>77.3</td>
<td>RCT</td>
<td>VRET: 26; WL: 18</td>
<td>Pre- and posttest measurement and 3- and 6-month FU</td>
</tr>
</tbody>
</table>

\(^a\)Refers to the total number of participants in the study.

\(^b\)Refers to the number of participants in each treatment condition.
Details of the VR Interventions

All the studies included VR-based exposure therapy. Nearly all the studies (15/18, 83%) tested a unique VR intervention except for the studies by Jeong et al [55] and Kim et al [47,48] (Table 2). VR hardware included standard computers, smartphones, and HMDs. In total, 17% (3/18) of the studies [49,51,61] did not identify the headset brands. The studies used custom-built software that immersed participants in VR environments that simulated social situations increasing in difficulty with audio, video, text, and interactivity. The treatment lengths ranged from 1 to 14 sessions of exposure therapy, with a mode of 8. A total of 17% (3/18) of the studies [53,55,64] terminated the sessions early if habituation occurred before the completion of the sessions. Participant VR use time ranged from 5 minutes to 3 hours per session, and 17% (3/18) of the studies delivered the VR in a single session [49,59,62]. All VR interventions were tested with therapist or facilitator guidance even though 22% (4/18) [47,48,55,64] were designed to be delivered as self-help.
### Table 2. Details of the virtual reality (VR) interventions.

<table>
<thead>
<tr>
<th>Study</th>
<th>Virtual environments</th>
<th>Headset</th>
<th>Treatment length (duration)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anderson et al [51]</td>
<td>Conference room, classroom, and auditorium</td>
<td>_a</td>
<td>4 exposure sessions (30 min each)</td>
</tr>
<tr>
<td>Amfred et al [50]</td>
<td>Supermarket, meeting, cafeteria, party, and auditorium</td>
<td>Oculus Go</td>
<td>8 exposure sessions (45 min each)</td>
</tr>
<tr>
<td>Bouchard et al [52]</td>
<td>Meeting room, job interview, apartment, coffee shop, neighbors, store, and neutral</td>
<td>eMagin Z800</td>
<td>8 exposure sessions (20-30 min each)</td>
</tr>
<tr>
<td>Geraets et al [53]</td>
<td>Street, bus, café, and supermarket</td>
<td>Sony HMZ-T1</td>
<td>14 exposure sessions (40 min each)</td>
</tr>
<tr>
<td>Hur et al [54]</td>
<td>College student group</td>
<td>HTC Vive</td>
<td>6 exposure sessions (5-8 min each)</td>
</tr>
<tr>
<td>Jeong et al [55]</td>
<td>School, business, and daily life</td>
<td>Samsung Gear VR powered by Oculus</td>
<td>ET&lt;sup&gt;b&lt;/sup&gt; (1-8 exposure sessions); NT&lt;sup&gt;c&lt;/sup&gt; (9-10 exposure sessions); SE&lt;sup&gt;d&lt;/sup&gt; (11-17 exposure sessions)</td>
</tr>
<tr>
<td>Kampmann et al [56]</td>
<td>Audience, stranger, clothes shopping, job interview, journalist interview, restaurant, and blind date</td>
<td>nVisor SX</td>
<td>7 exposure sessions (60 min each)</td>
</tr>
<tr>
<td>Kim et al [47]</td>
<td>School, business, and daily life</td>
<td>Samsung Gear VR powered by Oculus</td>
<td>8 exposure sessions</td>
</tr>
<tr>
<td>Kim et al [57]</td>
<td>College student group</td>
<td>HTC Vive</td>
<td>6 exposure sessions</td>
</tr>
<tr>
<td>Kim et al [48]</td>
<td>School, business, and daily life</td>
<td>Samsung Gear VR powered by Oculus</td>
<td>8 exposure sessions</td>
</tr>
<tr>
<td>Kovar [58]</td>
<td>Public speaking, telephone call, receiving criticism, job interview, refusal of job offer or unwanted product, and working lunch</td>
<td>HTC Vive</td>
<td>8 exposure sessions</td>
</tr>
<tr>
<td>Lindner et al [59]</td>
<td>Board room, conference room, and classroom</td>
<td>Oculus Go</td>
<td>1 exposure session (180 min)</td>
</tr>
<tr>
<td>Moldovan and David [49]</td>
<td>Presentation and interview</td>
<td>_a</td>
<td>1 exposure session (90 min)</td>
</tr>
<tr>
<td>Perandré and Haydu [60]</td>
<td>Food court in shopping center</td>
<td>Oculus Rift</td>
<td>8 exposure sessions</td>
</tr>
<tr>
<td>Price and Anderson [61]</td>
<td>Conference room, classroom, and auditorium</td>
<td>_a</td>
<td>8 exposure sessions</td>
</tr>
<tr>
<td>Rubin et al [62]</td>
<td>Conference room and auditorium</td>
<td>Oculus Rift DK2</td>
<td>1 exposure session (45 min)</td>
</tr>
<tr>
<td>Trahan et al [63]</td>
<td>Grocery store</td>
<td>Plastic HMD&lt;sup&gt;e&lt;/sup&gt; bracket for mobile phone</td>
<td>12 exposure sessions (12-15 min each)</td>
</tr>
<tr>
<td>Zainal et al [64]</td>
<td>Dinner party and job interview</td>
<td>Pico Goblin VR</td>
<td>8 exposure sessions (25-30 min each)</td>
</tr>
</tbody>
</table>

<sup>a</sup>Brand not reported.  
<sup>b</sup>ET: early termination.  
<sup>c</sup>NT: normal termination.  
<sup>d</sup>SE: session extension.  
<sup>e</sup>HMD: head-mounted display.

### Research Designs and Comparators

Table 1 summarizes the research designs and comparators. Nearly half (8/18, 44%) of the studies appraised participant VR use through RCT designs. Comparators included exposure group therapy, in vivo exposure, early and extended termination, attention guidance training using VR, psychotherapy, and waitlist control. All studies (18/18, 100%) had pre- and posttest assessments of user outcomes, although 39% (7/18) also had follow-up assessments, with the longest being 12 months [51].

### Effectiveness Measures and Outcomes

The details of the effectiveness measures and outcomes of VR treatment for SAD are summarized in Table 3. VR treatment effect sizes across all studies that reported them ranged from medium to large. Almost all studies (15/18, 83%) demonstrated a decrease in symptoms following VR treatment.
Table 3. Details on social anxiety measures and virtual reality (VR) effectiveness outcomes.

<table>
<thead>
<tr>
<th>Study and measures</th>
<th>VR effectiveness outcomes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Anderson et al [51]</strong></td>
<td></td>
</tr>
<tr>
<td>PRCS&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Significant improvement in confidence as a speaker from before to after treatment ($d^b=1.19$; $P=.01$), with benefits maintained at the 3- and 6-month follow-ups.</td>
</tr>
<tr>
<td>FNE-B&lt;sup&gt;c&lt;/sup&gt;</td>
<td>Significant decrease in fear of negative evaluation from before to after treatment ($d=0.29$; $P=.01$), with benefits maintained at the 3- and 6-month follow-ups.</td>
</tr>
<tr>
<td>BAT&lt;sup&gt;d&lt;/sup&gt;</td>
<td>Significant improvement in speech length ($d=0.78$; $P=.01$) and peak anxiety ($d=0.70$; $P=.02$) from before to after treatment.</td>
</tr>
<tr>
<td><strong>Arnfred et al [50]</strong></td>
<td></td>
</tr>
<tr>
<td>NSQ&lt;sup&gt;e&lt;/sup&gt;</td>
<td>The virtual environments effectively induced immersion and anxiety in some but not all participants with social anxiety disorder.</td>
</tr>
<tr>
<td><strong>Bouchard et al [52]</strong></td>
<td></td>
</tr>
<tr>
<td>LSAS-SR&lt;sup&gt;f&lt;/sup&gt;</td>
<td>Significant decrease in social anxiety symptoms from before to after treatment compared with waitlist ($P&lt;.001$) that was maintained at the 6-month follow-up.</td>
</tr>
<tr>
<td>BAT</td>
<td>Significant decrease in behavioral avoidance from before to after treatment ($P&lt;.001$).</td>
</tr>
<tr>
<td>SPS&lt;sup&gt;g&lt;/sup&gt;</td>
<td>Significant decrease in social phobia from before to after treatment ($P&lt;.001$) that was maintained at the 6-month follow-up.</td>
</tr>
<tr>
<td>SIAS&lt;sup&gt;h&lt;/sup&gt;</td>
<td>Significant decrease in social anxiety symptoms from before to after treatment ($P&lt;.001$) that was maintained at the 6-month follow-up.</td>
</tr>
<tr>
<td>FNE&lt;sup&gt;i&lt;/sup&gt;</td>
<td>Significant decrease in fear of negative evaluation from before to after treatment ($P&lt;.001$) that was maintained at the 6-month follow-up.</td>
</tr>
<tr>
<td><strong>Geraets et al [53]</strong></td>
<td></td>
</tr>
<tr>
<td>SIAS</td>
<td>Significant decrease in social interaction anxiety from before to after treatment ($d=0.9$; $P=.008$) that was maintained at the 6-month follow-up ($d=1.3$; $P=.003$).</td>
</tr>
<tr>
<td><strong>Hur et al [54]</strong></td>
<td></td>
</tr>
<tr>
<td>SPS</td>
<td>Significant decrease in social phobia symptoms from before to after treatment ($P=.005$).</td>
</tr>
<tr>
<td>PERS&lt;sup&gt;j&lt;/sup&gt;</td>
<td>Significant decrease in negative postevent rumination from before to after treatment ($P&lt;.001$).</td>
</tr>
<tr>
<td><strong>Jeong et al [55]</strong></td>
<td></td>
</tr>
<tr>
<td>FNE-B</td>
<td>Significant decrease in fear of negative evaluation from first to last session for the early, normal, and extended termination groups ($P&lt;.001$).</td>
</tr>
<tr>
<td>LSAS</td>
<td>Significant decrease in social anxiety symptoms from first to last session for the normal ($P&lt;.001$) and extended termination groups ($P=.002$).</td>
</tr>
<tr>
<td>SPS</td>
<td>Significant decrease in social phobia symptoms from first to last session for the normal ($P=.001$) and extended termination groups ($P&lt;.001$).</td>
</tr>
<tr>
<td>SIAS</td>
<td>Significant decrease in social interaction anxiety from first to last session for the normal ($P&lt;.001$) and extended termination groups ($P=.006$).</td>
</tr>
<tr>
<td><strong>Kampmann et al [56]</strong></td>
<td></td>
</tr>
<tr>
<td>LSAS-SR</td>
<td>Significant decrease in social anxiety symptoms from before to after treatment compared with waitlist ($d=0.55$; $P=.01$) that was maintained at the 3-month follow-up.</td>
</tr>
<tr>
<td>FNE-B</td>
<td>No significant change in fear of negative evaluation compared with waitlist group from before to after treatment or the 3-month follow-up.</td>
</tr>
<tr>
<td>BAT</td>
<td>Significant increase in speech length from before to after treatment compared with waitlist ($d=0.56$; $P=.02$) that was maintained at the 3-month follow-up; however, there was no significant difference in speech performance.</td>
</tr>
<tr>
<td><strong>Kim et al [47]</strong></td>
<td></td>
</tr>
<tr>
<td>HADS&lt;sup&gt;k&lt;/sup&gt;</td>
<td>Significant decrease in anxiety symptoms from before to after treatment ($P&lt;.001$).</td>
</tr>
<tr>
<td>LSAS-SR</td>
<td>Significant decrease in social anxiety symptoms from before to after treatment ($P&lt;.001$).</td>
</tr>
<tr>
<td>SIAS</td>
<td>Significant decrease in social interaction anxiety from before to after treatment ($P&lt;.001$).</td>
</tr>
<tr>
<td><strong>Kim et al [57]</strong></td>
<td></td>
</tr>
<tr>
<td>SPS</td>
<td>Significant decrease in social phobia symptoms from before to after treatment ($P&lt;.001$).</td>
</tr>
</tbody>
</table>
**VR effectiveness outcomes**

<table>
<thead>
<tr>
<th>Study and measures</th>
<th>VR effectiveness outcomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIAS</td>
<td>Significant decrease in social interaction anxiety from before to after treatment ($P&lt;0.001$).</td>
</tr>
<tr>
<td>FNE-B</td>
<td>Significant decrease in fear of negative evaluation from before to after treatment ($P=0.004$).</td>
</tr>
<tr>
<td>KSAD(^\text{1})</td>
<td>Significant decrease in social avoidance and distress from before to after treatment ($P&lt;0.001$).</td>
</tr>
<tr>
<td>LSAS(^\text{m})</td>
<td>Significant decrease in social anxiety symptoms from before to after treatment ($P=0.04$).</td>
</tr>
</tbody>
</table>

**Kim et al [48]**

<table>
<thead>
<tr>
<th>Measures</th>
<th>VR effectiveness outcomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>HADS</td>
<td>No significant changes in anxiety symptoms from before to after treatment.</td>
</tr>
<tr>
<td>LSAS-SR</td>
<td>Significant decrease in social anxiety symptoms from before to after treatment ($P&lt;0.01$).</td>
</tr>
</tbody>
</table>

**Kovar [58]**

Higher average decrease in symptoms on all these measures in the VR treatment group compared with the non-VR treatment group.

**Lindner et al [59]**

<table>
<thead>
<tr>
<th>Measures</th>
<th>VR effectiveness outcomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSAS(^\text{q})</td>
<td>Significant decrease in self-rated public speaking anxiety following the first 3-hour session ($d=0.77$; $P=0.006$).</td>
</tr>
<tr>
<td>LSAS-SR</td>
<td>Significant decrease in social anxiety symptoms from before to after treatment ($P=0.001$).</td>
</tr>
<tr>
<td>FNE-B</td>
<td>Significant decrease in fear of negative evaluation from before to after treatment ($P=0.04$).</td>
</tr>
</tbody>
</table>

**Moldovan and Price [49]**

<table>
<thead>
<tr>
<th>Measures</th>
<th>VR effectiveness outcomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>FNE-B</td>
<td>Significant decrease in fear of negative evaluation from before to after treatment ($P&lt;0.05$).</td>
</tr>
<tr>
<td>SSPS(^\text{z})</td>
<td>Significant decrease in negative self-statements from before to after treatment ($P&lt;0.05$).</td>
</tr>
<tr>
<td>LSAS</td>
<td>Significant decrease in social anxiety symptoms from before to after treatment ($P&lt;0.05$).</td>
</tr>
</tbody>
</table>

**Perandré and Haydu [60]**

<table>
<thead>
<tr>
<th>Measures</th>
<th>VR effectiveness outcomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPIN and BAI</td>
<td>Decrease in anxiety symptoms reported on both measures from the pretest measurement to the 3-month follow-up from treatment for both participants.</td>
</tr>
</tbody>
</table>

**Price and Anderson [61]**

<table>
<thead>
<tr>
<th>Measures</th>
<th>VR effectiveness outcomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSPS</td>
<td>Significant improvements on positive and negative self-statements from before to after treatment ($P&lt;0.01$).</td>
</tr>
<tr>
<td>PRCA-SF(^{a})</td>
<td>Significant decrease in public speaking anxiety from before to after treatment ($P&lt;0.01$).</td>
</tr>
</tbody>
</table>

**Rubin et al [62]**

<table>
<thead>
<tr>
<th>Measures</th>
<th>VR effectiveness outcomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRPSA(^{1})</td>
<td>Significant decrease in fear of public speaking from before to after treatment ($d=-1.11$) and at the 1-week follow-up ($d=-1.68$).</td>
</tr>
<tr>
<td>LSAS-SR</td>
<td>Significant decrease in general symptoms of social anxiety from before to after treatment ($d=-0.60$) and at the 1-week follow-up ($d=-2.07$).</td>
</tr>
</tbody>
</table>

**Trahan et al [63]**

<table>
<thead>
<tr>
<th>Measures</th>
<th>VR effectiveness outcomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>SUDS(^{u})</td>
<td>No significant change in subjective distress from before to after treatment for the participant ($P=0.21$).</td>
</tr>
<tr>
<td>SADS</td>
<td>Score decrease of 52.6% in social anxiety from before to after treatment for the participant.</td>
</tr>
</tbody>
</table>

**Zainal et al [64]**

<table>
<thead>
<tr>
<th>Measures</th>
<th>VR effectiveness outcomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAD composite(^{v})</td>
<td>Significant decrease in social anxiety symptoms from before to after treatment compared with the waitlist group ($g^w=-4.77$; $P&lt;0.001$). No significant within-group change at the 3- ($g=0.12$) and 6-month follow-ups ($g=-0.13$).</td>
</tr>
</tbody>
</table>
VR effectiveness outcomes

Study and measures | VR effect size outcomes
--- | ---
**MAST**<sup>a</sup> | Significant decrease in job interview anxiety from before to after treatment compared with the waitlist group ($g = -4.17; P < .001$). No significant within-group change at the 3- ($g = -0.10$) and 6-month follow-ups ($g = -0.53$).

---

<sup>a</sup>PRCS: Personal Report of Confidence as a Speaker.

<sup>b</sup>Cohen $d$ effect size.

<sup>c</sup>FNE-B: brief Fear of Negative Evaluation Scale.

<sup>d</sup>BAT: behavioral avoidance task.

<sup>e</sup>NSQ: nonstandardized questions.

<sup>f</sup>LSAS-SR: Liebowitz Social Anxiety Scale–Self-Report.

<sup>g</sup>SPS: Social Phobia Scale.

<sup>h</sup>SIAS: Social Interaction Anxiety Scale.

<sup>i</sup>FNE: Fear of Negative Evaluation Scale.

<sup>j</sup>PERS: Post-Event Rumination Scale.

<sup>k</sup>HADS: Hospital Anxiety and Depression Scale.

<sup>l</sup>KSAD: Korean Social Avoidance and Distress Scale.

<sup>m</sup>LSAS: Liebowitz Social Anxiety Scale.

<sup>n</sup>SPIN: Social Phobia Inventory.

<sup>o</sup>SADS: Social Avoidance and Distress Scale.

<sup>p</sup>BAI: Beck Anxiety Inventory.

<sup>q</sup>PSAS: Public Speaking Anxiety Scale.

<sup>r</sup>SSPS: Self-Statements During Public Speaking scale.

<sup>s</sup>PRCA-SF: Personal Report of Communication Apprehension–Short Form.

<sup>t</sup>PRPSA: Personal Report of Public Speaking Anxiety.

<sup>u</sup>SUDS: Subjective Units of Distress Scale.

<sup>v</sup>SAD composite: average standardized scores of the Social Phobia Diagnostic Questionnaire and the SIAS.

<sup>w</sup>Hedges $g$ effect size.

<sup>x</sup>MASI: Measure of Anxiety in Selection Interviews.

User Experience With the VR Interventions

The average attrition rate was 11.36% across all studies in the active VR treatment phase, with a range of 0% to 45.2% (Table 4). A total of 22% (4/18) of the studies reported the use of an intention-to-treat analysis. To measure VR user experience, 56% (10/18) of the studies used standardized measures, and 11% (2/18) of the studies used nonstandardized questions. A total of 67% (8/12) of these studies reported positive VR user experience findings in various areas of presence, usability, acceptability, or satisfaction. Low levels of simulator sickness were reported in 75% (3/4) of the studies that used standardized questions; however, 25% (1/4) of these studies reported higher levels of simulator sickness in participants with SAD than in controls without SAD [47]. No other safety issues, such as physical injury, user collision, postural complaints, headset discomfort, seizures, or infection, were reported.
### Table 4. Virtual reality (VR) interventions and user experience outcomes.

<table>
<thead>
<tr>
<th>Study</th>
<th>Measures</th>
<th>VR user experience findings</th>
<th>Attrition (%)</th>
<th>ITT&lt;sup&gt;a&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anderson et al [51]</td>
<td>CSQ&lt;sup&gt;b&lt;/sup&gt;</td>
<td>High satisfaction with VR was reported after treatment and maintained at the 12-month follow-up.</td>
<td>5/30 (17)</td>
<td>Yes</td>
</tr>
<tr>
<td>Amfred et al [50]</td>
<td>NSQ&lt;sup&gt;c&lt;/sup&gt;</td>
<td>A high level of presence in virtual environments for some participants but not all. There were technical issues with setting up and storing away equipment for the group. Wearing the HMD&lt;sup&gt;d&lt;/sup&gt; in front of strangers was more anxiety provoking than the virtual environments for some participants. All patients found VR to be a meaningful addition to their therapy sessions, with several wanting more exposure.</td>
<td>0/9 (0)</td>
<td></td>
</tr>
<tr>
<td>Bouchard et al [52]</td>
<td>SSQ&lt;sup&gt;f&lt;/sup&gt;, PQQ&lt;sup&gt;g&lt;/sup&gt;, and GPQ&lt;sup&gt;h&lt;/sup&gt;</td>
<td>No significant increases in simulator sickness after exposure sessions (P&gt;.20). Good level of presence that increased with a higher number of exposures.</td>
<td>2/17 (12)</td>
<td>Yes</td>
</tr>
<tr>
<td>Geraets et al [53]</td>
<td>—</td>
<td>VR treatment was well tolerated and deemed acceptable for most participants.</td>
<td>2/15 (13)</td>
<td></td>
</tr>
<tr>
<td>Hur et al [54]</td>
<td>—</td>
<td></td>
<td>16/73 (21)</td>
<td></td>
</tr>
<tr>
<td>Jeong et al [55]</td>
<td>—</td>
<td></td>
<td>52/115 (45)</td>
<td></td>
</tr>
<tr>
<td>Kampmann et al [56]</td>
<td>—</td>
<td>Simulator sickness led one patient to drop out.</td>
<td>5/20 (25)</td>
<td></td>
</tr>
<tr>
<td>Kim et al [47]</td>
<td>SSQ</td>
<td>Participants with SAD&lt;sup&gt;i&lt;/sup&gt; experienced significantly more simulator sickness than participants without SAD (P=.003).</td>
<td>2/54 (4)</td>
<td></td>
</tr>
<tr>
<td>Kim et al [57]</td>
<td>—</td>
<td></td>
<td>9/74 (12)</td>
<td></td>
</tr>
<tr>
<td>Kim et al [48]</td>
<td>SSQ</td>
<td>Low levels of simulator sickness.</td>
<td>3/24 (13)</td>
<td></td>
</tr>
<tr>
<td>Kovar [58]</td>
<td>—</td>
<td></td>
<td>0/10 (0)</td>
<td></td>
</tr>
<tr>
<td>Lindner et al [59]</td>
<td>NEQ&lt;sup&gt;j&lt;/sup&gt;</td>
<td>High stress levels and low levels of satisfaction.</td>
<td>3/23 (13)</td>
<td>Yes</td>
</tr>
<tr>
<td>Moldovan and David [49]</td>
<td>ITQ&lt;sup&gt;k&lt;/sup&gt; and PQ</td>
<td>No moderating effect of immersion and presence on pre- and posttest anxiety.</td>
<td>0/32 (0)</td>
<td></td>
</tr>
<tr>
<td>Perandré and Haydu [60]</td>
<td>SPI&lt;sup&gt;l&lt;/sup&gt;</td>
<td>High sense of presence reported by both participants.</td>
<td>0/2 (0)</td>
<td></td>
</tr>
<tr>
<td>Price and Anderson [61]</td>
<td>—</td>
<td>From a randomly selected subset of videotaped sessions (14%), high participant compliance was found, with 92% of the VR treatment protocol being completed.</td>
<td>0/33 (0)</td>
<td></td>
</tr>
<tr>
<td>Rubin et al [62]</td>
<td>—</td>
<td></td>
<td>2/21 (10)</td>
<td></td>
</tr>
<tr>
<td>Trahan et al [63]</td>
<td>SUS&lt;sup&gt;m&lt;/sup&gt;</td>
<td>High usability reported by the participant.</td>
<td>0/1 (0)</td>
<td></td>
</tr>
<tr>
<td>Zainal et al [64]</td>
<td>NSQ, IPQ&lt;sup&gt;n&lt;/sup&gt; and SSQ</td>
<td>Acceptable presence and low levels of simulator sickness. High levels of homework compliance. Participants (85%) would recommend it to others with SAD. High levels of acceptability and usability.</td>
<td>9/44 (21)</td>
<td></td>
</tr>
</tbody>
</table>

---

<sup>a</sup>IT: intention-to-treat analysis.
<sup>b</sup>CSQ: Client Satisfaction Questionnaire.
<sup>c</sup>NSQ: nonstandardized questions.
<sup>d</sup>HMD: head-mounted display.
<sup>e</sup>Not reported.
<sup>f</sup>SSQ: Simulator Sickness Questionnaire.
<sup>g</sup>PQ: Presence Questionnaire.
<sup>h</sup>GPQ: Gatineau Presence Questionnaire.
<sup>i</sup>SAD: social anxiety disorder.
<sup>j</sup>NEQ: Negative Effects Questionnaire.
<sup>k</sup>ITQ: Immersive Tendencies Questionnaire.
<sup>l</sup>SPI: Sense of Presence Inventory.
<sup>m</sup>SUS: System Usability Scale.
<sup>n</sup>IPQ: Igroup Presence Questionnaire.

**Quality Assessment Results**

Multimedia Appendix 2 [47-64] contains a table of quality assessment results for the included studies. In all RCT studies [48,49,51,52,56,61,62,64], randomization was reported, but schedule details were unclear in 11% (2/18) of the studies [48,61]. All RCT studies reported comparable baseline group
analyses. In total, 38% (3/8) of the RCT studies reported complete outcome data, which is defined as ≥80% [49,56,64]. All but the RCT studies by Kim et al [48], Moldovan and David [49], Price and Anderson [61], and Rubin et al [62] reported blinding of outcome assessors, which was applied at the pretest measurements. All RCT studies except those by Bouchard et al [52], Kamppmann et al [56], and Rubin et al [62] reported that participants adhered to their assigned VR interventions.

In the quantitative descriptive studies [53,60,63], the sampling strategy was relevant to the research question except in 33% (1/3) of the studies, in which details were unclear [53]. All quantitative descriptive study samples were representative of the target population, and the measures fulfilled the inclusion criteria. Nonresponse bias was low in all studies except one (2/3, 67%) [60]. Statistical analyses were appropriate to answer the research question in 33% (1/3) of the studies [53] but unclear in the other 2 [60,63].

In the quantitative nonrandomized studies [47,54,55,57-59], participants were representative of the target population, measurements were appropriate regarding both the outcome and intervention, and there were complete outcome data (defined as ≥80%) in all but 2 studies (4/6, 67%) [54,55]. Confounds were accounted for in the design and analysis of 50% (3/6) of the studies [54,55,59]. In total, 67% (2/3) of the quantitative nonrandomized studies reported that the intervention was administered as intended [54,55].

In the single qualitative interview study [50], the qualitative approach was appropriate to answer the research question; the data collection methods were adequate to address the research question; findings were adequately derived from the data; the interpretation of the results was sufficiently substantiated by the data; and there was coherence between qualitative data sources, collection, analysis, and interpretation.

**Discussion**

**Principal Findings**

**Overview**

SAD is a common and debilitating anxiety disorder that affects occupational and social functioning [2]. Current in vivo–based exposure therapies require significant time, resources, and effort, which results in limited treatment dissemination [6]. VR technology provides an alternative modality for treating SAD [19]; however, contemporary evidence on the user experience of VR for SAD is sparse. This systematic review was conducted to provide a comprehensive and up-to-date account of the available evidence regarding the effectiveness and user experience (ie, safety, usability, acceptability, and attrition) of VR interventions for the treatment of SAD.

**Effectiveness of VR Interventions for SAD**

Our review found that VR interventions can effectively treat SAD in adult populations, which is congruent with the existing literature [6,20-24]. It is interesting to note that, although our search terms and inclusion criteria were open to any VR-based intervention for treating people with SAD (eg, providing relaxation, cognitive distraction, exposure therapy, and psychoeducation), all the included interventions were intended for exposure therapy. This indicates that VRET dominates the research field of VR-based interventions for SAD.

Studies including follow-up measures highlight the maintenance of SAD symptom improvement from 1 week [62] to 1 year [51], indicating that VRET can provide effective short- and long-term treatment for SAD symptoms. This is impressive given that the study showing maintained benefits for up to 1 year involved only 4 treatment sessions [51]. However, it is important to note that this study only included participants with a fear of public speaking as the primary social fear as opposed to other social situations (eg, going to dinner with friends), limiting the generalizability of the findings [65]. Nevertheless, our findings suggest that VRET can be a rapidly effective treatment for SAD with the potential to provide long-term symptom improvement.

**Safety of VR Interventions for SAD**

Simulator sickness was a common measure of safety in the reviewed studies. Participant simulator sickness was reportedly low in most studies. However, it was found that participants with an SAD diagnosis were more prone to simulator sickness when compared with participants without SAD in one study [47]. This could be because patients with anxiety tend to experience greater motion discomfort [66,67]. For example, patients with anxiety may be more susceptible to irregular breathing and hyperventilation, leading to dizziness and nausea when exposed to fear-inducing cues. This may exacerbate the body’s interpretation of disparities in visual and vestibular systems as possible deadly causes (ie, poison) and potentially lead to nausea and vomiting [25]. Another safety consideration is the absence of other physical injuries (eg, collisions with real-world objects, poor posture, headset discomfort, and seizures) reported in the reviewed studies, which supports VR as a safe SAD treatment.

However, although the research safety findings are encouraging, the limitations of these studies are important to note. For example, most studies screened out participants who were unable to tolerate the VR environment and HMD or those who had a history of seizures. This would result in a sampling bias in favor of VR safety. Furthermore, all studies except for one [63] were conducted in controlled settings (ie, hospitals and clinics) that were supervised by clinicians, further reducing risks that would otherwise be significant when using VR alone. For example, an individual purchasing and using a VR system at home may collide with real-world objects without the intervention of a third party. As such, more research is required to evaluate the safety of VR for SAD in nonclinical, unsupervised settings.

**Usability of VR Interventions for SAD**

There was large variability in the VR software used for SAD. This is likely due to the infancy of VR for SAD. With such variability, it is inevitable that reports of usability will vary according to the hardware and software used, with some programs being easier to use than others.

A distinct hindrance in evaluating the usability of VR for SAD was the lack of an existing framework. The studies largely used nonstandardized questions and qualitative feedback to determine usability, making it difficult to generalize findings across
multiple studies. Although most studies did not comment on aspects of usability, those that did provided valuable information on the usability of VR for SAD. Studies in which practitioners delivered VR therapy to individual participants reported high levels of usability, such as the ease of setting up and navigating the hardware and software. However, reports of VR use in a group setting described low levels of usability, significant amounts of time spent on setting up and storing the equipment, and loss of focus on the exposure experience when therapists were helping others with their HMDs [50].

The differences in usability between individualized and group settings highlight important requirements for the use of VR interventions for SAD. Primarily, VR technology for SAD needs to be easy to learn by patients, and it is important that errors are limited in frequency and severity and that patients can recover from errors largely autonomously. As such, we propose a “VR usability framework” for the measurement of usability of VR for SAD that borrows elements from the usability heuristics by Nielsen [33]: (1) “learnability,” assessing how easy it is for a patient to set up and learn the VR technology; (2) “errors,” assessing the frequency, severity, and recoverability of errors autonomously by the patient; and (3) “memorability,” how easy it is to re-establish proficiency after a period of absence.

Using the VR usability framework, current trends show variability in the usability of VR for SAD. VR used in group therapy has a steep learning curve and requires substantial input from therapists to work through errors, and it is difficult to re-establish proficiency in it after a period of absence (eg, some participants wished they could take the equipment home) [50]. In contrast, VR used in individualized therapy is easy to learn, patients can autonomously handle errors, and they are familiar with the technology upon return [63,64]. Thus, VR may be more user-friendly in one-on-one therapy as opposed to a group setting, as articulated by the VR usability framework.

Acceptability of VR Interventions for SAD

The results we found regarding high VR acceptability in adult patients with SAD are congruent with earlier research by Saxena [35]. Empirical findings indicate that VR for SAD is generally acceptable to adult patients with SAD, with high scores on standardized measures of satisfaction reported by most patients. Positive qualitative responses suggest that VR allowed patients to gain more insights into their anxiety and a better understanding of the social situations that they would normally avoid or be too emotionally activated to observe [50]. For example, in real life, an individual with social anxiety may avoidantly play with their phone when someone sits next to them in a cafeteria rather than perceive the encounter as a valued learning experience. Therefore, it is likely that many adults with SAD who willingly undergo VR therapy will find the experience acceptable.

Conversely, one study [59] found that some patients reported that their expectations for the treatment were not fulfilled, and some reported feeling more stress during VR. It was also found that positive expectations of VR effectiveness as well as a positive working alliance with the therapist were significantly correlated with positive emotional changes [49]. Therefore, VR treatment may not be acceptable for all adults with SAD based on individual differences regarding their previous VR experience, their perceptions of VR therapy helpfulness, their level of distress tolerance to exposure to digital stimuli before habituation [68], and the nature of their relationship with the therapist offering VR treatment.

Attrition of VR Interventions for SAD

This review found that the attrition rate across most studies was relatively low and within acceptable levels (≤20%) [46]. Indeed, attrition rates for the use of VR interventions for SAD were found to be substantially lower than estimates from VRET in anxiety disorders [43]. Considering this, it appears that patients with SAD continue with treatment more than other patients with anxiety.

There may be several reasons for the low average attrition rate finding. First, patients with SAD may prefer to learn more about social situations in a VR space. An individual with SAD may be curious about learning about social situations but may struggle to overcome the anxiety associated with placing themselves in an environment where negative evaluation is possible. By engaging with VR, patients with SAD have the knowledge that they can exit the simulation at any point, giving them the opportunity to learn about social situations without real-world social consequences. Second, patients with SAD may be more tolerant of the potentially negative effects of VR (eg, simulator sickness) when compared with the general population with anxiety [43]. Third, patients with SAD may be more hesitant to drop out of therapy for fear of negative evaluation by examiners. For instance, patients with SAD may be more likely to remain in a study because of social desirability bias—the tendency to respond in a certain way to avoid criticism [69].

It is important to note that the observed attrition rates are heterogeneous. Some studies reported proportionally higher attrition rates than others [54,55]. This may be due to the differences in the number of sessions involved in different studies. For example, some studies were composed of single sessions [49,59,62], whereas the study with the highest attrition had 9 to 17 sessions [55]. As it takes longer to deliver all sessions, there is more opportunity for participants to drop out. Furthermore, attrition was defined in this review as those who did not complete measurements during or after intervention use, including completion of follow-up measures. Considering that some studies included follow-up measures of 3 months after the intervention or longer, it is plausible that participants may not have re-engaged in these measures for several reasons. These could potentially include both therapy-related factors (eg, intolerance of VR-induced anxiety, simulator sickness, and low satisfaction levels) or factors outside of therapy (eg, moving away, becoming too busy in everyday life, and major life events).

Recommendations

Safety

With regard to safety, the primary issue identified in this review was simulator sickness. Several factors appear to be related to the susceptibility to simulator sickness. If simulator sickness is exacerbated by physiological symptoms of anxiety (eg,
hyperventilation leading to dizziness] [66,67], it may be helpful to target these symptoms with clinical treatment before using VR technology. This is in line with other studies exploring attrition in anxiety disorders [43], which found that VRET attrition occurs early in treatment because of factors such as dizziness. As such, VR protocols for SAD should aim to improve retention at the beginning of treatment using a phase-based approach that includes strategies to tolerate negative emotions before immersion in VR. These may include implementation of relaxation and grounding techniques [70] or the prescription of antinausea medications. Changes in VR technology can also be applied to reduce simulator sickness [48]. Blurring or lowering the resolution of a VR image has been shown to reduce simulator sickness and improve the sense of reality [71,72].

Future Research

Presently, there are many research gaps in the literature regarding the user experience of VR for SAD. The development of a standardized measure to assess the usability of VR for SAD has the potential to identify prominent issues with usability and aid in the development of future VR programs. This measure may include elements identified in the VR usability framework discussed previously to assess learnability, error recoverability, and memorability. This may be applied to technical developments in VR that would likely improve VR’s “plug-n-play” capability for SAD and other anxiety disorder treatments. Future research should also delve deeper into the study of simulator sickness in patients with SAD when compared with both healthy controls and patients with other anxiety disorders. This may lead to valuable information on reducing simulator sickness, thereby reducing the levels of attrition and improving the user experience of VR for SAD. Finally, there were no studies found that specifically targeted a child or adolescent population. Given that the onset of SAD typically occurs around adolescence [1], future studies should evaluate the efficacy of early intervention of VR for SAD, particularly given adolescents’ success with VR for psychological distress [32].

Limitations

This review has several limitations. First, our review did not perform a cost-benefit analysis of the hardware identified (eg, HMDs). Affordability could have implications for the acceptability of VR among consumers with SAD. Second, we included only English-language studies, and there may have been pertinent articles published in other languages. Third, this study conducted a qualitative review of studies with different designs. Although the MMAT [46] was used to assess the quality of the studies, there is still a risk of subjective reviewer bias in addressing its criteria. Finally, studies may have been missed in our search because of obscure nomenclature (eg, research publications that did not clearly specify the use of a VR intervention for SAD in their title and abstract).

Conclusions

Our review findings showed that VRET interventions can generally provide an effective, safe, usable, acceptable, and low-attrition treatment option for adults with SAD. Nevertheless, there are research gaps evident when appraising user experience outcomes. These include the need to conduct more VR research with children and adolescents with SAD. We also do not yet know the specific causes of elevated simulator sickness in patients with SAD compared with participants without SAD or how effective other VR-based interventions beyond exposure therapy (eg, focused on mindfulness, relaxation, or cognitive distraction) are in the treatment of SAD. Further experimental studies (eg, pilot feasibility studies and RCTs) are required to explore these domains.

Acknowledgments

The authors acknowledge the financial support provided by Charles Sturt University.

Conflicts of Interest

None declared.

Multimedia Appendix 1

PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2009 checklist. [PDF File (Adobe PDF File), 146 KB - mental_v11i1e48916_app1.pdf]

Multimedia Appendix 2

Quality assessment results of the included studies. [PDF File (Adobe PDF File), 174 KB - mental_v11i1e48916_app2.pdf]

References


https://jment.jmir.org/2024/1/e48916

Jamr MENTAL HEALTH Shahid et al


Abbreviations

CAVE: cave automatic virtual environment
HMD: head-mounted display
MMAT: Mixed Methods Appraisal Tool
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RCT: randomized controlled trial
SAD: social anxiety disorder
VR: virtual reality
VRET: virtual reality exposure therapy

©Simon Shahid, Joshua Kelson, Anthony Saliba. Originally published in JMIR Mental Health (https://mental.jmir.org), 08.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Mental Health, is properly cited. The complete bibliographic information, a link to the original publication on https://mental.jmir.org/, as well as this copyright and license information must be included.
The Effect of Explicit Suicide Language in Engagement With a Suicide Prevention Search Page Help-Seeking Prompt: Nonrandomized Trial

Sandersan Onie1,2, PhD; Susanne Oliver Armstrong1, PhD; Natasha Josifovski1, PhD; Patrick Berlinquette3, BA; Nicola Livingstone1, BDM; Sarah Holland1, BCom; Coco Finemore1, BDS; Nyree Gale1, MCP; Emma Elder1, MEd; George Laggis1, BCom; Cassandra Heffernan1, BN; Adam Theobald1, MPsych; Michelle Torok1, PhD; Fiona Shand1, PhD; Mark Larsen1, DPhil

1Black Dog Institute, University of New South Wales, Sydney, Australia
2Faculty of Psychology, Universitas Airlangga, Surabaya, Indonesia
3BERLIN SEM, New York, NY, United States

Corresponding Author:
Sandersan Onie, PhD
Black Dog Institute
University of New South Wales
Hospital Road
Sydney, 2031
Australia
Phone: 61 432359134
Email: s.onie@blackdog.org.au

Abstract

Background: Given that signage, messaging, and advertisements (ads) are the gateway to many interventions in suicide prevention, it is important that we understand what type of messaging works best for whom.

Objective: We investigated whether explicitly mentioning suicide increases engagement using internet ads by investigating engagement with campaigns with different categories of keywords searched, which may reflect different cognitive states.

Methods: We ran a 2-arm study Australia-wide, with or without ads featuring explicit suicide wording. We analyzed whether there were differences in engagement for campaigns with explicit and nonexplicit ads for low-risk (distressed but not explicitly suicidal), high-risk (explicitly suicidal), and help-seeking for suicide keywords.

Results: Our analyses revealed that having explicit wording has opposite effects, depending on the search terms used: explicit wording reduced the engagement rate for individuals searching for low-risk keywords but increased engagement for those using high-risk keywords.

Conclusions: The findings suggest that individuals who are aware of their suicidality respond better to campaigns that explicitly use the word “suicide.” We found that individuals who search for low-risk keywords also respond to explicit ads, suggesting that some individuals who are experiencing suicidality search for low-risk keywords.

(JMIR Ment Health 2024;11:e50283) doi:10.2196/50283

KEYWORDS

suicide; suicide prevention; Google; Google Ads; internet search; explicit wording; mental health; suicidal; advertisement; advertisements; messaging; prevention signage; campaign; campaigns; distress; engagement; prompt; prompts; information seeking; help seeking; searching; search

Introduction

Over the past decade, there has been a growth in the different types of interventions for people contemplating suicide. For example, apps can help people keep themselves safe during a suicidal crisis [1]; phone booths are installed at frequently used locations, which give rapid access to a suicide hotline [2]; or an online banner containing a suicide hotline number may appear if individuals search for suicide-related terms using a search
engine [3]. Despite these help-seeking pathways, recent research has shown that less than half of individuals contemplating suicide seek professional help before a suicide attempt [4]. Thus, in parallel to developing new interventions, it may be important to understand how best to promote help seeking and access to help for individuals contemplating suicide for existing services and products.

Help-seeking prompts may function as a gateway to these interventions and come in various forms. In this case, a help-seeking prompt refers to any media encouraging service use. For example, a person may prompt an individual to see their general practitioner (GP) for their mental health, a description on an app store may prompt them to download a suicide prevention app, signage at frequently used locations may point them toward a phone booth, or text on an internet search page may encourage them to call a hotline. All these help-seeking prompts play an important role in being the first point of engagement for the individual, introducing the intervention and promoting its use. If the advertisement (ad) cannot encourage the individual to engage with the intervention, it cannot fulfill its role.

One way we can investigate how different types of prompt messaging affect engagement is by using internet search ads. Previous studies have shown that individuals may search for suicide-related terms on the internet before a suicide attempt [5], that search volumes for a particular region correspond to the suicide rate for that region [6], and that internet searches are used to seek help [7]. For example, an individual may use the internet to find the closest crisis center, a local psychiatrist, or an app to help manage their suicidal thoughts. Searching for “suicide help” yields over 1 billion results on Google, and previous studies have shown a high frequency of searches, with over 120,000 searches over a 19-day study period for suicide- and distress-related keywords in Australia alone [7]. Thus, the search page may be an ideal place to intervene and investigate what messaging is most effective for individuals contemplating suicide.

Internet search ads are triggered when keywords from a prepared list are used, presenting an ad at the top of the search results that links to a relevant page or intervention. Internet search ads also offer the ability to assess how different types of messaging perform with people in different cognitive states, reflected by their search terms. For example, individuals searching for keywords associated with suicide but not explicitly communicating suicidality (eg, loneliness, hopelessness) may engage differently with a particular ad wording than individuals searching for keywords explicitly indicating suicidality or seeking help. This allows us to assess what messaging is best for different risk or distress levels.

The reach and effectiveness of these ads are measured with the impressions (how many times an ad is shown), clicks (the number of clicks on the ad), and conversions (specific behaviors performed on the linked website). Engagement is specifically measured with the click rate (the proportion of individuals who saw the ad and then engaged with the website), the conversion rate (the proportion of individuals who clicked on the ad and then engaged with the website), and the total conversion rate (the proportion of individuals who saw the ad and then engaged with the website) [7].

A previous study has highlighted that one major discussion area regarding communicating with an individual contemplating suicide is the explicit use of the word “suicide” in the ad [8]. In one component of this study, lived experience advisors indicated that the use of the word “suicide” in ads might alienate some individuals who may indeed be experiencing thoughts of suicide but may not recognize, acknowledge, or identify their thoughts as being of suicide. The advisors elaborated that by not using the word “suicide,” we may be able to reach individuals at a precrisis phase for early intervention.

Conversely, other lived experience advisors from the same study communicated that it is imperative to be clear on the subject matter by using the word “suicide.” These advisors communicated that by being explicit, we can overcome the stigma associated with the word and the individual contemplating suicide may be more likely to engage with the service as it is specific to their needs or current situation. This is further supported by contemporary suicide first aid programs, which encourage the explicit use of the word “suicide” for the same reason [9]. Furthermore, all lived experience advisors highlighted the importance of understanding what wording is effective for different suicide risk levels to maximize engagement. For example, nonexplicit wording may be particularly effective for individuals contemplating suicide but not in a suicidal crisis, as they may not identify their thoughts being that of suicide—and vice versa.

In this study, we sought to compare engagement with two internet ad campaigns, one with explicit suicide wording in its ad and the other with nonexplicit suicide wording. We also investigated whether the pattern of engagement differed by the type of keyword searched (low risk, high risk, help seeking, or means specific). In addition, we examined engagement by gender, age, and time of day.

First, we hypothesized that the campaign with explicit wording related to suicide would perform better for individuals searching for high-risk, help-seeking, or means-specific keywords. Given that these individuals are explicitly experiencing suicidal ideation and can identify and communicate it, they may respond better to an ad that is explicit in what issue it is addressing. Second, we hypothesized that for individuals searching for low-risk-keywords, nonexplicit wording would have higher engagement as the explicit wording may alienate individuals who do not identify as having suicidal thoughts. Third, given that the key manipulation is in the ad’s wording and not the linked web pages, we hypothesized that we should see this increased engagement in the click rate but not the conversion rate (and, as such, also see increased engagement in the total conversion rate).

**Methods**

**Study Design**

The study used a 2-arm quasi-experimental design (explicit vs nonexplicit wording) with 4 initial pathways (individuals searching for different types of keywords: low risk, high risk,
help seeking, and means specific; these categories are described in more detail later).

Participants

Individuals over the age of 18 years and currently residing in Australia were included in the study. Google infers the age and gender of individuals through multiple sources of information, primarily the age and gender inputted when creating a Google account, as well as past browsing history (eg, websites visited and engagements) collected through website cookies.

Ethical Considerations

The University of New South Wales Human Research Ethics Committee approved this study (HC210827).

Figure 1. Schematic of Google Ads campaign.

Keywords

Together with lived experience advisors, researchers, and a Google Ads agent, we generated 4 lists of keywords: low-risk keywords, which included keywords people are likely to search for when in distress or situations associated with suicide, without explicitly mentioning suicide (eg, “feeling so alone,” “debt”); high-risk keywords, which included keywords explicitly communicating suicidal ideation or help seeking (eg, “I want to die”); help-seeking keywords, which included keywords explicitly searching for help for suicidal thoughts (eg, “suicide help”); and means-specific keywords, which were related to searching or using specific means [10]. When keywords are entered into Google Ads, the ad is also triggered when semantically similar search terms are entered; thus, the total

Intervention

The first arm used ads without explicit suicide wording and ran from March 2 to March 21, 2022. The second arm used ads with explicit suicide wording and ran from August 21 to August 31, 2022. The first arm was run as part of another larger study pertaining to the effectiveness of Google Ads campaigns in reaching individuals thinking of suicide and has been reported elsewhere [7]. The 2 arms of the trial were run sequentially, not concurrently. Thus, there was no randomization present. Full details of the keyword generation, ad and landing page codeign process, and content of the landing pages and linked pages can be found elsewhere [8]. A schematic of the campaign can be found in Figure 1.
number of keywords that trigger the ad is substantially larger than what is inputted.

**Advertisements**

The ads were codesigned alongside a group of lived experience advisors and investigators. The codesign process yielded 6 ads, 3 (50%) explicit suicide wording and 3 (50%) without explicit suicide wording, while controlling as much as possible for other content. Each ad was as closely matched as possible to its counterpart; that is, the first ad in the explicit condition had the same messaging and content as the first ad in the nonexplicit condition, save for explicit suicide wording in the first sentence. We developed more than 1 ad for each condition to allow our findings to be more generalizable and to reflect larger themes found in our codesign process [8] rather than specific wording.

All the ads across both conditions were controlled for the number of characters and number of words, with a range of 38-43 words and 215-234 characters. The absolute difference in the character and word count between each ad and its counterpart was between 1 and 5 characters and 1 and 3 words, respectively.

The codesigned text for the nonexplicit ads is specified next. All character counts include spaces.

### Lived Experience Hope Exchange.

Find the right support for you. A support hub designed for you by people who want to assist you to find the help you need. Hear personal stories about what has helped other people during difficult times. (237 characters, 41 words)

Looking for some support? Designed with Lived Experience. Our Hope Exchange has been designed by people who may understand how you’re feeling. There are lots of ways to seek help. We want to find the right one for you. (220 characters, 39 words)

Find the right support for you. Lived Experience Hope Exchange. We want to help you to find the help that you need and value during challenging times. Hear stories and advice from people who may have felt the way you’re feeling now. (234 characters, 42 words)

The matched text for the explicit wording ads was as follows:

### Are you feeling suicidal? Lived Experience Hope Exchange.

Support designed for you. Designed by people who want to assist you to find the help you need. Hear personal stories about what has helped other people during difficult times. (235 characters, 38 words)

Help for suicidal thoughts. Looking for some support? Designed with Lived Experience. Designed by people who may understand how you’re feeling. There are lots of ways to seek help. We want to find the right one for you. (221 characters, 38 words)

Dealing with suicidal thoughts. Find the right support for you. Lived Experience Hope Exchange. We want to help you to find the help that you need and value. Hear stories and advice from people who may have felt the way you’re feeling now. (241 characters, 43 words)

In each campaign condition, ads were shown to users independently of which category of keywords were searched. When an ad was triggered, 1 of 3 ads in that condition would be randomly shown, resulting in equal presentations across the study.

### Landing Page

In collaboration with lived experience advisors, we codesigned a series of landing pages containing lived experience stories, calming and distracting activities, and links to support services and hotlines with descriptions of what the individual will likely experience when engaging in these services. Details of the pages can be found elsewhere [8].

### Outcomes

Data on impressions, clicks, the click rate (clicks/impressions), conversions, the conversion rate (conversions/clicks), the cost per click, and the cost per conversion were extracted from Google Ads in a deidentified, aggregated form. The total conversion rate was manually calculated (conversions/impressions). Currently, the total volume of searches for each category is not available through Google Ads.

The primary outcome was the click rate (engagement with the ad) as our manipulation was on the search page rather than on the landing page. Our secondary outcomes were the total conversion rate (total engagement with the campaign, that is, all things being equal, the conversion rate per impression) and the conversion rate (engagement with the landing page).

Conversions contained behaviors the investigators, the lived experience advisors, and the collaborative team considered positive. Triggering any of these conditions was considered a conversion, including:

- Clicking the Get Help button to see available support services
- Downloading any file pertaining to the modules to help de-escalate a crisis or for self-help for suicidality
- Clicking on a link to call a support service
- Spending more than 2 minutes on the website that was designed to promote help seeking and de-escalate crises, as an indication that the individual was engaging with content on the website

### Statistical Analyses

In the main analysis, each combination of outcome metric (click rate, conversion rate, and total conversation rate) and keyword type (high risk, low risk, and help seeking) was considered separately. The outcome metric rates associated with the explicit and nonexplicit wording were compared using an incidence rate ratio (IRR) from the `rateratio` function in the `fmsb` package in R (R Foundation for Statistical Computing), which calculates the exact mid-p double-sided P value and calculates the CI using the exact Poisson method [11,12].

Interaction terms were considered if at least 1 significant difference between explicit and nonexplicit wording was identified for an outcome measure. In this case, a keyword...
category with a significant difference due to wording was compared to the other keyword categories. To test the interactions, we first calculated a difference term between the explicit and nonexplicit conditions for a single keyword type. For example, we calculated the difference term for the click rate and low-risk keywords as follows:

This difference term was compared with the other keyword difference terms using the rate_difference function in the fmsb package in R, which uses a chi-squared test to test for a significant incidence rate difference (IRD) [12]. Thus, this analysis assessed interaction by assessing a difference of a difference.

Where a significant difference of differences (ie, an interaction) was identified in the outcomes, post hoc tests were conducted to determine whether there were underlying differences in the relevant outcome metrics for the explicit or nonexplicit wording conditions or both.

Results

Campaign Metrics

A total of 153,768 impressions, 7263 clicks, and 1657 conversions were achieved during the study periods. The engagement metrics, reported by trial condition, age, and gender, are reported in Table 1.

Due to the exceptionally low numbers in the means-specific group (n=11, 0.01%, impressions; n=1, 0.01%, click; and 0 conversions), these campaigns were excluded from subsequent analyses.
Table 1. Engagement metrics by trial condition, age, and gender.

<table>
<thead>
<tr>
<th>Trial condition, gender, and age (years)</th>
<th>Impressions (N=153,768), n (%)</th>
<th>Clicks (N=7263), n (%)</th>
<th>Click rate, %</th>
<th>Conversions (N=1657), n (%)</th>
<th>Conversion rate, %</th>
<th>Total conversion rate, %</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Nonexplicit trial, gender male</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18-24</td>
<td>6191 (4.03)</td>
<td>308 (4.24)</td>
<td>4.97</td>
<td>39 (2.35)</td>
<td>12.66</td>
<td>0.63</td>
</tr>
<tr>
<td>25-34</td>
<td>6955 (4.52)</td>
<td>261 (3.59)</td>
<td>3.75</td>
<td>52 (3.14)</td>
<td>19.92</td>
<td>0.75</td>
</tr>
<tr>
<td>35-44</td>
<td>6770 (4.40)</td>
<td>313 (4.31)</td>
<td>4.62</td>
<td>77 (4.65)</td>
<td>24.60</td>
<td>1.14</td>
</tr>
<tr>
<td>45-54</td>
<td>6975 (4.54)</td>
<td>398 (5.48)</td>
<td>5.71</td>
<td>122 (7.36)</td>
<td>30.65</td>
<td>1.75</td>
</tr>
<tr>
<td>55-64</td>
<td>4644 (3.02)</td>
<td>257 (3.54)</td>
<td>5.53</td>
<td>93 (5.61)</td>
<td>36.19</td>
<td>2.00</td>
</tr>
<tr>
<td>≥65</td>
<td>2695 (1.75)</td>
<td>169 (2.33)</td>
<td>6.27</td>
<td>36 (2.17)</td>
<td>21.30</td>
<td>1.34</td>
</tr>
<tr>
<td>All ages</td>
<td>34,230 (22.26)</td>
<td>1706 (23.49)</td>
<td>4.98</td>
<td>419 (25.29)</td>
<td>24.56</td>
<td>1.22</td>
</tr>
<tr>
<td><strong>Nonexplicit trial, gender female</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18-24</td>
<td>13,243 (8.61)</td>
<td>796 (10.96)</td>
<td>6.01</td>
<td>89 (5.37)</td>
<td>11.18</td>
<td>0.67</td>
</tr>
<tr>
<td>25-34</td>
<td>17,448 (11.35)</td>
<td>708 (9.75)</td>
<td>4.06</td>
<td>135 (8.15)</td>
<td>19.07</td>
<td>0.77</td>
</tr>
<tr>
<td>35-44</td>
<td>17,996 (11.70)</td>
<td>869 (11.96)</td>
<td>4.83</td>
<td>180 (10.86)</td>
<td>20.71</td>
<td>1.00</td>
</tr>
<tr>
<td>45-54</td>
<td>18,963 (12.33)</td>
<td>1013 (13.95)</td>
<td>5.34</td>
<td>311 (18.77)</td>
<td>30.70</td>
<td>1.64</td>
</tr>
<tr>
<td>55-64</td>
<td>12,675 (8.24)</td>
<td>737 (10.15)</td>
<td>5.81</td>
<td>187 (11.29)</td>
<td>25.37</td>
<td>1.48</td>
</tr>
<tr>
<td>≥65</td>
<td>6326 (4.11)</td>
<td>398 (5.48)</td>
<td>6.29</td>
<td>98 (5.91)</td>
<td>24.62</td>
<td>1.55</td>
</tr>
<tr>
<td>All ages</td>
<td>86,651 (56.35)</td>
<td>4521 (62.25)</td>
<td>5.22</td>
<td>1000 (60.35)</td>
<td>22.12</td>
<td>1.15</td>
</tr>
<tr>
<td><strong>Explicit trial, gender male</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18-24</td>
<td>1825 (1.19)</td>
<td>62 (0.85)</td>
<td>3.40</td>
<td>4 (0.24)</td>
<td>6.45</td>
<td>0.22</td>
</tr>
<tr>
<td>25-34</td>
<td>2281 (1.48)</td>
<td>62 (0.85)</td>
<td>2.72</td>
<td>21 (1.27)</td>
<td>33.87</td>
<td>0.92</td>
</tr>
<tr>
<td>35-44</td>
<td>2052 (1.33)</td>
<td>64 (0.88)</td>
<td>3.12</td>
<td>13 (0.78)</td>
<td>20.31</td>
<td>0.63</td>
</tr>
<tr>
<td>45-54</td>
<td>1941 (1.26)</td>
<td>76 (1.05)</td>
<td>3.92</td>
<td>19 (1.15)</td>
<td>25.00</td>
<td>0.98</td>
</tr>
<tr>
<td>55-64</td>
<td>1351 (0.88)</td>
<td>54 (0.74)</td>
<td>4.00</td>
<td>15 (0.90)</td>
<td>27.78</td>
<td>1.11</td>
</tr>
<tr>
<td>≥65</td>
<td>841 (0.55)</td>
<td>46 (0.63)</td>
<td>5.47</td>
<td>8 (0.48)</td>
<td>17.39</td>
<td>0.95</td>
</tr>
<tr>
<td>All ages</td>
<td>10,291 (6.69)</td>
<td>364 (5.01)</td>
<td>3.54</td>
<td>80 (4.83)</td>
<td>21.98</td>
<td>0.78</td>
</tr>
<tr>
<td><strong>Explicit trial, gender female</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18-24</td>
<td>3436 (2.23)</td>
<td>104 (1.43)</td>
<td>3.03</td>
<td>16 (0.97)</td>
<td>15.38</td>
<td>0.47</td>
</tr>
<tr>
<td>25-34</td>
<td>4474 (2.91)</td>
<td>108 (1.49)</td>
<td>2.41</td>
<td>21 (1.27)</td>
<td>19.44</td>
<td>0.47</td>
</tr>
<tr>
<td>35-44</td>
<td>4489 (2.92)</td>
<td>140 (1.93)</td>
<td>3.12</td>
<td>30 (1.81)</td>
<td>21.43</td>
<td>0.67</td>
</tr>
<tr>
<td>45-54</td>
<td>4824 (3.14)</td>
<td>136 (1.87)</td>
<td>2.82</td>
<td>37 (2.23)</td>
<td>27.21</td>
<td>0.77</td>
</tr>
<tr>
<td>55-64</td>
<td>3381 (2.20)</td>
<td>97 (1.34)</td>
<td>2.87</td>
<td>28 (1.69)</td>
<td>28.87</td>
<td>0.83</td>
</tr>
<tr>
<td>≥65</td>
<td>1992 (1.30)</td>
<td>87 (1.20)</td>
<td>4.37</td>
<td>26 (1.57)</td>
<td>29.89</td>
<td>1.31</td>
</tr>
<tr>
<td>All ages</td>
<td>22,596 (14.69)</td>
<td>672 (9.25)</td>
<td>2.97</td>
<td>158 (9.54)</td>
<td>23.51</td>
<td>0.70</td>
</tr>
<tr>
<td><strong>Combined trial</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All ages, male</td>
<td>44,521 (28.95)</td>
<td>2070 (28.50)</td>
<td>4.65</td>
<td>499 (30.11)</td>
<td>24.11</td>
<td>1.12</td>
</tr>
<tr>
<td>All ages, female</td>
<td>109,247 (71.05)</td>
<td>5193 (71.50)</td>
<td>4.75</td>
<td>1158 (69.89)</td>
<td>22.30</td>
<td>1.06</td>
</tr>
</tbody>
</table>

Click Rate

The click rate using explicit and nonexplicit wording for each keyword category is shown in Figure 2. A significant difference between explicit and nonexplicit wording for low-risk keywords was found (IRR=1.848, 95% CI 1.718-1.987, P<.001), in which there was a higher click rate for nonexplicit versus explicit keywords (5.11% vs 2.77%). For help-seeking keywords, there was a significantly higher click rate for explicit versus nonexplicit keywords (3.50% vs 4.93%; IRR=0.715, 95% CI 0.599-0.854, P<.001). A similar pattern was observed for high-risk keywords, with a higher click rate for explicit versus nonexplicit keywords, although this was only marginally nonsignificant (6.33% vs 8.85%; IRR=0.711, 95% CI 0.503-1.005, P=.052).
Next, we investigated the presence of 2-way interactions between the 3 groups (low risk and high risk, low risk and help seeking, high risk and help seeking). All 2-way interactions were significant (high risk and help seeking: IRD=0.0109, 95% CI 0.000146-0.0217, $P=.047$; high risk and low risk: IRD=0.0301, 95% CI 0.0301-0.0453, $P<.001$; low risk and help seeking: IRD=0.0472, 95% CI 0.0405-0.0541, $P<.001$).

Post hoc tests revealed a significant difference in the click rate between the low- and high-risk keywords in the explicit condition (2.77% vs 4.93%; IRR=0.5612, 95% CI 0.4144-0.7784, $P<.001$) and the nonexplicit condition (5.11% vs 3.50%; IRR=1.459, 95% CI 1.216-1.766, $P<.001$). Significant differences were also found in the click rate between the low-risk and help-seeking keywords in the explicit condition (2.77% vs 8.85%; IRR=0.312, 95% CI 0.263-0.373, $P<.001$) and the nonexplicit condition (5.11% vs 6.33%; IRR=0.807, 95% CI 0.740-0.881, $P<.001$).

Conversion Rate
A graphical representation of the conversion rate data is shown in Figure 3. There were no significant differences between the nonexplicit and explicit conditions (low risk: 21.86% vs. 21.17%; IRR=1.0326, 95% CI 0.882-1.209, $P=.69$; high risk: 31.36% vs 43.18%; IRR=0.726, 95% CI 0.418-1.263, $P=.25$; help seeking: 29.74% vs 25.16%; IRR=1.182, 95% CI 0.835-1.674, $P=.55$). Thus, interaction effects were not explored.
Total Conversion Rate

A graphical representation of the total conversion rate can be seen in Figure 4. There was a significant difference in the total conversion rate for low-risk keywords (IRR=1.908, 95% CI 1.630-2.234, \(P<.001\)), in which the nonexplicit wording had a higher rate (1.12% vs 0.59%), and high-risk keywords (IRR=0.846, 95% CI 0.297-0.896, \(P=.02\)), in which the explicit wording had a higher rate (1.10% vs 2.13%); however, there was not enough evidence to suggest a difference between explicit and nonexplicit wording (1.88% vs 2.23%) when an individual was searching for help-seeking keywords (IRR=0.846, 95% CI 0.597-1.197, \(P=.34\)).

Possible interactions between the 3 groups (low risk and high risk, low risk and help seeking, high risk and help seeking) were explored. A significant interaction was identified between low-risk and high-risk keywords (IRD=0.00875, 95% CI 0.00613-0.0114, \(P<.001\)), as well as between low-risk and help-seeking keywords (IRD=0.0156, 95% CI 0.00901-0.0222, \(P<.001\)). However, there was no significant interaction between high-risk and help-seeking keywords (IRD=-0.00685, 95% CI -0.0141 to 0.00345, \(P=.06\)).

Post hoc tests revealed a significant difference in the total conversion rate between the low- and high-risk keywords in the explicit condition (0.59% vs 2.13%; IRR=0.275, 95% CI 0.171-0.442, \(P<.001\)) but not in the nonexplicit condition (1.12% vs 1.10%; IRR=1.017, 95% CI 0.733-1.411, \(P=.92\)).

A summary of the analysis outcomes can be seen in Table 2.

Figure 4. Total conversion rate by keyword type and condition. *\(P<.05\).
### Table 2. Summary of analysis outcomes.

<table>
<thead>
<tr>
<th>Metric and keyword type</th>
<th>Significant comparisons</th>
<th>Significant interactions</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Click rate</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low risk</td>
<td>Nonexplicit &gt; explicit</td>
<td>Low risk × high risk</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Low risk × help seeking</td>
</tr>
<tr>
<td>High risk</td>
<td>Explicit &gt; nonexplicit</td>
<td>Low risk × high risk</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Low risk × help seeking</td>
</tr>
<tr>
<td>Help seeking</td>
<td>Explicit &gt; nonexplicit</td>
<td>Low risk × high risk</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Low risk × help seeking</td>
</tr>
<tr>
<td></td>
<td></td>
<td>High risk × help seeking</td>
</tr>
<tr>
<td><strong>Conversion rate</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low risk</td>
<td>N/S&lt;sup&gt;b&lt;/sup&gt;</td>
<td>—</td>
</tr>
<tr>
<td>High risk</td>
<td>N/S</td>
<td>—</td>
</tr>
<tr>
<td>Help seeking</td>
<td>N/S</td>
<td>—</td>
</tr>
<tr>
<td><strong>Total conversion rate</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low risk</td>
<td>Nonexplicit &gt; explicit</td>
<td>Low risk × high risk</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Low risk × help seeking</td>
</tr>
<tr>
<td>High risk</td>
<td>Explicit &gt; nonexplicit</td>
<td>Low risk × high risk</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Low risk × help seeking</td>
</tr>
<tr>
<td>Help seeking</td>
<td>N/S</td>
<td>Low risk × high risk</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Low risk × help seeking</td>
</tr>
</tbody>
</table>

<sup>a</sup>Marginal significance (P<.06).

<sup>b</sup>N/S: not significant.

<sup>c</sup>Not available. This was used when interaction analyses were not conducted due to nonsignificant comparisons.

### Discussion

#### Principal Findings

In this study, we compared the impact of explicit and nonexplicit suicide wording in an online ad campaign and webpage targeting those searching for suicide- and distress-related keywords. Analysis of the click rate revealed that for low-risk keywords, nonexplicit wording had a higher click rate; for high-risk keywords, there was marginal evidence that explicit wording had a higher click rate; and for help-seeking keywords, explicit wording had a higher click rate. For the conversion rate, there was no evidence of any differences between conditions. Analysis of the total conversion rate revealed that for low-risk keywords, nonexplicit wording had a higher total conversion rate; for high-risk keywords, explicit wording had a higher total conversion rate; and for help-seeking keywords, there was no evidence of a difference.

Further analysis of the click rate revealed 2-way interactions between low- and high-risk keywords, low-risk and help-seeking keywords, and high-risk and help-seeking keywords, suggesting that the effect of explicit wording in the ad differed among these groups. Further exploratory analysis revealed significant differences between low-risk and high-risk keywords, as well as between low-risk and help-seeking keywords, in the explicit and nonexplicit conditions. These findings further support the suggestion that the effect of wording differentially impacts engagement in keyword groups, rather than an interaction emerging due to the manipulation only affecting one group but not the other. Together, these findings suggest that ads with explicit suicide language are less likely to be clicked on than those with nonexplicit language when individuals are searching for low-risk keywords. The reverse was observed when searching for high-risk or help-seeking keywords (although the former did not reach significance), where ads with explicit language were more likely to be clicked on. This pattern of findings may partly contribute to the pattern of findings in the total conversion rate, in which analysis revealed 2-way interactions between low- and high-risk keywords and between low-risk and help-seeking keywords, suggesting that the effect of explicit wording in the ad differed among these groups.

Exploratory analysis revealed that the interaction between low- and high-risk keywords may be driven by a higher total conversion rate when explicit wording is used with high-risk versus low-risk keywords, whereas there was no apparent difference between high-risk and low-risk keywords when nonexplicit wording was used. These findings suggest that having explicit wording has opposite effects, depending on the search terms used; explicit wording reduces the total conversion rate for individuals searching for low-risk keywords but increases the total conversion rate for those using high-risk keywords. There is no evidence to suggest that explicit or...
nonexplicit wording affects the total conversion rate when help-seeking keywords are used. These findings support both recommendations from lived experience advisors, both for and against the use of explicit wording, as perhaps both are appropriate for individuals in different cognitive states.

Overall, these findings suggest that individuals who search for help-seeking or high-risk suicide keywords respond more to ads and campaigns with explicit suicide wording, demonstrated by the higher click rate and total conversion rate, respectively. Alternatively stated, an ad campaign targeting individuals searching for high-risk keywords is likely to lead to more desirable behaviors if explicit wording is used. This may be because individuals may respond more strongly to a campaign that specifically targets their current situation. Given that the ad explicitly communicates and labels their current issue, the individuals may be more inclined to seek help. As the nonexplicit ad contained no indication that the campaign is for suicidality, the users may not have been certain that the webpage was able to meet their needs.

Furthermore, the high-risk search terms in this study excluded individuals who were explicitly searching for help for their suicidality. Thus, our findings suggest that explicitly naming the issue can improve help seeking for individuals at high risk of suicide but who may not be actively seeking help, as revealed in the total conversion rate. This finding is consistent with current practices in suicide prevention first aid (eg, applied suicide prevention skills training [9]), where directly addressing and asking about suicide are strongly encouraged. This study demonstrated that the benefit of explicit wording for individuals in crisis is generalizable beyond suicide first aid and direct face-to-face communication. This finding may have implications for other forms of communication when addressing individuals at high risk of suicide.

The results also showed that individuals searching for low-risk suicide terms respond less to campaigns with explicit suicide wording. Since the low-risk keywords in this study were broad (eg, loneliness), many individuals searching for these terms may not have been suicidal—hence the lower click rate. However, despite being low, we still observed a click rate by individuals searching for low-risk keywords on the explicit campaign compared to the industry standard of 3.17% [13]. This suggests that individuals who may not be searching for explicit keywords may be experiencing suicidality and that targeting low-risk keywords is still beneficial. Conversely, we may have seen an elevated click rate for the low-risk, nonexplicit condition due to people clicking on the ad not realizing it was for individuals experiencing suicidality. Nevertheless, in the low-risk, nonexplicit condition, we still observed a conversion rate of over 20%, relative to the industry standard of 3.75%, suggesting that the landing page was still fulfilling a need. This may be because by using explicit suicide wording, we may alienate individuals who, for various reasons, may not recognize, identify, or acknowledge that their feelings are those of suicide. Thus, when visiting the page, their needs are met. However, one possibility is that some individuals searching for low-risk keywords are unaware that they are suicidal, and using the word “suicide” may help bring awareness to these underlying feelings [8]. Thus, explicit and nonexplicit keywords may have their own benefits; however, our findings suggest that overall, using nonexplicit keywords will reach more people. Although many people searching for low-risk keywords may not be experiencing suicidality, we must ensure that little effort is needed to access suicide-related resources and help, given that there are still individuals searching for low-risk keywords who are experiencing suicidality. Further research is needed to understand how best to tailor the ads to individuals searching for low-risk keywords.

The findings suggest that individuals experiencing suicidality and who could explicitly communicate it have higher engagement patterns with a campaign when the word “suicide” is used in the ad regardless of whether they are explicitly seeking help. However, for individuals experiencing general distress but not searching for suicide-specific terms, using explicit suicide wording leads to lower engagement with the campaign. Thus, in response to the finding that some lived experience advisors advocated for the explicit use of the word “suicide,” while others advocated against it, perhaps both are true for individuals in different cognitive states.

We recommend that the development and design of help-seeking prompts for suicide consider at what stage the individual is. If the prompt is intended for individuals with a lower risk of suicide, such as public media campaigns, then the use of the word “suicide” may decrease engagement; however, if the prompt is intended for individuals at high risk of suicide, whether they are or are not actively seeking help, then the use of the word “suicide” is likely to increase engagement. Furthermore, these findings suggest the need to codesign with a range of individuals who have experienced the spectrum of suicidality to understand their needs, the thought process, and how they speak about and internally conceptualize their distress and suicide to formulate different terms that promote help seeking and engagement.

Given that this pattern of finding has been found across 2 modalities (suicide first aid and internet ads), future research should seek to investigate the generalizability of these findings to other help-seeking prompts in suicide prevention, for example, signage at frequently used locations, safety planning app notifications, or the wording on the suicide hotline banner if individuals search for suicide-related terms. Further, future research should further understand what type of ad wording works for whom. For example, previous evidence has suggested that men and women respond differently to tailored ad campaigns [14]. Using Google Ads, we can further investigate what type of ad best engages men and women.

There was a low number of searches and engagement for means-specific keywords in this study. There may be several possibilities for this. First, these numbers may reflect true rates and only a few individuals were searching for these terms. Another possibility is that individuals at the planning stage do not primarily turn to search engines but may use other means of information seeking, as means selection has been found to be influenced by prior familiarity with the means itself [15]. Another possibility is that more individuals were searching for these terms but our current keyword list and Google’s function of generating permutations of the keyword list could not capture...
the range of search terms. This could be rectified by experimenting with keyword setting iterations in Google Ads to capture a wider variety of expressions. Finally, one possibility is that there are cultural differences in search behaviors. For example, a previous study found high incidences of method-related searches in Japan [16], suggesting that cultural differences may exist. Given that a key finding in previous research was that individuals who attempted suicide had searched for means-specific keywords [5], future research should investigate how individuals search for information regarding means and how we can best intervene at this stage.

Future studies may also investigate whether these findings generalize to other psychological or health domains. For example, an individual who is acutely depressed and can identify it may respond better to an ad for therapy explicitly communicating that this is for individuals experiencing depression. Conversely, an individual who may also be experiencing depressive symptoms but is unfamiliar that they are experiencing depressive symptoms or is from a background where depression is strongly stigmatized may respond better to an ad without explicitly using the word “depression.” Future studies should carefully consider conversion actions, as we cannot measure directly whether an individual is processing the information presented to them, and thus, we use proxy measures common to the marketing field (eg, time spent on a page). Given the growing interest in online interventions, standardized methods for measuring engagements should be established.

Limitations
The study has several limitations. One limitation is that due to resourcing constraints, there was a large difference in the number of participants in the 2 conditions. Furthermore, the data were collected at different times of the year, so there may be seasonal or cohort effects; future studies should run campaigns with explicit and nonexplicit wording at the same time. In addition, as research with this type of data is still in its infancy, future research should focus on understanding whether, when, and how search metrics should be normalized against variations in time. Another limitation is that we could only infer cognitive states from search results but did not measure suicidality directly. For example, a person may search for a high-risk keyword as part of a study but not be suicidal. Thus, there is added noise in the data.

Strengths
Our study also has several strengths. The study was run nationwide, allowing us to sample the entire target population rather than just a specific subset. The components of this study, such as the ad wording, landing page, and keywords used to trigger the ad, were co-designed with individuals with lived and living experiences of suicide. Furthermore, by providing data on click rates and conversion rates, we obtained greater mechanistic insight into the findings for the total conversion rate.

Conclusion
Our study demonstrates different engagement levels with an online suicide prevention campaign due to the word “suicide” in the search page ad. Future research should further explore what type of messaging works best for whom, and when paired with the flexibility of the advertising industry, we may be one step closer to ensuring that each person is met with a message that leads to the highest probability of them engaging services, using resources, or seeking help.

Acknowledgments
SO led the conceptualization, methodology, data curation, formal analysis, funding acquisition, project administration, and writing—original draft. SOA co-led writing—original draft. ML led supervision and co-led funding acquisition, formal analysis, and writing—review and editing. MT and FS co-led funding acquisition and writing—review and editing. PB led resources and software. All other authors supported the methodology and investigation. This study was supported by the Australian Government Department of Health–funded National Suicide Prevention Research Fund, managed by Suicide Prevention Australia, National Health and Medical Research Council (NHMRC) Centre of Research Excellence in Suicide Prevention (APP1152952).

Conflicts of Interest
None declared.

References


Google Ads co-design. Center for Open Science. URL: https://osf.io/d6m45/ [accessed 2024-02-24]


Irvine M. Google Ads benchmarks for your industry [updated!]. WordStream. 2024. URL: https://www.wordstream.com/blog/ws/2016/02/29/google-adwords-industry-benchmarks [accessed 2024-02-23]


Abbreviations

- ad: advertisement
- IRD: incidence rate difference
- IRR: incidence rate ratio

©Sandersan Onie, Susanne Oliver Armstrong, Natasha Josifovski, Patrick Berlinguette, Nicola Livingstone, Sarah Holland, Coco Finemore, Nyree Gale, Emma Elder, George Laggis, Cassandra Heffernan, Adam Theobald, Michelle Torok, Fiona Shand, Mark Larsen. Originally published in JMIR Mental Health (https://mental.jmir.org), 19.03.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Mental Health, is properly cited. The complete bibliographic information, a link to the original publication on https://mental.jmir.org/, as well as this copyright and license information must be included.
Action Opportunities to Pursue Responsible Digital Care for People With Intellectual Disabilities: Qualitative Study

Nienke M Siebelink¹, PhD; Kirstin N van Dam¹,², MSc; Dirk R M Lukkien³,⁴, MSc; Brigitte Boon¹,²,⁵, PhD; Merlijn Smits⁶, PhD; Agnes van der Poel¹, PhD

¹Academy Het Dorp, Arnhem, Netherlands
²Tranzo, Tilburg School of Social and Behavioral Sciences, Tilburg University, Tilburg, Netherlands
³Vilans, Utrecht, Netherlands
⁴Copernicus Institute of Sustainable Development, Utrecht University, Utrecht, Netherlands
⁵Siza, Arnhem, Netherlands
⁶PBLQ, The Hague, Netherlands

Corresponding Author:
Nienke M Siebelink, PhD
Academy Het Dorp
Kemperbergerweg 139e
Arnhem, 6816RP
Netherlands
Phone: 31 0651656387
Email: nienke.siebelink@academyhetdorp.nl

Abstract

Background: Responsible digital care refers to any intentional systematic effort designed to increase the likelihood of a digital care technology developed through ethical decision-making, being socially responsible and aligned with the values and well-being of those impacted by it.

Objective: We aimed to present examples of action opportunities for (1) designing “technology”; (2) shaping the “context” of use; and (3) adjusting the behavior of “users” to guide responsible digital care for people with intellectual disabilities.

Methods: Three cases were considered: (1) design of a web application to support the preparation of meals for groups of people with intellectual disabilities, (2) implementation of an app to help people with intellectual disabilities regulate their stress independently, and (3) implementation of a social robot to stimulate interaction and physical activity among people with intellectual disabilities. Overall, 26 stakeholders participated in 3 multistakeholder workshops (case 1: 10/26, 38%; case 2: 10/26, 38%; case 3: 6/26, 23%) based on the “guidance ethics approach.” We identified stakeholders’ values based on bottom-up exploration of experienced and expected effects of using the technology, and we formulated action opportunities for these values in the specific context of use. Qualitative data were analyzed thematically.

Results: Overall, 232 effects, 33 values, and 156 action opportunities were collected. General and case-specific themes were identified. Important stakeholder values included quality of care, autonomy, efficiency, health, enjoyment, reliability, and privacy. Both positive and negative effects could underlie stakeholders’ values and influence the development of action opportunities. Action opportunities comprised the following: (1) technology: development of the technology (eg, user experience and customization), technology input (eg, recipes for meals, intervention options for reducing stress, and activities), and technology output (eg, storage and use of data); (2) context: guidelines, training and support, policy or agreements, and adjusting the physical environment in which the technology is used; and (3) users: integrating the technology into daily care practice, by diminishing (eg, “letting go”) to increase the autonomy of people with intellectual disabilities), retaining (eg, face-to-face contact), and adding (eg, evaluation moments) certain behaviors of care professionals.

Conclusions: This is the first study to provide insight into responsible digital care for people with intellectual disabilities by means of bottom-up exploration of action opportunities to take account of stakeholders’ values in designing technology, shaping the context of use, and adjusting the behavior of users. Although part of the findings may be generalized, case-specific insights and a complementary top-down approach (eg, predefined ethical frameworks) are essential. The findings represent a part of an ethical discourse that requires follow-up to meet the dynamism of stakeholders’ values and further develop and implement action opportunities.
opportunities to achieve socially desirable, ethically acceptable, and sustainable digital care that improves the lives of people with intellectual disabilities.

(International Journal of Mental Health 2024;11:e48147) doi:10.2196/48147

KEYWORDS
ethics; value-based health care; digital technology; intellectual disability; digital care

Introduction

Digital Care

As digital tools have shown great potential to enhance health care and well-being services, digital care plays a central role in the policies and plans of governments and care organizations to continue to provide good care efficiently [1-3]. Digital care refers to technology and data that inform and improve health care provision [4]. Unfortunately, today, digital care is often not aligned with the needs and values of its users and other stakeholders [5,6]. Not aligning digital care with stakeholders' needs and values results in low technology uptake. Although the importance of involving all relevant stakeholders in digital care innovation is widely accepted [7,8], they are insufficiently involved and, often, involved very late in digital care innovation [9]. Consequently, time and effort are wasted [10], and the clinical appropriateness and usability of digital care are compromised [11].

Digital care is found to have a significant impact on people’s lives, especially for those with intellectual disabilities who often receive life-long care. For people with intellectual disabilities, technologies are not only applied to promote health but also to enhance independence and quality of life, such as being able to participate in the society, which creates more educational, vocational, and leisure opportunities [12-14]. In long-term care organizations, integrating technology in daily practice for people with intellectual disabilities requires insight into the needs and values of the people with intellectual disabilities themselves, as well as those of their care professionals who guide them through daily life and need to adapt their guiding strategies, the IT support staff of the care organization who provide technical support for the digital solutions, the human resources professionals who need to integrate the technology in their regular training programs within the care organization, and the data specialists who need to make decisions about incorporating the data provided by the digital care technology. Considering the increasing influence of digital care on several domains of the life of people with intellectual disabilities [13], the design and implementation of technologies should be well considered.

Responsible Design and Implementation

Ideally, the design and implementation of digital care for people with intellectual disabilities are “responsible”: to include any intentional systematic effort designed to increase the likelihood of a digital care technology developed through ethical decision-making, being socially responsible and aligned with the values and well-being of those influenced by it [15]. Values, defined as “convictions or matters that people feel should be strived for in general and not just for themselves to be able to lead a good life or realise a good society” [16], are commonly considered within ethical discourse. These values function as moral compasses that guide certain actions, for example, in the design and implementation of technology. However, the ethics of digital care is not a common subject of study [17,18]. There is limited empirical evidence describing how to address stakeholders’ values within their context, in this case, the context of long-term care for people with intellectual disabilities, even though it is broadly recognized that responsible design and implementation of digital care require insight into and sensitivity toward specific contexts of use [17,19,20]. There is a need for context-specific studies about how certain values matter to the stakeholders of particular technologies and how these values can be accounted for in technology design and implementation.

Guidance Ethics

The “guidance ethics approach” [21] is a relatively new method for reflection about and guidance for the responsible design and implementation of technologies in the context of use, developed by the ECP (Platform for the Information Society), the Netherlands. The approach is applied in various fields, such as municipalities, government, security, police, and health care sector, and regarding various cases, such as a biofeedback app for people with profound intellectual and multiple disabilities and challenging behavior or the use of artificial intelligence for nighttime monitoring in disability care [22]. The guidance ethics method involves a multistakeholder workshop, in which stakeholders’ values are identified based on an exploration of the experienced and expected positive and negative effects of using a specific technology. Subsequently, workshop participants mutually formulate action opportunities to account for these values in the specific context of use. In this study, we used this method to identify stakeholders’ values and formulate action opportunities for the responsible design and implementation of specific technologies used in long-term care for people with intellectual disabilities.

The guidance ethics approach—more extensively described in the Methods section—has several advantages compared with other research methods focused on ethics in design. One of the most well-known methodologies considering ethics through values in design is “Value Sensitive Design” [23]. Although the idea of embedding values in technology originated from this method, it does not provide the tools to empirically study values. However, guidance ethics is practical and hands-on, allowing stakeholders in a workshop to contribute to identifying the values affected by the use of technology. There are tools, such as the “interactive technology assessment” [24], that also provide hands-on tools, but these solely focus on studying the values of 1 user. Guidance ethics enables to involve a diverse group of stakeholders to identify a comprehensive set of values [7] and facilitates stakeholders to better understand the position of others [25]. Although there are methods, such as an
evidence-informed, deliberative process approach to a health technology assessment [7,26] that involves multiple stakeholders also, this method, in contrary to guidance ethics, does not translate insights into concrete action opportunities for responsible technology use. To the best of our knowledge, guidance ethics is the only method that enables the study of values involving multiple stakeholders and directly translates these into concrete action opportunities.

Objective

In this study, we applied the guidance ethics approach to three digital care technologies that are currently being developed or implemented within care organizations for people with intellectual disabilities:

1. The design of “Kookapp for groups” (developed by care organizations Amerpoort and Reinaerde and IT company Ilionx, Utrecht): this is a web application to support group workers with the preparation of healthy and tasty meals for groups of people with intellectual disabilities, from choosing recipes and buying ingredients to cooking and serving the meals.

2. The implementation of the “SignaLEREN” app (developed by care organization Koraal and IT company Ivengi, Maastricht): this app is used by people with intellectual disabilities or autism spectrum disorder to regularly gauge their emotional state; in the case of increased stress, they can choose a personalized stress-reducing activity within the app, such as watching a video clip or listening to certain music.

3. The implementation of SARA (developed by SARA Robotics, Eindhoven): this is a social robot that provides activities (eg, exercises, games, and music) during day care to stimulate interaction and physical activity among older people with intellectual disabilities.

With these 3 cases, we aimed to present examples of action opportunities to guide the responsible use of digital care for people with intellectual disabilities.

Methods

Participants

The 4 care organizations of the 3 cases participated in the Innovation Impulse Disability Care, a 3-year program initiated by the Dutch Ministry of Health, Welfare, and Sport in 2019. This program aimed to accelerate digital transformation in long-term care by providing support in implementing technology in the everyday practice of 26 disability care organizations [27]. In each organization, the implementation started by defining a topical care issue from the perspective of and together with people with a disability, followed by the selection of a technology that contributed to the solution of this care issue. In addition, organizations evaluated their IT and organizational readiness to implement the selected technology [28,29]. The care issues included, for example, improving day structure [30] or sleep-wake patterns [31], lowering stress levels, and increasing independent living [32]. Digital care technologies included sensors, domotics, social robotics, and apps. The Innovation Impulse program also entailed researching the factors influencing the implementation (NM Siebelink, unpublished data, 2024).

All 26 care organizations were invited to apply for participation in this study. Guidance ethics workshops were conducted in 4 care organizations, for the 3 cases described previously. Project leaders of the Innovation Impulse program within each care organization invited a purposefully diverse group of workshop participants from their organizations, for example, people with intellectual disabilities, relatives, care professionals, policy advisors, managers, members of the board of directors, IT staff, and technology developers.

In total, 26 individuals participated in this study. Participants’ characteristics are presented in Table 1. No personal data such as sex or age were collected. Almost all participants had some knowledge about the specific technology: of the 26 participants, 13 (50%) considered themselves informed, 9 (35%) had practical experience with the technology, and 4 (15%) were unfamiliar with the respective technology before participating in the workshop.

<table>
<thead>
<tr>
<th>Participant</th>
<th>Case 1: Kookapp for groups (n=10), n (%)</th>
<th>Case 2: SignaLEREN app (n=10), n (%)</th>
<th>Case 3: social robot, SARA (n=6), n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>People with intellectual disabilities (or a representative)</td>
<td>2 (20)</td>
<td>0 (0)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Management or policy maker</td>
<td>2 (20)</td>
<td>1 (10)</td>
<td>1 (17)</td>
</tr>
<tr>
<td>IT staff or technology developer</td>
<td>1 (10)</td>
<td>2 (20)</td>
<td>2 (33)</td>
</tr>
<tr>
<td>Care professional or team leader</td>
<td>4 (40)</td>
<td>3 (30)</td>
<td>2 (33)</td>
</tr>
<tr>
<td>Other (eg, project leader or consultant)</td>
<td>1 (10)</td>
<td>4 (40)</td>
<td>1 (17)</td>
</tr>
</tbody>
</table>

Procedure and Materials

This study had a qualitative research design, using guidance ethics workshops to collect data. In total, three 3.5-hour multiple stakeholder workshops were conducted by trained workshop leaders from ECP (2 per workshop). The workshops were attended in person in May, June, and September 2022. Data were collected by means of a questionnaire (described in this section) completed on paper by the participants during the workshop. In addition, the information that the workshop leaders
wrote on the flip charts was collected by taking photographs of the flip charts. In total, 2 researchers (KNvD and NMS or AvdP) were present during each workshop to observe and explain the study and the questionnaire; they did not engage in the workshop.

The questionnaire—constructed by researchers (NMS and KNvD) for this study—followed the workshop outline (Figure 1 [21,33]). In stage 1 of the workshop (case), the project leader presented the case, that is, information about the technological solution, its aim, the way it works, for which target group, and in which daily (care) process. Thereafter, data about the participants’ characteristics and their familiarity with the technology were collected.

Figure 1. Outline of the guidance ethics approach (adapted from Verbeek and Tijink 2020 [21], which is published under Creative Commons Attribution 4.0 International License [33]).

In stage 2 (dialogue), participants were first asked to call out all actors that are or should be affected by or involved with the use of the technology; the workshop leaders wrote these actors on a flip chart. Second, participants were asked to write down any positive and negative effects of the technology they could think of, not only from their own perspective but also any effect that came to mind. Next, all of them were asked to mention an effect until all effects were called out. Again, the workshop leader wrote these effects on a flip chart, and the effects were discussed, supplemented, and clustered by the workshop leaders and participants. Third, the workshop leaders identified values based on the clustered effects, and these values were adjusted in discussion with the participants. Finally, in stage 2, each participant determined the top 3 values that they deemed most important for their professional role in the particular case. These values were marked on the flip chart and discussed, after which the top 3 values of the total group were determined.

For stage 3 (action opportunities), participants were divided into 3 subgroups with diverse stakeholders in each subgroup. These subgroups were invited to come up with action opportunities to achieve a highly value-driven use of the respective technology in the context of the specific case, using the top 3 values of the group as starting point. A slight deviation from the protocol was that the group of the third case (social robot, SARA)—which was relatively small—was not divided into subgroups in stage 3 and did not explicitly focus on the top 3 values. The subgroups were instructed to come up with action opportunities for the following (respectively): the design of the technology (technology), shaping the environment or context of use (context), and adjusting the behavior of the users (users). Workshop leaders explained the meaning of “action opportunities” by using the example of the technology “car.” Driving a car should be safe (value); therefore, cars have seat belts and automatic brakes (action opportunities for the technology to improve safety), traffic lights and other infrastructure guide drivers (action opportunities for the environment to improve safety), and drivers practice driving and learn the rules before receiving a license (action opportunities for the user’s behavior to improve safety).

Participants were asked to write down the action opportunities that came to mind, which were then discussed and collected on a flip chart in the subgroups. Each subgroup presented their action opportunities, which were discussed plenarily. At the end of the workshop, participants were asked to write down in the questionnaire any new effects, values, or action opportunities that came to mind that were not mentioned during the workshop.

Ethical Considerations

Participants were informed about the study and privacy statement, and they provided consent by completing the questionnaire anonymously. The local Medical Research Ethics Committee Oost-Nederland deemed the research in the Innovation Impulse program not subject to the Medical Research Involving Human Subjects Act (“Wet medisch-wetenschappelijk onderzoek met mensen”; file 2021-8293).
Analyses

Data from the questionnaires about positive and negative effects, values, and action opportunities were entered in Excel (Microsoft Corporation; 2018). The data set was checked for completeness using notes of the workshop observations, pictures of the flip charts from the workshops, and the workshop reports made by ECP’s workshop leaders. Analyses were conducted using a bottom-up approach: participants’ descriptions were the starting point leading to the derivation of themes. First, 2 researchers (KNvD and NMS) independently derived themes from the “effects data” per case. That is, effects regarding a similar subject were given a descriptive name. For example, “joyous end users” and “end users can experience more enjoyment” were named “enjoyment of end users,” which was then considered an effect theme. The 2 researchers compared and discussed their effect themes until consensus was reached. Next, the effect themes of all 3 cases were written on digital Post-it notes and visually arranged, so that related effect themes were near each other. Digital Post-it notes on which values were presented were added for each theme from which the values were abstracted. Furthermore, 2 researchers (KNvD and NMS) also derived themes from the action opportunity data per case. Next, analyses and discussions were conducted regarding which values were represented by the action opportunity themes. For example, the action opportunity theme “Give the person with intellectual disability some self-direction in the use of the technology” is mainly related to the value “autonomy,” whereas the action opportunity theme “keep the goal in mind and deploy technology as a means rather than a goal in itself” is related to the value “quality of care.” The preliminary results were presented and discussed in an interpretation meeting with workshop leaders and project leaders from the care organizations (participants from all 3 workshops were present). Input and feedback from this meeting were used for further analyses through an iterative process.

Results

Overview

The 3 workshops provided insight into how effects were translated into values and, subsequently, how values were translated into action opportunities for technology, context, and users of a specific technology. The numbers of collected effects, values, and action opportunities for each of the 3 cases are presented in Table 2. An overview of their content is provided in Multimedia Appendix 1; for readability, effects and action opportunities were shortened, and similar ones were combined in the overview.

Data about values and action opportunities were missing from a participant who could only attend the first part of the workshop. Furthermore, the person with intellectual disability from case 1 formulated effects and action opportunities together with a care professional.

Table 2. Number of participants per case and the amount of data collected.

<table>
<thead>
<tr>
<th>Case</th>
<th>Participants (n=26), n (%)</th>
<th>Effects (n=232), n (%)</th>
<th>Values (n=33), n (%)</th>
<th>Action opportunities (n=156), n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1—Kookapp for groups</td>
<td>10 (38.5)</td>
<td>102 (43.9)</td>
<td>13 (39.4)</td>
<td>74 (47.4)</td>
</tr>
<tr>
<td>2—SignaLEREN app</td>
<td>10 (38.5)</td>
<td>66 (28.4)</td>
<td>7 (21.2)</td>
<td>54 (34.6)</td>
</tr>
<tr>
<td>3—Social robot, SARA</td>
<td>6 (23.1)</td>
<td>64 (27.6)</td>
<td>13 (39.4)</td>
<td>28 (17.9)</td>
</tr>
</tbody>
</table>

Examples of Action Opportunities

Table 3 presents examples of action opportunities for (1) designing “technology”; (2) shaping the “context” of use; and (3) adjusting the behavior of the “users” to guide the responsible use of digital care for people with intellectual disabilities. Given that describing all results (which can be found in Multimedia Appendix 1) is beyond the scope of this paper, Table 3 highlights 1 example per case based on one of the most prominent values in that case, and we have described the effects and action opportunities linked to that value. Following the examples, we have reflected about general observations within and overarching the 3 cases.
Table 3. Examples of combinations of effects, a value, and action opportunities for the 3 cases.

<table>
<thead>
<tr>
<th>Cases and categories</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Case 1: Kookapp for groups—a web application to support healthy cooking for groups of people with intellectual disabilities</strong></td>
<td></td>
</tr>
<tr>
<td>Selected value</td>
<td>• Quality of care</td>
</tr>
<tr>
<td>Positive effects</td>
<td>• Connectedness through choosing, cooking, and eating together</td>
</tr>
<tr>
<td></td>
<td>• Continuity of meal quality; not being dependent on care professionals’ skills</td>
</tr>
<tr>
<td></td>
<td>• Awareness of the importance of good nutrition</td>
</tr>
<tr>
<td></td>
<td>• More time for the people with intellectual disabilities</td>
</tr>
<tr>
<td></td>
<td>• Equality (differences regarding meals for people with intellectual disabilities between care organizations become small when they use the web application)</td>
</tr>
<tr>
<td>Negative effects</td>
<td>• Excessive focus on health compared with enjoying tasty food</td>
</tr>
<tr>
<td>Action opportunities—technology</td>
<td>• None mentioned regarding the selected value</td>
</tr>
<tr>
<td>Action opportunities—context (care organization)</td>
<td>• Evaluate efficiency, health, and eating pleasure continuously</td>
</tr>
<tr>
<td></td>
<td>• Provide the care professionals with instructions about how to use the web application along with a manual</td>
</tr>
<tr>
<td>Action opportunities—users (care professionals)</td>
<td>• Invest in understanding the web application to use it properly</td>
</tr>
<tr>
<td></td>
<td>• Have a backup plan for situations in which the web application does not work</td>
</tr>
<tr>
<td></td>
<td>• Know the dietary preferences and needs of each person with intellectual disability in the group</td>
</tr>
<tr>
<td></td>
<td>• Know what to do when a person with intellectual disability does not want to participate (or experiences less fun) in cooking with the web application</td>
</tr>
<tr>
<td><strong>Case 2: SignaLEREN app—an app to support people with intellectual disabilities in autonomously dealing with stress or anxiety</strong></td>
<td></td>
</tr>
<tr>
<td>Selected value</td>
<td>• Autonomy</td>
</tr>
<tr>
<td>Positive effects</td>
<td>• The following were the positive effects for the people with intellectual disabilities:</td>
</tr>
<tr>
<td></td>
<td>• More self-direction, independence, and personal autonomy</td>
</tr>
<tr>
<td></td>
<td>• An extra support option besides support from care professionals</td>
</tr>
<tr>
<td></td>
<td>• Increased awareness of own stress and its causes</td>
</tr>
<tr>
<td>Negative effects</td>
<td>• Counterproductive effects of the app on stress if it does not work</td>
</tr>
<tr>
<td></td>
<td>• Less autonomy when using the app feels obligatory</td>
</tr>
<tr>
<td></td>
<td>• The app as a barrier to seeking contact with the care professional</td>
</tr>
<tr>
<td>Action opportunities—technology</td>
<td>• Enable people with intellectual disabilities to make choices themselves:</td>
</tr>
<tr>
<td></td>
<td>• Set the regularity of question pop-ups in the app</td>
</tr>
<tr>
<td></td>
<td>• Disregard the notifications at unsuitable moments by choosing the response option “I don’t want to answer this question (right now)”</td>
</tr>
<tr>
<td></td>
<td>• Delete data</td>
</tr>
<tr>
<td></td>
<td>• Schedule an appointment with their care professional via the app</td>
</tr>
<tr>
<td>Action opportunities—context</td>
<td>• Give people with intellectual disabilities access to the back end of the app, so that they can adjust specific content in the app themselves</td>
</tr>
<tr>
<td>Action opportunities—users</td>
<td>• Give people with intellectual disabilities some self-direction regarding the use of the app:</td>
</tr>
<tr>
<td></td>
<td>• Discuss what using the app entails and what happens with the data</td>
</tr>
<tr>
<td></td>
<td>• Give guidance in setting up and using the app at their own pace</td>
</tr>
<tr>
<td></td>
<td>• Evaluate app use and effects frequently during coaching moments</td>
</tr>
<tr>
<td></td>
<td>• Support people with intellectual disabilities in a different way, eg, redirect them to the app first, “Have you completed the app?”</td>
</tr>
<tr>
<td><strong>Case 3: Social robot, SARA—a robot to support the physical and social activities for people with intellectual disabilities</strong></td>
<td></td>
</tr>
<tr>
<td>Selected value</td>
<td>• Privacy</td>
</tr>
<tr>
<td>Positive effects</td>
<td>• None mentioned regarding the selected value</td>
</tr>
<tr>
<td>Negative effects</td>
<td>• Risk of privacy infringement owing to storage and sharing of personal data</td>
</tr>
<tr>
<td></td>
<td>• Insufficient insight into what data are collected and stored when the functionalities of the robot are expanded</td>
</tr>
</tbody>
</table>
### Observations About and Differences and Similarities Among the Cases

#### Effects

Part of the effects that were collected was case specific. For example, effects regarding healthy food were mentioned only in case 1 (Kookapp for groups), effects regarding insight into the stress of the person with intellectual disability were mentioned only in case 2 (SignaLEREN app), and effects regarding activation or development of cognitive skills of the person with intellectual disability were mentioned only in case 3 (social robot, SARA). Apart from case-specific effects, several general themes that were extracted from the effects across all 3 cases are presented in Table 4.

<table>
<thead>
<tr>
<th>Cases and categories</th>
<th>Examples</th>
</tr>
</thead>
</table>
| Action opportunities—technology | • Differentiate between accounts for administrators and care professionals using the robot to restrict access to personal data  
• Establish a maximum storage period for personal data before they are automatically deleted |
| Action opportunities—context | • Train the care professionals using the robot in accordance with the privacy law  
• Restrict the number of people who have access to personal data  
• Read and reconsider the consent statements annually with people with intellectual disabilities (or their representatives) who use the robot—taking into account any changes in the functionalities of the robot and therefore storage of other data  
• Revise the organization’s privacy policy  
• When people with intellectual disabilities use the robot to have contact with relatives, the privacy of both should be protected:  
  • Use headphones during this contact  
  • Create a cozy private “corner” in the location |
| Action opportunities—users | • None mentioned regarding the selected value |
Table 4. General (ie, not case specific) effect themes, values, and action opportunity themes.

<table>
<thead>
<tr>
<th>Positive effects</th>
<th>Actors to whom the themes mainly apply</th>
</tr>
</thead>
<tbody>
<tr>
<td>Customization of care</td>
<td>People with intellectual disabilities</td>
</tr>
<tr>
<td>Increase in self-reliance or self-direction</td>
<td>People with intellectual disabilities</td>
</tr>
<tr>
<td>Ease of work or job satisfaction</td>
<td>Care professionals</td>
</tr>
<tr>
<td>Efficiency or labor saving or time saving</td>
<td>Care organizations</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Negative effects</th>
<th>People with intellectual disabilities, care professionals, and care organizations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dependency on IT infrastructure</td>
<td>People with intellectual disabilities, care professionals, and care organizations</td>
</tr>
<tr>
<td>Risks related to the storage of privacy-sensitive data</td>
<td>People with intellectual disabilities, care professionals, and care organizations</td>
</tr>
<tr>
<td>Perception of having “yet another system”</td>
<td>Care professionals</td>
</tr>
<tr>
<td>Cost or time investment</td>
<td>Care organizations</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Values</th>
<th>People with intellectual disabilities, care professionals, and care organizations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quality of care</td>
<td>People with intellectual disabilities</td>
</tr>
<tr>
<td>Autonomy</td>
<td>People with intellectual disabilities</td>
</tr>
<tr>
<td>Privacy</td>
<td>People with intellectual disabilities, care professionals, and care organizations</td>
</tr>
<tr>
<td>Job satisfaction</td>
<td>Care professionals</td>
</tr>
<tr>
<td>Efficiency or affordability of care</td>
<td>Care organizations and society</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Action opportunities</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Keep the content (recipes, interventions, and activities) up to date</td>
<td>Care professionals</td>
</tr>
<tr>
<td>Train the care professionals regarding how to use the technology well</td>
<td>Care organizations</td>
</tr>
<tr>
<td>Connect the use of the technology to goals in the individual care plan or electronic health record</td>
<td>Care professionals and technology developers</td>
</tr>
<tr>
<td>Focus on upscaling (more users of the technology in the care organization)</td>
<td>Care organizations</td>
</tr>
</tbody>
</table>

Values

Table 5 shows the values that were identified as the top 3 values during the 3 workshops. Note that only in case 1 (Kookapp for groups), participants chose 4 values. Values identified in all 3 cases were “quality of care,” “autonomy,” “privacy,” “job satisfaction,” and “efficiency or affordability of care” (Table 4). In all 3 cases, certain values apply to a specific actor. For example, “job satisfaction” applied to care professionals, whereas “autonomy” was primarily related to people with intellectual disabilities. Other values applied to several actors (eg, “reliability”) or an actor group; for example, “efficiency or affordability of care” was related to the care organization or even society.

Table 5. Top values identified during “guidance ethics” workshops, based on the personal top 3 values of all participants per case.

<table>
<thead>
<tr>
<th>Case</th>
<th>Top values</th>
</tr>
</thead>
<tbody>
<tr>
<td>1—Kookapp for groups</td>
<td>Quality of care, efficiency, health, and enjoyment</td>
</tr>
<tr>
<td>2—SignaLEREN app</td>
<td>Quality of care, autonomy, and reliability</td>
</tr>
<tr>
<td>3—Social robot, SARA</td>
<td>Quality of care, autonomy, and privacy</td>
</tr>
</tbody>
</table>

Action Opportunities

Action opportunities regarding the technology comprised the development of the technology itself (eg, optimization of user experience and customization), input into the technology (eg, recipes for meals in case 1—Kookapp for groups; intervention options for reducing tension in case 2—SignaLEREN app; and activities in case 3—social robot, SARA), and output of the technology (eg, storage and use of data). Action opportunities regarding the context covered the need for guidelines, training and support, policy or agreements, and adjustments to the physical environment in which the technology is used. Action opportunities regarding the users mainly focused on how care professionals can integrate the technology into their daily care practice. Some behaviors need to be diminished (eg, care professionals need to “let go” instead of “take over” to give the...
person with intellectual disability more autonomy), some behaviors must be retained (eg, face-to-face contact moments), and some behaviors need to be added (eg, evaluation moments). Although some general themes across all 3 cases were identified (Table 4), most action opportunities were context specific.

For case 1 (Kookapp for groups), most action opportunities were listed for the value “user convenience.” Action opportunities included optimizing the user experience with the web application (eg, using icons and less text) and providing resources (eg, placing magnets on the kitchen wall to hold tablets while cooking). Action opportunities that stood out because they were mentioned by several participants were related to an attractive design and ease of operation of the web application. Although the value “health” was a top value in case 1, relatively few action opportunities were formulated for this value.

For case 2 (SignaLEREN app), none of the values stood out, but 4 values were evenly represented among most action opportunities: “quality of care,” “autonomy” (Table 3), “reliability,” and “efficiency.” “Quality of care,” “autonomy,” and “reliability” were the top 3 values. For “quality of care,” action opportunities included keeping the goal in mind and deploying the app as a means rather than a goal in itself (eg, personal goals of the person with intellectual disability as starting point for the conversation about how to use the app), training the care professionals on the use of the app, and having a person-centered approach (ie, customizing the app). Action opportunities for the value “reliability” included continuous provision of easily accessible support (eg, assigning SignaLEREN coaches and arranging a 24-h helpdesk) and the maintenance of the app organized within the own organization. Finally, action opportunities for the value “efficiency” included integrating the app in the care process (eg, embedding the use of the app in a particular care methodology), scaling up the use and adoption of the app (eg, deploying the app with all care professionals to whom it applies), and extracting and using data from the app (eg, built-in notifications in the app for when the person’s stress level is likely to become very high). The more frequently mentioned action opportunities included giving the person with intellectual disability access to the personal settings of the app (eg, frequency of prompts and data access rights), securing face-to-face contact of the person with intellectual disability and care professional, using data from the app to provide insights into stress level trends, connecting the app with the electronic health record, and assigning a SignaLEREN coach.

For case 3 (social robot, SARA), most action opportunities were related to the value “privacy” (Table 3). Action opportunities that stood out because they were mentioned by several participants included linking the use of the robot to individual care goals, connecting the robot with the electronic health record, and expanding the content that the robot can present. Although the value “autonomy” was a top value in case 3, relatively few action opportunities were related to this value, whereas relatively many action opportunities were linked to “effectiveness” (eg, optimizing the content that the robot can present and recurrent evaluation), which was not chosen as a top value.

Discussion

Relevance

Often, the development, implementation, and use of digital care does not entail an intentional and systematic effort to include the ethical considerations of all involved stakeholders [34]. Therefore, new technologies and the processes to integrate them into daily practice are often not aligned with the values and well-being of those influenced by them [15]. Instead, ethics is merely considered a separate area of attention (eg, a separate line of investigation or work package within projects) discussed by a distinct group of experts [35].

This study illustrates the types of insights that are gained when various stakeholders are involved in the reflection about the ethical impact of specific technologies and how this impact can be influenced for the better. This is illustrated using 3 cases of different digital care technologies for people with intellectual disabilities: a web application for cooking for groups (Kookapp for groups), an app for stress regulation (SignaLEREN app), and a robot for interaction and physical activity (social robot, SARA). Our findings may help researchers, innovators, and users of technology to move from a rather abstract thinking about ethics and responsible innovation toward effective practical approaches in which all stakeholders can be involved.

Principal Findings

In a short amount of time (three 3.5-h workshops), relatively much information was gathered in a multistakeholder setting regarding (1) positive and negative effects for various stakeholders of a specific digital care technology for people with intellectual disabilities; (2) values underlying these effects; and (3) action opportunities to take into account important values in the design, implementation, and use of the specific technology. The effects were primarily case specific, as they described the implementation of a technology in a specific context, but several general themes were also recognized. The latter included the effects of the technology on customization of care, dependency on IT infrastructure, self-reliance or self-direction of people with intellectual disabilities, risk of privacy infringement, care professionals’ ease of work, workload, and efficiency and investment. When all the effects were abstracted into values, several values were identified in all 3 cases and were found to be related to the general effects. These values were quality of care, autonomy, privacy, job satisfaction, and efficiency or affordability of care. Most action opportunities were related to the top values from the respective cases, as can be expected when the guidance ethics approach (stage 3) is followed. Hence, many action opportunities from case 1 (Kookapp for groups) were related to “enjoyment,” “efficiency,” and “quality of care.” However, relatively few action opportunities involved the top value, “health.” Notably, most action opportunities from case 1 were related to “user convenience”; however, this was not identified as a top value. In case 2 (SignaLEREN app), most action opportunities were related to the top values (“quality of care,” “autonomy,” and “reliability”) and to “efficiency.” In case 3 (social robot, SARA), the top values, “quality of care” and “privacy,” were well represented among the action opportunities, but few were related...
to the top value, “autonomy,” and relatively many were related to “efficiency” and “effectivity.”

Although action opportunities can only be described in relation to specific sociomaternal contexts [36] (ie, specific technologies in their contexts of use), our study reveals that, at a higher level, there are similarities regarding effects, values, and action opportunities for different cases. Thus, it may be wise for stakeholders of digital care technologies to not only learn how technologies can be responsibly used within their own context but also seek inspiration from similar contexts in which other technologies are used and from different contexts in which the same or comparable technologies are used. However, caution should be exercised when generalizing case-specific effects, values, and action opportunities to a broad scope.

**Comparisons With Previous Studies**

This is the first study to provide a broad overview of actual action opportunities for responsible digital care for people with intellectual disabilities. In their 2020 reports, the Dutch Centre of Ethics and Health advised the Dutch Ministry of Health, Welfare, and Sport about ethics regarding digital care such as apps and robots [37,38]. The themes discussed in the report regarding apps are cost savings, increase of autonomy, increase of well-being, unrest, information overload, decrease of human contact, overemphasis on health that can lead to medicalization, and increase of differences in health and inequality [37]. Notably, the diametrically opposed side of inequality was raised in our study, namely that differences between care organizations would decrease if they would organize meals using the Kookapp for groups. The report regarding care robots discusses meaningful contact, dignity, autonomy, dependency, privacy, and justice [38]. Apart from information overload and justice, all themes also appeared in ≥1 of the 3 cases in our study. This shows that our results covered most of the essential topics that ethics experts recognized.

Studies of ethics and health often describe themes that have a positive or negative load (eg, increase of autonomy or increase of inequality, respectively) or identify ethical harms [17,18,37,39]. Our data revealed that, in most cases, 2 sides of the same coin were considered in the multistakeholder setting, for instance, technology as a facilitator and a burden for care professionals’ work (all cases); cost or labor savings and high costs or time investment (all cases); positive and negative effects of the focus on a health theme (Kookapp for groups: awareness of the importance of healthy food vs a lot of emphasis on health at the cost of enjoying tasty food; SignaLEREN app: improving the stress signaling plan vs risk of medicalization of normal stress); and increase and risks of autonomy (SignaLEREN app: person with intellectual disability is less dependent on care professional but possibly also less “visible”). The advantage of using values—instead of themes or harms—as a starting point for fostering responsible use of digital care is that values are neutral and hence facilitate the consideration of both sides of the same coin [40].

**Strengths and Limitations**

As this study illustrates, the guidance ethics approach can be a valuable and low-key method to gain insight into different stakeholders’ experienced and expected positive and negative effects and values affected (or at stake) when using a specific technology and insight into action opportunities for responsible digital care. However, we recognize that the insights gained in this respect may fall short in terms of correctness (ie, being in agreement with facts or with what is generally accepted), concreteness (ie, being specific and detailed), and completeness (ie, the extent to which all relevant effects, values, and action opportunities have been identified) [41].

The correctness of our results about effects may be limited owing to, among others, a general lack of methodologically sound studies of the effects of digital care for people with intellectual disabilities. Hence, there was little to no evidence from scientific studies of the specific care technologies to be presented in stage 1 of the workshops. Therefore, the collected positive and negative effects are mainly based on subjective effects but from stakeholders with lived experience with the specific technology. In addition, an inherent characteristic of qualitative data analysis is that deriving themes from the data (including identifying values based on the effects of the technology) involves interpretation by the analyst. To limit subjectivity, values were discussed during the workshop with all participants, and themes were created independently by 2 researchers and discussed until consensus was reached. Another discussion point linked to correctness is that the values were discussed and presented as relatively stable entities, while they are neither stable nor singular [42,43]. Values may be affected by time and thus constantly defined and redefined (value dynamism), for instance, because users have gained experience with technology [41,43]. Although it may be challenging for researchers, innovators, and other stakeholders to continuously respond to this dynamism, a starting point could be to regularly collect the stakeholders’ perspectives about effects, values, and action opportunities.

The method used in this study has advantages regarding concreteness. For example, although the values are abstract, their definitions are embedded in the concrete effects from which they are derived. Moreover, the method results in relatively concrete output (action opportunities) compared with most ethics research on digital care [44,45], and action opportunities apply to the specific context of use. However, it was not always deducible from the workshop data what a participant specifically meant by an effect or action opportunity or to whom (eg, care professional or person with intellectual disability) specific insights applied. In the cases used in this study, it is not straightforward who is meant by the “user” of the care technology. To improve this, workshop leaders should be alert and ask participants to clarify whether they mean the care professional or the person with intellectual disability.

Regarding completeness, this study did not aim to be exhaustive in collecting effects, values, and action opportunities. However, we aimed to include a diverse sample of participants. Despite the accessibility of the workshops for people with intellectual disabilities, a participant with intellectual disability was included only in case 1 (Kookapp for groups). In case 3 (social robot, SARA), the person with intellectual disability withdrew on the morning of the workshop (with a valid reason), and in case 2 (SignaLEREN app), no person with intellectual disability was
invited. In addition, other relevant stakeholders were absent, for example, relatives of the people with intellectual disabilities, the board of directors, or representatives of health insurers. Although all relevant stakeholders that participants could think of were identified during the workshop and participants were asked to keep them all in mind, some perspectives may be missing in the results. Including people with disabilities requires special attention, as this is not common in the co-design or cocreation of digital care technologies [46]; however, this is upcoming [13,47,48]. To improve stakeholder inclusion in general, it may be useful to consult “design principles” for stakeholder engagement [49].

Furthermore, the method of deriving values from effects is suitable for identifying “proximal” values on the micro level that are specific to the technology and context of use [21], but mesolevel or macrolevel effects and more “distal” values may be missed [50] (such as “social justice,” which is an important theme in studies of ethics and digital care [45,51,52]). Whenever missed in bottom-up ethical dialogues (such as in this study), proximal and distal values (or principles) from predefined ethical frameworks could be brought in as “top-down” guidance. At the same time, the bottom-up approach is a strength of the guidance ethics approach, revealing important topics such as enjoyment of the person with intellectual disability or job satisfaction, which may be missed when an ethical theory is applied to a case instead [17,53]. In this sense, we argue that the top-down and bottom-up approaches are complementary.

Hence, we suggest moving back and forth between the perspectives of stakeholders affected by technology when implementing digital care and ethical frameworks or perspectives of experts about digital care ethics [19,54].

Finally, the action opportunities identified in this study require follow-up in practice. Responsible use of technology requires being continuously responsive and adaptive to new insights that are gained regarding effects, values, and action opportunities, from early design to local implementation and use [55,56]. In addition, it is conceivable that trade-offs between action opportunities need to be made owing to value conflicts (eg, autonomy vs duty of care [57]) and costs. Future studies may shed light on how action opportunities, once formulated, are further operationalized and applied by technology designers, user organizations, and individual end users of the technology and on what factors withhold stakeholders from doing so. Previous studies indicated that ethical concerns of stakeholders might considerably slow the pace of digital care innovation, implying that responsible innovation could be a core catalyst for the progress of digital care overall [18]. Through explicit attention to and communication about responsible digital care, not only are ethical concerns taken into account but also support and acceptance among the involved stakeholders are generated. This increases the chances for the successful implementation of socially desirable, ethically acceptable, and sustainable digital care that improves the lives of people with disabilities.

Acknowledgments
This study was conducted as a part of “Innovation Impulse Disability Care” (in Dutch: Innovatie-impuls Gehandicaptenzorg), a 3-year program funded by the Dutch Ministry of Health, Welfare, and Sport. The authors thank the workshop leaders of the Platform for the Information Society and all study participants for their participation.

Authors' Contributions
NMS, DRML, MS, BB, and AvdP were involved in conceptualization. NMS, BB, and AvdP contributed to the methodology. NMS and KNvD were responsible for the software, conducted the validation and formal analysis, were responsible for the resources, and performed data curation and visualization. NMS, KNvD, and AvdP conducted the investigation. NMS, KNvD, DRML, MS, BB, and AvdP wrote the original draft and were involved in reviewing and editing. AvdP and BB were involved in supervision and funding acquisition. NMS and KNvD were responsible for the software, conducted the validation and formal analysis, were responsible for the resources, and performed data curation and visualization. NMS, KNvD, DRML, MS, BB, and AvdP wrote the original draft and were involved in reviewing and editing. AvdP and BB were involved in supervision and funding acquisition. NMS and AvdP were involved in project administration.

Conflicts of Interest
None declared.

Multimedia Appendix 1
Overview of the effects, values, and action opportunities collected from the workshops for the 3 cases: Kookapp for groups; SignaLEEREN app; and social robot, SARA.

References


22. Casbes. Aanpak begeleidingsethiek. URL: https://begeleidingsethiek.nl/cases/ [accessed 2024-02-21]


33. Attribution 4.0 International (CC BY 4.0). Creative Commons. URL: https://creativecommons.org/licenses/by/4.0/ [accessed 2024-02-26]


Abbreviations
ECP: Platform for the Information Society

©Nienke M Siebelink, Kirstin N van Dam, Dirk R M Luktien, Brigitte Boon, Merlijn Smits, Agnes van der Poel. Originally published in JMIR Mental Health (https://mental.jmir.org), 28.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Mental Health, is properly cited. The complete bibliographic information, a link to the original publication on https://mental.jmir.org/, as well as this copyright and license information must be included.
Effectiveness of One Videoconference-Based Exposure and Response Prevention Session at Home in Adjunction to Inpatient Treatment in Persons With Obsessive-Compulsive Disorder: Nonrandomized Study

Ulrich Voderholzer\textsuperscript{1,2,3}, Prof Dr med; Adrian Meule\textsuperscript{1,2,4}, PD Dr; Stefan Koch\textsuperscript{1}, PhD; Simone Pfeuffer\textsuperscript{1}, Dr med; Anna-Lena Netter\textsuperscript{5}, PhD; Dirk Lehr\textsuperscript{6}, Prof Dr; Eva Maria Zisler\textsuperscript{1,2}, MSc

\textsuperscript{1}Schoen Clinic Roseneck, Prien am Chiemsee, Germany
\textsuperscript{2}Department of Psychiatry and Psychotherapy, University Hospital, LMU Munich, Munich, Germany
\textsuperscript{3}Medical Center, Department of Psychiatry and Psychotherapy, Faculty of Medicine, University of Freiburg, Freiburg, Germany
\textsuperscript{4}Institute of Medical Psychology, Faculty of Medicine, LMU Munich, Munich, Germany
\textsuperscript{5}MindDoc Health GmbH, Munich, Germany
\textsuperscript{6}Leuphana Universität Lüneburg, Lüneburg, Germany

Corresponding Author:
Eva Maria Zisler, MSc
Department of Psychiatry and Psychotherapy
University Hospital, LMU Munich
Nußbaumstraße 7
Munich, 80336
Germany
Phone: 49 15205886528
Email: Eva.Zisler@med.uni-muenchen.de

Abstract

Background: Therapist-guided exposure and response prevention (ERP) for the treatment of obsessive-compulsive disorder (OCD) is frequently conducted within clinical settings but rarely at places where patients are usually confronted with OCD symptom-provoking situations in daily life (eg, at home).

Objective: This study aimed to investigate patients’ views on 1 ERP session at home via videoconference and its impact on treatment outcome.

Methods: A total of 64 inpatients with OCD received 1 session of therapist-guided videoconference-based ERP at home in adjunction to a multimodal inpatient treatment between 2015 and 2020.

Results: Compared with 64 age- and sex-matched controls who received a multimodal inpatient treatment without 1 session of videoconference-based ERP at home, patients who received 1 session of videoconference-based ERP in adjunction to a multimodal inpatient treatment showed stronger reductions in OCD symptom severity from admission to discharge. Before the videoconference-based ERP session, patients reported high rationale credibility and treatment expectancy. After the videoconference-based ERP session, patients reported medium-to-high positive mood as well as depth and smoothness of the session, and they perceived the working alliance as high.

Conclusions: Results highlight the importance of administering therapist-guided ERP sessions in patients’ natural environment to enhance treatment response in OCD. Videoconference-based ERP as add-on to treatment as usual is, therefore, a promising approach to facilitate the application of ERP in patients’ natural environment and foster the generalization of ERP conducted in clinical settings.

\[(\text{JMIR Ment Health 2024;11:e52790}) \quad \text{doi:10.2196/52790}\]

KEYWORDS

obsessive-compulsive disorder; videoconference-based treatment; therapy; exposure; response prevention; OCD; prevention; inpatient; video; videoconference; therapist; therapists; mood; positive mood; environment; clinical setting
Introduction

Background

Obsessive-compulsive disorder (OCD) is a mental disorder characterized by intrusive and disturbing thoughts as well as repetitive patterns of behavior [1,2]. These are often multifaceted, that is, they include different obsessions and compulsions related to unwanted intrusive thoughts, fears of diseases, and contamination, among others [2,3]. OCD is a common disorder with a lifetime prevalence ranging from 1% to 3% and often has debilitating consequences on the daily functioning, well-being, and quality of life of affected persons as well as family members [4,5]. It usually emerges in late adolescence or early adulthood and has a chronic course if effective treatment is lacking [6,7]. Yet, OCD is often underrecognized and missed in primary care settings [8]. Thus, the duration of untreated illness in adults often exceeds 10 years, which creates a large treatment gap [9].

Exposure and Response Prevention in the Treatment of OCD

Cognitive-behavioral therapy (CBT) with exposure and response prevention (ERP) is the first-line, evidence-based psychotherapeutic treatment for OCD and is recommended as the psychotherapeutic method of choice [10-12]. ERP is a crucial element in CBT for OCD and requires patients to “engage in repeated, prolonged exposure to obsessions while refraining from compulsions” ([13], p. 85) [14]. Recent evidence stemming from inhibitory learning theory suggests that patients learn new associations during ERP (eg, “dirt—no fatal disease”), which in turn inhibit existing maladaptive associations (eg, “dirt—fatal disease”) [15]). This acquisition of associations is enabled by expectancy violation that is a mismatch between patients’ expectancy and outcome [15]. Although ERP is particularly useful in reducing OCD symptomatology, many patients find it difficult to endure upcoming unpleasant cognitions, feelings, and bodily sensations [16].

Besides the challenging nature of ERP itself, patients with OCD receiving CBT with ERP have to face a variety of difficulties [16]. First, patients are required to understand the underlying principles and measures of a treatment that is referred to as rationale credibility [17]. Second, patients need to expect that the treatment they are undergoing is effective [17-19]. Finally, it is beneficial if a positive working alliance is established between the patient and psychotherapist as it appears to predict treatment outcome [20]. Despite patients having to face various challenges when undergoing ERP, this psychotherapeutic intervention is highly effective for many people with OCD [21].

Home-Based ERP

The most commonly applied form of ERP is therapist-guided ERP in clinical settings (eg, at inpatient wards and in offices of psychotherapists), although the intervention can possibly be provided in several ways and facilities [22]. However, as persons with OCD often face the occurrence of obsessions and compulsions at home and feared situations or triggers cannot be replicated in a hospital or office, it can be hypothesized that home-based ERP may be beneficial in the treatment of OCD [23]. Although the theoretical framework of administering ERP at patients’ homes may sound reasonable, evidence on this treatment variant is mixed. Although some studies found that ERP at home was slightly advantageous in terms of symptom reduction [24,25], others reported that home-based ERP was no more effective than standard office-based ERP [23].

There are a variety of reasons why home-based ERP is not administered on a regular basis by the majority of psychotherapists. Specifically, many clinicians lack time or familiarity with this intervention [22,26,27]. Additionally, specialized hospitals are not always located close to the patient’s home, making therapist-guided ERP in the patient’s living environment difficult to conduct. Even with outpatient therapy close to the patient’s home, there is the challenge of therapists having to travel to the patient’s place, which is difficult to implement due to limited time resources of therapists. Therefore, patients often receive outpatient treatment that only includes a limited number of therapist-assisted exposure sessions, if any [22,28].

To achieve a better care situation for patients with OCD, there are 2 cost-effective ways of implementing therapist-guided home-based ERP. The first option is telephone-supported ERP, which was shown to be effective in 2 studies [29,30]. With advancing technologies, the second option is videoconference-based ERP, which can also be considered an adequate tool that comes with significant reductions in obsessive-compulsive symptoms, especially in persons with moderate OCD symptoms [27,31-33]. Videoconference-based psychotherapy has several advantages over in-person psychotherapy. First, by using videoconference-based psychotherapy, treatment with ERP can easily be delivered to patients who are homebound or living in rural areas [31,34]. Second, the administration of home-based in vivo exposures allows the generalization of treatment effects to other contexts [31,34]. Third, the therapist is in charge of accompanying and supporting the patient during ERP [35]. Fourth, therapist-assisted ERP has been shown to be more effective than non-therapist-assisted ERP [36], and using videoconference at home might allow for even more therapist-assisted ERP.

This Study

As research on videoconference-based ERP is still limited, we examined treatment effects in patients who received inpatient treatment with an additional videoconference-based ERP at home compared with an age- and sex-matched group of patients who received inpatient treatment without an additional videoconference-based ERP at home. Second, we assessed patients’ views on the current intervention (ie, treatment expectancy and rationale credibility) before undergoing the videoconference-based ERP session. Third, we examined patients’ evaluations of the videoconference-based ERP session (ie, depth, smoothness, positivity, and arousal) and working alliance with the therapists after having received videoconference-based ERP. We expected stronger reductions in OCD symptom severity in patients who received inpatient treatment with an additional videoconference-based ERP session at home compared with inpatients who received multimodal inpatient treatment without an additional videoconference-based ERP.
ERP session at home from admission to discharge. Furthermore, we expected high ratings on rationale credibility and treatment expectancy before as well as high ratings on satisfaction with the therapeutic sessions and quality of the therapeutic relationship from the patients’ perspective after the videoconference-based ERP session at home.

**Methods**

**Sample Characteristics**

This study was a nonrandomized, 2-group design study, in which a subset of patients who voluntarily participated in the study (videoconference exposure group) were compared with another subset of patients who did not participate in the study (control group). Although this design has disadvantages compared with a randomized controlled trial (RCT; see the Discussion section), it can be conducted more conveniently (eg, is less expensive and requires less resources) and may even have higher external validity as randomization may influence participation and outcomes when patients have a treatment preference [37].

Inpatients with OCD treated at the Schoen Clinic Roseneck (Prien am Chiemsee, Germany) between 2015 and 2020 were investigated. In Germany, inpatient treatment is indicated if at least 1 of the following factors applies: absence of or nonresponse to guideline-based disorder-specific outpatient treatment, danger to life, severe neglect, compulsive and avoidant behavior that is either severe or habitual resulting in an inability to maintain a normal daily routine and adherence to outpatient treatment, severe suffering and impairment of psychosocial functioning, psychological or somatic comorbidities aggravating outpatient treatment, and a particularly disease-promoting environment [10,38]. The treatment provided at the Schoen Clinic Roseneck adheres to the German S3 guidelines for the treatment of OCD [10]. Thus, the therapeutic concept is multimodal and consists of symptom-specific, individual CBT and ERP sessions, and other treatment elements, depending on indication (eg, psychopharmacological medication; see Table 1).

**Table 1. Sample characteristics (N=128).**

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Videoconference exposure group (n=64)</th>
<th>Control group (n=64)</th>
<th>Chi-square (df)</th>
<th>U</th>
<th>V</th>
<th>P value</th>
<th>r_{pb} (d)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Subtype of obsessive-compulsive disorder (ICD-10 code), n (%)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Obsessions-only subtype (F42.0)</td>
<td>0 (0)</td>
<td>1 (2)</td>
<td>1.62 (N/A)</td>
<td>N/A</td>
<td>0.11</td>
<td>.44</td>
<td>N/A</td>
</tr>
<tr>
<td>Compulsions-only subtype (F42.1)</td>
<td>8 (13)</td>
<td>11 (17)</td>
<td>.58 (N/A)</td>
<td>N/A</td>
<td>.07</td>
<td>.45</td>
<td>N/A</td>
</tr>
<tr>
<td>Mixed subtype (F42.2)</td>
<td>56 (88)</td>
<td>52 (81)</td>
<td>N/A</td>
<td>1888.00</td>
<td>N/A</td>
<td>.45</td>
<td>–0.08 (–0.18)</td>
</tr>
<tr>
<td><strong>Sex (female), n (%)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sex (female), n (%)</td>
<td>42 (66)</td>
<td>46 (72)</td>
<td>.58 (N/A)</td>
<td>N/A</td>
<td>.07</td>
<td>.45</td>
<td>N/A</td>
</tr>
<tr>
<td><strong>Age (years), mean (SD)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age (years), mean (SD)</td>
<td>26.95 (12.26)</td>
<td>29.28 (13.78)</td>
<td>N/A</td>
<td>1888.00</td>
<td>N/A</td>
<td>.45</td>
<td>–0.08 (–0.18)</td>
</tr>
<tr>
<td><strong>Length of stay (days), mean (SD)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Length of stay (days), mean (SD)</td>
<td>93.33 (30.54)</td>
<td>85.88 (38.40)</td>
<td>N/A</td>
<td>2302.50</td>
<td>N/A</td>
<td>.23</td>
<td>0.12 (0.22)</td>
</tr>
<tr>
<td><strong>Any comorbid mental disorder, n (%)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Any comorbid mental disorder, n (%)</td>
<td>44 (69)</td>
<td>43 (67)</td>
<td>0.04 (N/A)</td>
<td>N/A</td>
<td>.02</td>
<td>.85</td>
<td>N/A</td>
</tr>
<tr>
<td>Affective disorders</td>
<td>32 (50)</td>
<td>34 (53)</td>
<td>0.13 (N/A)</td>
<td>N/A</td>
<td>.03</td>
<td>.72</td>
<td>N/A</td>
</tr>
<tr>
<td>Anxiety disorders</td>
<td>11 (17)</td>
<td>16 (25)</td>
<td>1.17 (N/A)</td>
<td>N/A</td>
<td>.10</td>
<td>.28</td>
<td>N/A</td>
</tr>
<tr>
<td>Eating disorders</td>
<td>6 (9)</td>
<td>4 (6)</td>
<td>0.43 (N/A)</td>
<td>N/A</td>
<td>.06</td>
<td>.51</td>
<td>N/A</td>
</tr>
<tr>
<td>Antidepressant medicationc, n (%)</td>
<td>28 (49)</td>
<td>23 (51)</td>
<td>0.04 (N/A)</td>
<td>N/A</td>
<td>.02</td>
<td>.84</td>
<td>N/A</td>
</tr>
<tr>
<td><strong>Total score of Obsessive-Compulsive Inventory—Revised at admission, mean (SD)</strong></td>
<td>31.56 (12.62)</td>
<td>31.32 (14.89)</td>
<td>N/A</td>
<td>2095.00</td>
<td>N/A</td>
<td>.83</td>
<td>0.02 (0.02)</td>
</tr>
<tr>
<td><strong>Total score of Yale-Brown Obsessive-Compulsive Scale at admission, mean (SD)</strong></td>
<td>23.63 (5.33)</td>
<td>22.57 (6.51)</td>
<td>N/A</td>
<td>2202.50</td>
<td>N/A</td>
<td>.46</td>
<td>0.08 (0.18)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Videoconference exposure group (n=64)</th>
<th>Control group (n=64)</th>
<th>Chi-square (df)</th>
<th>U</th>
<th>V</th>
<th>P value</th>
<th>r_{pb} (d)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total score of Yale-Brown Obsessive-Compulsive Scale at admission, mean (SD)</strong></td>
<td>23.63 (5.33)</td>
<td>22.57 (6.51)</td>
<td>N/A</td>
<td>2202.50</td>
<td>N/A</td>
<td>.46</td>
<td>0.08 (0.18)</td>
</tr>
</tbody>
</table>

aICD-10: International Classification of Diseases, Tenth Revision.

bN/A: not applicable.

cInformation missing for 7 patients in the videoconference exposure group and 19 patients in the control group.

A total of 88 inpatients participated in this study, that is, received 1 videoconference-based ERP session at home in addition to inpatient treatment. As inpatient treatment at the Schoen Clinic Roseneck consists of 3 phases (psychoeducation and motivation, ERP, and transfer to the patients’ homes), participating patients were in the third phase of inpatient treatment. Psychotherapists at the hospital who had undergone technical training on videoconference-based ERP were authorized to offer the intervention to their patients. Patients were free to choose whether or not to receive the additional videoconference-based...
ERP session at home. On average, persons who received videoconference-based ERP at home had moderate OCD symptom severity according to the self-report version of the Yale-Brown Obsessive-Compulsive Scale (Y-BOCS; mean sum score 23.63, SD 5.33; Table 1; see recommendations by Cervin et al [39]).

Inpatients with OCD who were treated at the hospital within the same time period but who did not receive a videoconference-based ERP session at home were selected as the control group. Yet, these patients also received therapist-guided ERP in the hospital. Similar to the persons having received the videoconference-based ERP session, persons in the control group had, on average, moderate symptom severity according to the Y-BOCS (mean sum score 22.57, SD 6.51; Table 1; see recommendations by Cervin et al [39]). At the Schoen Clinic Roseneck, data from diagnostic assessments (eg, age, sex, diagnoses, medication, length of stay, and questionnaire scores) are automatically transferred to a database from which they can be exported without any identifying information by authorized employees. Thus, accessing individual patient charts is not necessary.

Figure 1. Participant flowchart.

Between 2015 and 2020, a total of 1471 patients with OCD were treated in the hospital who did not receive videoconference-based ERP at home, that is, did not take part in the study. Because of missing data, 1219 patients were available for matching with 65 of the 88 patients in the videoconference exposure group (Figure 1). Groups were matched based on propensity score matching without replacement using the FUZZY extension for SPSS (version 27.0; IBM Corp) [40]. Data were matched in regard to the variables age, sex, any comorbidity, length of stay, Obsessive-Compulsive Inventory–Revised (OCI-R) scores at admission, and Y-BOCS scores at admission. Using a match tolerance with which all 65 persons in the videoconference exposure group were retained did not result in well-matched groups (ie, groups still differed in age and length of stay). Thus, a match tolerance of 0.019 was chosen, which led to the exclusion of 1 person from the videoconference exposure group, resulting in a final sample size of 128 (ie, 64 persons per group; Table 1).

Figure 1.

Measures

**OCI-R Questionnaire**
The OCI-R [41,42] was used to examine obsessive-compulsive symptoms. The OCI-R is an 18-item self-report questionnaire with 6 subscales: washing, checking, ordering, obsessing, hoarding, and neutralizing. Responses are recorded on a 5-point scale ranging from 0 (not at all) to 4 (extremely) and refer to the extent of distress during the past month due to OCD symptoms. In a previous study, internal reliability coefficients
for the 6 subscales ranged between $\alpha = 0.76$ and 0.95. In this study, the internal reliability coefficient for the total scale was $\omega = 0.82$ at admission and $\omega = 0.86$ at discharge.

**Y-BOCS Questionnaire**

The self-report version [43] of the Y-BOCS [44,45] was used to examine OCD severity. The Y-BOCS is a 10-item self-report questionnaire comprising 2 subscales: obsessions and compulsions. Responses are recorded on a 5-point scale ranging from 0 (no symptoms) to 4 (extreme symptoms). Internal reliability coefficients ranged between $\alpha = 0.78$ and 0.88 in 2 validation studies [46,47] and between $\omega = 0.83$ and 0.91 in this study. Convergent validity has been supported by high correlations with other measures for obsessive-compulsive symptomatology, and divergent validity has been supported by moderate correlations with measures for related but distinct constructs such as worry [48-50].

**Credibility Expectancy Questionnaire**

The Credibility Expectancy Questionnaire (CEQ) [17] was used to assess the rationale credibility and treatment expectancy of the patient. The CEQ is a 6-item self-report questionnaire with 2 subscales: rationale credibility and treatment expectancy. Responses are recorded on a 9-point scale ranging from 1 (not at all) to 9 (very much). Internal reliability coefficients for the subscales ranged between $\omega = 0.71$ and 0.88.

**Session Evaluation Questionnaire**

The Session Evaluation Questionnaire (SEQ) [51] was used to examine the patients' satisfaction with the therapeutic sessions. The SEQ is a 21-item self-report questionnaire with 4 subscales: depth, smoothness, positivity, and arousal. Responses are recorded on a 7-point scale ranging from 1 (unpleasant) to 7 (pleasant). Internal reliability coefficients for the subscales ranged between $\omega = 0.61$ and 0.87. A closer inspection revealed that 2 items (1=slow, 7=fast; 1=moved, 7=composed) contributed to a low internal reliability of the arousal subscale. After removing those items, the remaining items of the arousal subscale had an internal reliability of $\omega = 0.76$. Thus, internal reliability coefficients for the subscales then ranged between $\omega = 0.76$ and 0.87.

**Working Alliance Inventory—Short Revised**

The Working Alliance Inventory—Short Revised (WAI-SR) [52] was used to examine the quality of the therapeutic relationship from the patient's perspective. The WAI-SR is a 12-item self-report questionnaire with 3 subscales: task, goal, and bond. Responses are recorded on a 7-point scale ranging from 1 (never) to 7 (always). Internal reliability coefficients for the subscales and the total scale ranged between $\omega = 0.84$ and 0.88.

**Procedure**

The videoconference app “VidyoMobile” by Vidyo, Inc was used to enable visual and auditory communication between the patient and therapist [25]. Patients were taught by a research staff member on how to use the smartphone, the tripod, and the videoconference app. Moreover, therapists prepared ERP sessions with their patients in close detail in a preceding session in the hospital. Before the ERP session, patients completed the CEQ. Patients received 1 videoconference-based ERP session each at home either on Friday afternoon or Monday morning. Each session had a duration of 2 hours on average. All videoconference-based ERP sessions were conducted by therapists specialized in CBT and ERP, and only the patient and the therapist were attending the session. The primary goal of the videoconference-based ERP session was to practice difficult situations associated with obsessions and compulsions in the patient’s home. The therapist’s role was to encourage the patient to face upcoming unpleasant feelings, emotions, and bodily sensations and to accompany them emotionally [16]. The exact execution of actions during ERP (ie, turning off the stove without checking, washing hands only once, etc) was not controlled by the therapist so as to give the patient a sense of personal responsibility in their own home.

After the ERP session, patients completed the SEQ and WAI-SR. In addition, after the videoconference-based ERP session, patients were asked to continue practicing the exposure exercise on their own. These exercises were not accompanied by the therapist, but debriefing followed in subsequent therapy sessions. Questionnaires assessing symptom severity (ie, OCI-R and Y-BOCS) were completed by the patients at admission and discharge.

**Data Analyses**

Group differences on categorical variables (OCD subtype, sex, comorbid mental disorders, and antidepressant medication) were tested with $\chi^2$ tests and on continuous variables (age, length of stay, and questionnaire scores at admission) with Mann-Whitney $U$ tests. Due to missing data at discharge (OCI-R: n=28, Y-BOCS: n=26), we examined changes of OCI-R and Y-BOCS total scores from admission to discharge as a function of a group with robust linear mixed models, which include cases with missing data in the maximum likelihood estimation. For this, we used R [53] and RStudio [54] and, specifically, the R package robustlmm [55]. The 2 models (1 for OCI-R scores and 1 for Y-BOCS scores) included fixed effects of time (admission vs discharge), group (videoconference exposure group vs control group), and their interaction term as well as a random intercept (ie, person-level random variability in scores at admission). As the package robustlmm does not produce parameter-specific $P$ values, we used the workaround by Geniole et al [56]. Specifically, nonrobust models were fitted with the lmer package [57]. $P$ values were obtained with the package lmerTest [58], and Satterthwaite-approximated degrees of freedom generated by the lmer models were combined with the output of the robustlmm model [56,59].

**Ethical Considerations**

The study was approved by the ethics committee of the Psychological Department of the Philipps University of Marburg, Germany. According to the guidelines by the institutional review board of the LMU Munich, retrospective analyses on already available anonymized data are exempt from requiring ethics approval. All participants in the videoconference exposure group signed informed consent before taking part in the study.
Results

As can be seen in Table 1, both groups did not significantly differ in age, sex, having any comorbid mental disorder, OCD subtype, antidepressant medication, OCI-R total scores at admission, and Y-BOCS total scores at admission. Robust linear mixed models revealed statistically significant interactions for group time for OCI-R ($b$=6.27; $P$=.01) and Y-BOCS ($b$=4.58; $P<.001$) scores, indicating that OCD symptom changes from admission to discharge differed as a function of group. As can be seen in Figures 2 and 3, the videoconference exposure group had larger OCD symptom reductions from admission to discharge than the control group. Descriptive statistics for obsessive-compulsive symptoms (total scores for OCI-R and Y-BOCS) at admission and discharge in the videoconference exposure and control groups are displayed in Table 2. On a scale ranging from 1 to 9, patients had mean (SD) values of 8.03 (0.74) on the subscale rationale credibility and 7.24 (1.13) on the subscale treatment expectancy on the CEQ. On a scale ranging from 1 to 7, patients had mean (SD) values of 5.87 (0.97) on the subscale depth, 3.60 (1.29) on the subscale smoothness, 4.61 (1.43) on the subscale positivity, and 4.11 (1.30) on the (reduced) subscale arousal on the SEQ. On a scale ranging from 1 to 7, patients had mean (SD) values of 6.25 (0.65) on the subscale therapeutic tasks, 6.52 (0.60) on the subscale therapeutic goals, 6.34 (0.75) on the subscale therapeutic bond, and 6.37 (0.57) on the total scale of the WAI-SR.

Figure 2. Mean sum scores of the Obsessive-Compulsive Inventory—Revised at admission and discharge as a function of group. The error bars indicate the SE of the mean. Effect sizes (rank biserial correlation coefficients $r_{rb}$ and Cohen $d$) refer to the changes within each group from admission to discharge.
Table 2. Descriptive statistics for obsessive-compulsive symptoms at admission and discharge in the videoconference exposure and control groups.

<table>
<thead>
<tr>
<th>Time point and statistic</th>
<th>Videoconference exposure group</th>
<th>Control group</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n (%) Mean (SD) Range</td>
<td>n (%) Mean (SD) Range</td>
</tr>
<tr>
<td>Admission</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Obsessive-Compulsive Inventory—Revised</td>
<td>64 (50) 31.56 (12.62) 6-56</td>
<td>64 (50) 31.32 (14.89) 3-59</td>
</tr>
<tr>
<td>Yale-Brown Obsessive-Compulsive Scale</td>
<td>64 (50) 23.63 (5.33) 9-36</td>
<td>64 (50) 22.57 (6.51) 5-35</td>
</tr>
<tr>
<td>Discharge</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Obsessive-Compulsive Inventory—Revised</td>
<td>57 (44.5) 13.54 (8.59) 1-35</td>
<td>43 (33.6) 20.67 (14.81) 3-56</td>
</tr>
<tr>
<td>Yale-Brown Obsessive-Compulsive Scale</td>
<td>56 (43.8) 11.48 (4.88) 1-25</td>
<td>46 (35.9) 14.92 (7.67) 1-31</td>
</tr>
</tbody>
</table>

Discussion

Summary of Results

This study showed that the group that had an additional therapist-guided, videoconference-based ERP session at home showed greater improvements during inpatient treatment for OCD, that is, displayed larger decreases in OCD symptomatology compared with treatment as usual. Obsessive-compulsive symptoms from admission to discharge decreased for patients who received a videoconference-based ERP session at home as well as for patients who received treatment as usual without a videoconference-based ERP session with medium to large effect sizes. Yet, obsessive-compulsive symptoms decreased even stronger for patients who have received inpatient treatment and a videoconference-based ERP session as an add-on. Furthermore, patients had high treatment expectancy and perceived the rationale as credible before...
receiving videoconference-based ERP. After undergoing videoconference-based ERP, patients perceived depth (ie, potency and value), smoothness of the session (ie, comfort and relaxation), and mood after the session (ie, positivity and arousal) as medium to high. Patients who received videoconference-based ERP rated working alliance (ie, agreement on therapeutic tasks and goals as well as therapeutic bond) with their therapist as high.

### Possible Mechanisms of Videoconference-Based ERP-Enhanced Symptom Reductions

Our results revealed that patients who received videoconference-based ERP at home in addition to a multimodal inpatient treatment had higher symptom reductions from admission to discharge with higher effect sizes than the control group. This might be significantly attributable to patients being able to generalize and extend their progresses achieved in the hospital to their own home; that is, with the help of the personal support of their therapist, they are more successful in giving up avoidance behavior at home as well [60]. Yet, it must be considered that there was no randomization in this study, which is why factors other than the additional videoconference-based ERP session might have also contributed to the reduction in OCD symptomatology from admission to discharge.

Alternative explanations for higher OCD symptom reductions in the videoconference exposure group might be that mostly patients who were highly motivated decided to participate in the additional videoconference-based ERP session or that the psychotherapists who treated patients receiving videoconference-based ERP were more motivated compared with other psychotherapists who treated the other patients with traditional ERP in the hospital only. Additionally, it might be possible that psychotherapists themselves expected that the additional ERP session at home would be beneficial for the patients and, thus, were highly engaged in the therapeutic sessions in the hospital as well, which particularly helped patients in reducing their OCD symptoms.

Despite methodological restrictions in nonrandomized study designs such as this study, there are also several disadvantages in RCTs that must be taken into account. First, participants are no passive recipients of interventions and do have treatment preferences. Patients with specific treatment preferences might, thus, refuse to take part in RCTs to avoid being randomized to the nonpreferred treatment, which reduces external validity [37]. Second, patients included in RCTs are strongly preselected, which was not the case in this study. Thus, the characteristics of patients included in this study correspond more to the real care situation. Third, internal validity of RCTs could be reduced as randomization to the nonpreferred treatment might influence patient adherence to the treatment protocol [37]. Accordingly, as this study was a nonrandomized study, patients were able to express and act on their treatment preferences as they could choose to receive the additional videoconference-based ERP session at home. This might have substantially increased patient adherence, which could, in turn, have been a factor contributing to reductions in OCD symptomatology. Furthermore, the 2 groups in this study were matched based on propensity score matching, which aims to account for absent randomization as it imitates some of the characteristics of an RCT [61]. Propensity score matching helps to strengthen causal arguments in observational studies by reducing selection bias [62].

Besides significant reductions in OCD symptomatology from admission to discharge in patients in the videoconference exposure group, the current results indicate that patients mainly had positive views on the videoconference-based ERP session, which became apparent in positive subjective ratings of the sessions. The positive effects of the videoconference-based ERP session on OCD symptomatology might be due to several change factors (ie, treatment expectancy and working alliance) that appear to be targeted in the videoconference setting to a sufficient degree. Several studies have provided evidence that treatment expectancy and understanding of the underlying treatment rationale are powerful predictors of psychotherapy outcome in general [18]. Additionally, as patients rated working alliance in the videoconference setting as high, this might also substantially contribute to the effects shown in this study. Previous studies have already shown that the videoconference setting enables the patient and psychotherapist to establish a strong and stable working alliance that is comparable to that in traditional face-to-face treatment [63,64]. Several studies even highlight that a positive working alliance is predictive of substantial decreases in symptomatology [31]. Although this study cannot show causal associations between working alliance and symptom reductions, a positive working alliance might substantially be linked to improvements of the patients’ condition in the face-to-face and videoconference setting.

### Limitations

As in every study, interpretation of the current results is limited to the persons and methods investigated. First, the examination of obsessive-compulsive symptoms was based on self-report, and—although the instruments used (OCI-R and Y-BOCS) are characterized by high validity and reliability—future studies may include therapist-rated measurements (eg, Y-BOCS interview version, Clinical Global Impression-Improvement Scale, and Global Assessment of Functioning [65,66]) as the inclusion of multiple views on the patients’ OCD symptomatology allows for an even more comprehensive evaluation. Second, due to limited material and human resources in the hospital, only a subset of patients treated at the hospital received an additional videoconference-based ERP session at home. Therefore, future studies might make the treatment available to a larger sample and replicate the effect. Third, future studies might examine the effects of multiple videoconference-based ERP sessions as the current add-on intervention included only 1 ERP session. Fourth, there was no randomization in this study. Hence, there might also be a number of factors other than the additional videoconference-based ERP session at home that might have contributed to significant OCD symptom reductions (eg, motivation to engage in ERP might have differed between the 2 groups and different therapists administered ERP sessions). Thus, conducting RCTs is recommended for future studies.
Conclusions

Altogether, this study showed that the group that received a 1-time home visit of videoconference-based ERP in conjunction to a multimodal inpatient treatment had greater improvements, that is, larger decreases in OCD symptomatology, during inpatient treatment of OCD compared with treatment as usual. In addition, patients’ ratings showed that the videoconference setting as well as working alliance with therapists was largely perceived as pleasant. Overall, it is recommended to provide patients with OCD with therapist-guided ERP at home. If it is not possible to accompany the intervention in person due to time constraints or other issues, videoconference-based therapy is a promising alternative to facilitate the application of ERP in patients’ natural environment and foster the generalization of treatment effects achieved in clinical settings.

Data Availability

The data sets generated during and/or analyzed during this study are available in the Open Science Framework repository (https://osf.io/pybhw/).

Conflicts of Interest

None declared.

References


https://mental.jmir.org/2024/1/e52790 JMIR Ment Health 2024 | vol. 11 | e52790 | p.438 (page number not for citation purposes)


Abbreviations

CBT: cognitive-behavioral therapy
CEQ: Credibility Expectancy Questionnaire
ERP: exposure and response prevention
OCD: obsessive-compulsive disorder
OCI-R: Obsessive-Compulsive Inventory—Revised
RCT: randomized controlled trial
SEQ: Session Evaluation Questionnaire
WAI-SR: Working Alliance Inventory—Short Revised
Y-BOCS: Yale-Brown Obsessive-Compulsive Scale
Original Paper

Design and Implementation of a Brief Digital Mindfulness and Compassion Training App for Health Care Professionals: Cluster Randomized Controlled Trial

Satish Jaiswal¹, PhD; Suzanna R Purpura¹, MS; James K Manchanda¹, BS; Jason Nan¹, BS; Nihal Azeez¹, BS; Dhakshin Ramanathan², PhD; Jyoti Mishra¹, PhD

¹Department of Psychiatry, University of California San Diego, San Diego, CA, United States
²Department of Mental Health, Veterans Affairs San Diego Medical Center, San Diego, CA, United States

Corresponding Author:
Satish Jaiswal, PhD
Department of Psychiatry
University of California San Diego
9452 Medical Center Dr
La Jolla, CA
San Diego, CA, 92037
United States
Phone: 1 858 232 2855
Email: s2jaiswal@health.ucsd.edu

Abstract

Background: Several studies show that intense work schedules make health care professionals particularly vulnerable to emotional exhaustion and burnout.

Objective: In this scenario, promoting self-compassion and mindfulness may be beneficial for well-being. Notably, scalable, digital app–based methods may have the potential to enhance self-compassion and mindfulness in health care professionals.

Methods: In this study, we designed and implemented a scalable, digital app–based, brief mindfulness and compassion training program called “WellMind” for health care professionals. A total of 22 adult participants completed up to 60 sessions of WellMind training, 5-10 minutes in duration each, over 3 months. Participants completed behavioral assessments measuring self-compassion and mindfulness at baseline (preintervention), 3 months (postintervention), and 6 months (follow-up). In order to control for practice effects on the repeat assessments and calculate effect sizes, we also studied a no-contact control group of 21 health care professionals who only completed the repeated assessments but were not provided any training. Additionally, we evaluated pre- and postintervention neural activity in core brain networks using electroencephalography source imaging as an objective neurophysiological training outcome.

Results: Findings showed a post- versus preintervention increase in self-compassion (Cohen $d=0.57$; $P=.007$) and state-mindfulness ($d=0.52$; $P=.02$) only in the WellMind training group, with improvements in self-compassion sustained at follow-up ($d=0.8$; $P=.01$). Additionally, WellMind training durations correlated with the magnitude of improvement in self-compassion across human participants ($p=0.52$; $P=.01$). Training-related neurophysiological results revealed plasticity specific to the default mode network (DMN) that is implicated in mind-wandering and rumination, with DMN network suppression selectively observed at the postintervention time point in the WellMind group ($d=–0.87$; $P=.03$). We also found that improvement in self-compassion was directly related to the extent of DMN suppression ($p=–0.368$; $P=.04$).

Conclusions: Overall, promising behavioral and neurophysiological findings from this first study demonstrate the benefits of brief digital mindfulness and compassion training for health care professionals and compel the scale-up of the digital intervention.

Trial Registration: Trial Registration: International Standard Randomized Controlled Trial Number Registry ISRCTN94766568, https://www.isrctn.com/ISRCTN94766568

(JMIR Ment Health 2024;11:e49467) doi:10.2196/49467

https://mental.jmir.org/2024/1/e49467

JMIR Ment Health 2024 | vol. 11 | e49467 | p.442
(page number not for citation purposes)
KEYWORDS
compassion; digital app; digital health; digital intervention; digital mental health; digital mindfulness; EEG; health workers; healthcare professionals; mindfulness; neuroplasticity; physicians; training

Introduction

Health care professionals receive intensive hands-on education and training so they may serve as resilient healers. Yet, with high workloads and multitasking demands, physicians can become vulnerable to workplace stress while making critical life-altering decisions for patients [1,2]. Evidence shows high rates of physician burnout (“workplace stress that has not been successfully managed,” International Classification of Diseases, World Health Organization), currently estimated at 44%, which is much higher than burnout in other professions [3,4]. Such chronic stress among physicians, in turn, may reduce the quality of physician-patient interactions and can lead to unprofessional behavior and attitudes [5,6]. Consistent with this, a study showed that burnout was associated with self-reported unprofessional conduct and less altruistic professional values among medical students at 7 US schools [7]. The COVID-19 pandemic also exacerbated distress in health care workers and has been shown to be associated with significant emotional pain, drop out from training, and reductions in work hours [8,9], which eventually affect the entire health care system. Hence, there is an emerging need for interventions that can help alleviate physician stress and prevent burnout [10].

Previous research has shown that once burnout develops, both individual-level and organizational strategies can result in clinically meaningful reductions in burnout [11], but organizational-level interventions (reducing work hours, increasing staffing, etc) have a larger overall effect size. [12] Notably, research has suggested that individual-level characteristics, particularly those related to mindfulness and self-compassion, may help to prevent burnout from developing [13-15]. Importantly, these are skills that can be developed. Research shows that when mindfulness and interoceptive-awareness exercises are provided as part of continuing medical education to primary care physicians, they help to reduce physician work-related stress and enhance well-being [11,16]. In a recent qualitative study among radio-oncologists, higher-trait mindfulness was found to be a protective factor from burnout and positively associated with life fulfillment [17]. Similarly, studies suggest that compassion training can promote well-being in medical students and improve the quality of clinical care [18]. Research has also shown that quality of care, health care costs, and the well-being of the clinician workforce are interlinked domains [19]. Hence, feasible, scalable, and effective mindfulness strategies delivered early may have far-reaching socioeconomic benefits for the health care system.

Compared to classroom approaches, digital training approaches can be highly scalable. Participants can flexibly engage with these trainings as per their convenience, so it is not a burden on busy work schedules [20]. Further, without the need for a teacher guide for digital training, the motivation for the practice is intrinsic and not reliant on teacher expertise. There are very few empirical studies among health care professionals that examine the efficacy of an app-based mindfulness program and its objective neural implications. A recent study in physician assistants showed a decrease in sleep dysfunction, enhanced connectivity between the medial prefrontal cortex and the superior temporal gyrus, as well as between regions critical for working memory after 8 weeks of intervention [21].

We have previously tested digital meditation approaches in adolescents and healthy young adults and shown positive neuro-cognitive outcomes [22,23]. Moreover, these digital trainings have integrated gamification, feedback, and rewards within closed-loop design systems to drive high user adherence [24].

In this study, we implemented a brief digital closed-loop training for health care professionals that involved an attention-to-breath practice with integrated compassion prompts. Respecting the time constraints of health professionals, each session provided 5-10 minutes of practice, and trainees had access to up to 60 digital sessions. Relative to a no-contact or business-as-usual control group, we evaluated the primary outcomes of change in mindfulness and self-compassion and monitored sessions of training engagement. While we additionally measured burnout, we did not expect this outcome to change with individual-focused training because it has been shown that organization-directed workplace interventions are more effective at addressing burnout [1,25] and burnout was also observed to be low in our sample. Thus, our main goal in this study was to examine whether mindfulness and compassion can be enhanced by brief digitally delivered practice sessions for health care professionals. Positive results from such a study may then serve as a rationale for future implementation integrated within an organizational framework to prevent burnout from developing.

Finally, this study also uniquely assayed objective neurophysiological plasticity associated with the training alongside subjective behavioral changes. For this, we measured electroencephalography (EEG)-based brain signals on an interoceptive attention-to-breathing task, assessed before and after training. The rationale for selecting such a task for neurophysiological measurements is that interoceptive attention to breathing is a core feature of several meditation practices [26,27]. On this task, we were particularly interested in neurophysiological activity within the default mode network (DMN), which has been shown to be modulated by meditation [28-31]. The DMN is a functional network that has not only been consistently associated with autobiographical memory and self-referencing but also on-task behavioral variability, mind-wandering, and rumination [32-37]. We hypothesized that mindfulness and compassion training would suppress DMN activity. Our ultimate analyses focused on whether objective modulation of the DMN relates to subjective behavioral changes in self-compassion and mindfulness.
Methods

Participants
A total of 43 human participants recruited in the study (mean age 28.77, SD 4.13; range 23-43 y; 20 male participants). All human participants were fluent in English. Participants were recruited from the University of California San Diego (UCSD) School of Medicine from Spring 2021 to Fall 2022 academic quarters through email advertisements and campus flyers.

Participants provided demographic data with regards to age, gender, and ethnicity. All participants were healthy adults, that is, they did not have any current medical diagnosis nor were taking any current psychotropic medications. Healthy status and affiliation to the UCSD School of Medicine were the only eligibility criteria.

Participants completed the Maslach Burnout Inventory (MBI) at the time of screening; MBI scores did not reflect high burnout in our sample as all scores were less than the midscore of the MBI score range (see the Results section).

Study Design
The study design was interventional and cluster randomized. Of the total 43 study participants, 22 were enrolled in the digital WellMind intervention group, and 21 were part of the no-contact control group. Participants were cluster randomized based on the academic quarter of enrollment to the WellMind or control group. Specifically, all participants recruited during Spring 2021, Spring 2022, and Fall 2022 academic quarters were assigned to the WellMind group, and the no-contact control group participants were recruited during Fall 2021. This was done because individuals within each academic quarter (but not across quarters) were working or studying together and, hence, knew each other professionally and could reveal components of the study intervention to each other. The WellMind group participants received the digital app intervention and had periodic email contact from our research team during the intervention, at about once every 2 weeks, to ensure compliance and help troubleshoot any issues faced by the participants. On the other hand, the no-contact control group had no interaction with the study research team or any digital training resource provided to them between their pre- and postintervention time points.

Sample Size and Power
The sample size within each group was powered to detect medium effect size for pre- or postintervention differences (Cohen $d>0.6$) at $\beta$ power of 0.8 and $\alpha$ level of 0.05. Between-group differences met criteria for investigating only large effect size outcomes (Cohen $d>0.8$) at $\beta$ power of 0.8 and $\alpha$ level of 0.05. Effect sizes were calculated a priori using the G*Power (Axel Buchner) software [38].

Intervention
The WellMind digital intervention was deployed on the BrainE platform, implemented in Unity Game Engine (Unity Technologies), and available on both iOS and Android phone devices [39]. This digital program is Health Insurance Portability and Accountability Act (HIPAA) compliant and secured by password protection, and each user interacts through an alphanumeric study ID that is not linked to any personal health information. Participants accessed the app in their own free time and engaged in breath-focused mindfulness training, with each session lasting 5-10 minutes for up to 60 sessions. The training was delivered in a game-like format and was performance adaptive. Specifically, individuals were requested to close their eyes, pay attention to their breathing, and tap the mobile screen after a specific number of breaths. The app monitored the consistency of tap responses. If the user was distracted based on the low consistency of breath monitoring taps, a gentle chime reminded the user to let go of the distraction and revert their attention back to mindful breathing. Initially, at level 1, participants tapped the screen after each breath. If they were able to do this consistently for 3 repeats of level 1 of 1 minute duration each, they graduated to level 2 and tracked 2 breaths at a time for 2 minutes, and so on. Thus, in the performance-adaptive task, the level reflected the number of minutes spent at that level and the number of breaths the participant was requested to repeatedly monitor. The maximum achievable level was level 10, that is, monitoring 10 breaths at a time for up to 10 minutes. When the user graduated to the max level, they stayed at this level until the end of all assigned sessions, that is, 60 sessions. Also within the game-like format, when the participant opened their eyes at the end of a level, a peaceful nature scene would slowly unfold as a form of training reward.

Overall, this digital meditative practice is considered closed loop because of its performance-adaptive feature [22,23]. Consistent attention to breathing is emphasized over other types of breathing techniques, such as deep breathing. The moment-to-moment performance tracking further allows quantification of the attentive focus during each session, which is not possible with traditional nondigital meditation.

Finally, the training also introduced standard compassion cultivation instructions as audio and text prompts before the start of each session’s breath practice. Prompts were updated every 6 sessions, with a total of 10 prompts gradually increasing in complexity over 60 sessions. These prompts were designed per guidance from the Compassion Cultivation Training program [40] and included (1) settling the mind, (2) compassion for a loved one, (3) compassion for oneself, (4) loving kindness for oneself, (5) embracing common humanity, (6) embracing common humanity continued, (7) cultivating compassion for oneself and others, (8) cultivating compassion for others continued, (9) active compassion, and (10) integrated compassion cultivation practice. The WellMind training app and study design are summarized in Figure 1. Participants received in-app notifications once a day, reminding them to complete their training.
Figure 1. Brief digital mindfulness and compassion training. (A) The WellMind app delivered closed-loop, that is, performance-adaptive, attention to breath training. Common instruction across 10 levels of training is shown. At each level, the user tapped the mobile screen after (certain level) number of breaths while keeping their eyes closed. Feedback included auditory chimes to guide consistent performance and signal the end of training. A distinct, calming nature scene is unveiled at the end of each block in the session, along with focus feedback based on consistency of performance. Levels of breath monitoring were tied to 10 levels of cultivating compassion instructions. (B) The study design incorporated behavioral and neurophysiological assessments at pre- and posttraining and continued behavioral assessments at follow-up conducted in the WellMind training group relative to a no-contact control group.

Behavioral Assessments
At baseline (T1), postintervention completion (T2; or a 3-month no-contact period for the control group), and at follow-up (T3; 6 months following baseline), participants completed validated behavioral self-report scales of self-compassion: a 12-item self-compassion scale [41], and mindfulness: a 14-item Mindful Attention Awareness Scale [42]. These measures served as the primary outcomes. MBI measures were obtained as exploratory outcomes at T1 and T2. The Cronbach $\alpha$ measure of reliability was calculated for each of these behavioral measures at baseline.

Neurocognitive Assessments
In addition, participants completed an objective neurophysiological assessment of interoceptive attention to breathing at T1 and T2. For these assessments, all participants made individual study visits at the Neural Engineering and Translational Labs at the UCSD. Assessments were deployed on the BrainE platform with simultaneous EEG [43], delivered on a laptop (running on the Windows 10 operating system) at a comfortable viewing distance. The Lab Streaming Layer protocol was used to time stamp all user response events in this assessment [44].

In the interoceptive attention to breathing task, participants were instructed to close their eyes, breathe naturally, and respond every 2 breaths by tapping on the spacebar [45,46]. The computer screen appeared gray for the 5-minute duration of the task, implemented in two 2.5-minute blocks. A beep signaled the end of the task, at which time participants opened their eyes. The median response time (RT) on the interoceptive task was monitored for all human participants so that we could identify and contrast neurophysiological activity on high consistency, that is, attentive breath monitoring trials (trials with RT $\leq$ 1 median absolute deviation of median RT) versus low consistency, that is, distracted trials (trials with RT $>$ 1 median absolute deviation of median RT) in each human participant.

EEG data were collected using a 24-channel cap with saline soaked electrodes following the 10-20 system and a wireless SMARTING amplifier (mBrainTrain). The signals were digitized with a sampling rate of 500 Hz and 24-bit resolution and stored as .xdf files.

Behavioral Data Analyses
For behavioral subjective scales, scores on self-compassion and mindfulness were calculated at T1, T2, and T3 and for MBI at T1 and T2. T2 versus T1 and T3 versus T1 scores were compared within each group using 2-tailed paired $t$ tests or its nonparametric equivalent Wilcoxon signed rank test depending on the distribution of the behavioral scores; the normality of distributions was checked using the Levene test. For mindfulness, we compared state mindfulness across the 3 time points, which is a component of the dispositional trait mindfulness scale [47], as we expected state mindfulness but not trait mindfulness to be malleable with training. Cohen $d$ effect sizes were calculated for both within and between-group differences. Repeated measures analyses of variance comparing between-group behavioral differences were not conducted in this first study given the large effect sizes.
(d>0.8) needed to observe significant group differences with adequate power.

To investigate the relationship between outcome gains and training engagement, behavioral changes in the WellMind training group were correlated with the number of training sessions completed by participants using Spearman correlations.

**Neurocognitive Data Analyses**

We applied a uniform processing pipeline to all EEG data published in several of our studies [43,45,46,48-53]. This included (1) EEG channel data processing and (2) cortical source localization of the EEG data to estimate source-level neural activity. Details of this analysis are provided in the Multimedia Appendix 1 [51,53-68].

Alpha band EEG data were trial averaged for high versus low consistency (ie, attended vs distracted) breath monitoring trials on the interoceptive attention task. These trials were compared for within-group pre- versus postintervention activity differences in the fronto-parietal network (FPN), cingulo-opercular network (CON), and DMN using paired t tests. Effect sizes were also calculated for neural data, reported as Cohen d: 0.2=small, 0.5=medium, and 0.8=large [69]. Given that we have observed large effect size neural outcomes (d>0.8) in our previous digital training studies [22,23,70], repeated measures ANOVA was conducted to analyze between-group post- versus preintervention network effects; the Greenhouse-Geisser significance correction was applied to adjust for lack of sphericity. Finally, Spearman correlations were used to analyze neurobehavioral associations.

**Ethical Considerations**

Each participant gave written informed consent in accordance with the Declaration of Helsinki before participating in the experiment. All the experimental procedures were approved by the institutional review board of the University of California San Diego (protocol #180140). All data is de-identified, and up to US $150 in compensation was provided as an e-Gift card to all participants for completing all aspects of the study, including assessments and intervention procedures.

**Results**

**Baseline Group Comparisons**

There were no demographic differences between the 2 groups for age, gender, and ethnicity (Table 1). Age comparisons were made using the Wilcoxon sum rank test, and gender and ethnicity comparisons were made using chi-square tests.

<table>
<thead>
<tr>
<th>Demographics and baseline behaviors</th>
<th>WellMind (n=22)</th>
<th>Control (n=21)</th>
<th>Group difference P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (years), mean (SD)</td>
<td>27.91 (3.15)</td>
<td>29.67 (4.96)</td>
<td>.32</td>
</tr>
<tr>
<td>Gender n (%)</td>
<td></td>
<td></td>
<td>.55</td>
</tr>
<tr>
<td>Men</td>
<td>8 (36)</td>
<td>12 (57)</td>
<td></td>
</tr>
<tr>
<td>Women</td>
<td>14 (64)</td>
<td>8 (38)</td>
<td></td>
</tr>
<tr>
<td>Ethnicity n (%)</td>
<td></td>
<td></td>
<td>.19</td>
</tr>
<tr>
<td>Asian</td>
<td>8 (36)</td>
<td>8 (38)</td>
<td></td>
</tr>
<tr>
<td>Black or African American</td>
<td>0 (0)</td>
<td>1 (5)</td>
<td></td>
</tr>
<tr>
<td>American Indian or Alaska Native</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>10 (46)</td>
<td>9 (43)</td>
<td></td>
</tr>
<tr>
<td>More than 1 ethnicity</td>
<td>3 (14)</td>
<td>2 (10)</td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>1 (5)</td>
<td>1 (5)</td>
<td></td>
</tr>
<tr>
<td>Trait mindfulness, mean (SD)</td>
<td>3.18 (0.66)</td>
<td>3.14 (0.77)</td>
<td>.84</td>
</tr>
<tr>
<td>Self-compassion, mean (SD)</td>
<td>2.69 (0.57)</td>
<td>3.09 (0.73)</td>
<td>.05</td>
</tr>
<tr>
<td>MBI(^a) emotional exhaustion, mean (SD)</td>
<td>16.27 (4.45)</td>
<td>24.05 (10.22)</td>
<td>.02(^b)</td>
</tr>
<tr>
<td>MBI personal accomplishment, mean (SD)</td>
<td>21.41 (3.63)</td>
<td>32.86 (6.73)</td>
<td>&lt;.001(^c)</td>
</tr>
<tr>
<td>MBI depersonalization, mean (SD)</td>
<td>5.36 (3.33)</td>
<td>9.29 (7.93)</td>
<td>.23</td>
</tr>
</tbody>
</table>

\(^a\)MBI: Maslach Burnout Inventory.  
\(^b\)P<.05.  
\(^c\)P<.001.

Our main outcome variables of mindfulness and self-compassion also did not show significant group differences at baseline. Burnout measures on the MBI showed significantly less emotional exhaustion and sense of personal accomplishment in the WellMind group relative to the control group at baseline, but no differences in MBI depersonalization. Overall, burnout levels in the WellMind group were low, that is, less than a scale midscore of 18 for emotional exhaustion, greater than a scale
A midscore of 16 for sense of personal accomplishment, and less than a scale midscore of 10 for depersonalization.

All behavioral measures had consistently high Cronbach $\alpha$ measured across all human participants at baseline (mindfulness: $\alpha=.81$, self-compassion: $\alpha=.86$, and MBI: $\alpha=.88$).

**Behavioral Results**

Participants in the WellMind group completed 40.64 (SD 17.79) training sessions on average, or 68% (40.64/60) of the total 60 sessions; the number of training sessions completed by different participants is shown in Figure 2A. The number of training sessions completed directly related to the final level achieved in training across participants ($p=.76, P<.001$; Figure 2B). Additionally, we found that the number of training sessions was also significantly related to post- versus preintervention increase in self-compassion ($p=.52, P=.01$; Figure 2C) and trended toward significance in relation to post- versus preintervention increase in mindfulness ($p=.38, P=.08$; Figure 2D), but this was not statistically significant.

**Figure 2.** WellMind training sessions completed and relationship with outcomes. (A) The number of training sessions completed by each WellMind participant. (B) The final level of breath training is significantly related to the number of completed training sessions. (C) Post- versus prechange in self-compassion is significantly related to number of completed training sessions. (D) Post- versus prechange in state mindfulness showed a trend in its relationship to the number of completed training sessions. The Spearman correlation results are shown.

The within-group changes in primary measures of self-compassion and state mindfulness and exploratory measures of burnout (MBI scores) between post- versus preintervention and follow-up versus preintervention sessions are shown in Tables 2 and 3 for the WellMind and control groups. A significant increase in self-compassion and mindfulness was observed at post- versus preintervention sessions only in the WellMind group, and the significant increase in self-compassion was sustained at follow-up (Figure 3). As the self-compassion scale has 2 components for compassionate self-responding (CSR) and uncompassionate self-responding (USR) [71], we also analyzed whether either CSR or USR undergo significant training-related change. CSR reflects self-kindness, common humanity, and mindfulness, while USR reflects the opposite constructs of self-judgment, isolation, and overidentification.

We found that the WellMind training-related improvement in self-compassion was exclusively driven by a reduction in USR (post vs preintervention change: $-0.42$, SD 0.55; $P=.002$; and follow-up vs preintervention change: $-0.82$, SD 1.27; $P=.009$), but there was no significant change in CSR at post ($P=.17$) or follow-up ($P=.07$). Baseline burnout in our sample was low, and there were no significant post- versus preintervention changes in MBI scores in both groups.

Cohen $d$ effect sizes for within-group differences were calculated as the mean difference between post- versus preintervention (or follow-up vs preintervention) outcomes expressed in pooled SD units. Cohen $d$ effect sizes were also calculated for between-group differences. Effect sizes for significant self-compassion and mindfulness outcomes were in the medium range (Tables 2 and 3).
Table 2. Summary of behavioral outcomes obtained at baseline (T1) and 3 months (postintervention; T2) after baseline. The primary outcome measures for the study, self-compassion and state mindfulness, showed significant changes in the WellMind group at postintervention time point and are noted. The control group did not show any significant changes. Data were checked for normality, and pre- versus postintervention within-group differences were appropriately compared using a paired \( t \) test if normal, or else using the nonparametric Wilcoxon sign rank test. Both within-group and between-group Cohen \( d \) effect sizes were calculated.

<table>
<thead>
<tr>
<th>Pre- versus postintervention outcomes</th>
<th>WellMind</th>
<th>Control</th>
<th>Between-group effect size</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>T1, mean (SD)</td>
<td>T2, mean (SD)</td>
<td>Effect size</td>
</tr>
<tr>
<td>Self-compassion</td>
<td>2.69 (0.57)</td>
<td>3.02 (0.58)</td>
<td>0.57</td>
</tr>
<tr>
<td>State mindfulness</td>
<td>2.94 (0.79)</td>
<td>3.37 (0.85)</td>
<td>0.52</td>
</tr>
<tr>
<td>MBI emotional exhaustion</td>
<td>16.27 (4.45)</td>
<td>15 (5.86)</td>
<td>-0.24</td>
</tr>
<tr>
<td>MBI personal accomplishment</td>
<td>21.41 (3.63)</td>
<td>21.64 (4.41)</td>
<td>0.06</td>
</tr>
<tr>
<td>MBI depersonalization</td>
<td>5.36 (3.33)</td>
<td>4.5 (3.71)</td>
<td>-0.24</td>
</tr>
</tbody>
</table>

\( aP<.001. \)  
\( bP<.05. \)  
\( ^cMBI: \) Maslach Burnout Inventory.

Table 3. Summary of behavioral outcomes obtained at baseline (T1) and 6 months after baseline (follow-up; T3). Data were checked for normality, and pre- versus postintervention within-group differences were appropriately compared using a paired \( t \) test if normal, or else using the nonparametric Wilcoxon sign rank test. Both within-group and between-group Cohen \( d \) effect sizes were calculated.

<table>
<thead>
<tr>
<th>Pre- versus postintervention outcomes</th>
<th>WellMind</th>
<th>Control</th>
<th>Between-group effect size</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>T1, mean (SD)</td>
<td>T3, mean (SD)</td>
<td>Effect size</td>
</tr>
<tr>
<td>Self-compassion</td>
<td>2.69 (0.57)</td>
<td>3.14 (0.55)</td>
<td>0.80</td>
</tr>
<tr>
<td>State mindfulness</td>
<td>2.94 (0.79)</td>
<td>3.41 (1.08)</td>
<td>0.50</td>
</tr>
</tbody>
</table>

\( aP<.001. \)
Figure 3. Post- versus preintervention (T2 vs T1) and follow-up versus preintervention (T3 vs T1) self-compassion and state mindfulness outcomes in the WellMind (n=22) and control (n=21) group; data at follow-up were missing for 2 participants in the WellMind group and for 1 participant in the control group. Both (A) self-compassion and (B) state mindfulness scores significantly improved in the WellMind group. Bar plots show the change in score mean and standard error about the mean for error bars, with the actual distribution of scores shown as scatter points. As self-compassion and state mindfulness measures had normal distributions, *P value results are from paired t tests between pre- and postintervention or preintervention and follow-up assessments.

Neurocognitive Results

We analyzed post- versus preintervention modulation of source-localized neural activity on the interoceptive attention to breathing assessment in brain regions of interest collated within canonical cognitive control networks, specifically the FPN, CON, and DMN. There were no between-group differences in neural network activity at baseline (P>.05). The within-group changes in network activity between post- versus preintervention sessions are shown in Table 4 for the WellMind and control groups. Per our hypothesis, a significant decrease in activity in the mind-wandering and rumination-associated DMN was observed at post versus pre sessions only in the WellMind group (Figure 4). Besides this, the only other significant change observed was a decrease in post- versus preintervention FPN activity in the control group. Cohen d effect sizes were calculated for both within-group and between-group differences (Table 4). Notably, the reduction in post- versus preintervention DMN activity in the WellMind group showed both large within-group and between-group effect sizes (DMN between group repeated measures ANOVA: session interaction: $F_{1,32}=10.11; P=.003; \eta^2=0.24$; main effects of group or session were not significant, $P>.30$; also, the other networks, FPN and CON, did not show any significant between-group effects, $P>.10$).

Table 4. Summary of neural activity outcomes obtained at baseline (T1) and at 3 months (postintervention; T2). The primary network of interest for the study, the default mode network (DMN), showed significant change in the WellMind group at post- versus preintervention and is noted. Besides, the control group showed a significant change in fronto-parietal network (FPN) activity. Mean (SD) data are shown in 10-4 cortical source arbitrary units for all variables. Post- versus preintervention within-group and between-group Cohen d effect sizes are calculated.

<table>
<thead>
<tr>
<th>Pre- versus postintervention outcomes</th>
<th>WellMind</th>
<th>Control</th>
<th>Between-group Cohen d effect size</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>T1, mean (SD)</td>
<td>T2, mean (SD)</td>
<td>Effect size</td>
</tr>
<tr>
<td>FPN</td>
<td>3.08 (3.14)</td>
<td>1.96 (1.45)</td>
<td>-0.46</td>
</tr>
<tr>
<td>CON$^b$</td>
<td>2.55 (2.18)</td>
<td>1.44 (1.76)</td>
<td>-0.56</td>
</tr>
<tr>
<td>DMN</td>
<td>11.92 (12.55)</td>
<td>1.44 (11.66)</td>
<td>-0.87</td>
</tr>
</tbody>
</table>

$^aP<.05$.  
$^b$CON: cingulo-opercular network.
**Figure 4.** Training-related neurophysiological changes evaluated on the attention-to-breath monitoring assessment. (A) Schematic of task instructions. (B) A power frequency plot of scalp channel data across all participants and sessions showed peak processing in the $\alpha$ frequency band (8-12 Hz). (C) Source-reconstructed electroencephalography data were analyzed for 3 networks: frontoparietal network (FPN), the cingulo-opercular network (CON), and the default mode network (DMN); regions of interest averaged within each network are shown. (D) Comparisons of the WellMind versus control group network activity showed significant reduction in activity only for the WellMind group in the DMN; bar plots show mean and standard error about the mean $\alpha$ band activity (y-axes: 10–4 cortical source activity arbitrary units) within the 0- to 4-second epoch before breath responses, as well as the relative response on low versus high consistency (i.e., distracted versus attended) trials at pre- and postintervention time points and the post-pre difference. Actual activity distributions are shown as scatter points.

**Neurobehavioral Associations**

We conducted neurobehavioral correlation analyses between neural measures of post- versus preintervention change in DMN activity and change in primary behavioral outcomes of self-compassion or state mindfulness. Spearman partial correlations were implemented that accounted for participant groups and their baseline burnout score differences (MBI; Table 1). We found a significant relationship between change in DMN network activity and change in self-compassion ($\rho=-0.368; P=.04; \text{Figure 5}$) but not state mindfulness ($P>.50$). Specifically, individuals who showed the largest improvement in self-compassion also showed the greatest DMN suppression. We also verified that these effects were driven by the relationship between DMN suppression and reduction in USR ($\rho=0.37; P=.04$), but there was no significant relationship with change in CSR ($P=.35$).

**Figure 5.** Relationship between post- versus preintervention change in self-compassion scores and change in default mode network (DMN) network activity. Residuals are plotted using Spearman partial correlation, taking into account group assignment to the WellMind or control group and baseline burnout score differences between groups.
We also found that the extent of training-related DMN suppression in the WellMind group was significantly related to their baseline DMN activity ($p = -0.74$; $P = .002$); this result did not change when controlling for the number of WellMind training sessions completed, and no such relationship was observed for the control group ($P = .71$). These results suggest that baseline DMN activity could be a predictive marker for the extent of training-related neural plasticity in this network, which in turn relates to improvement in self-compassion.

**Discussion**

In this study, we developed WellMind, a digital mindfulness and compassion training, and implemented it with health care professionals. Respecting their real-world time constraints, the training was brief (5-10 minutes per session) and available for up to 60 sessions. The novelty of the training lay in its closed-loop, that is, performance-adaptive mechanics and quantitative feedback design applied to attentive breathing, with levels of compassion cultivation instructions. To account for the practice effects of repeat assessments, we included a no-contact control group in the study design. We found that the WellMind intervention significantly improved the primary behavioral outcomes of self-compassion and mindfulness, with the improvements in self-compassion sustained at follow-up. No such behavioral effects were observed in the control group, and overall effect sizes were in the medium range. Concomitantly, we also found training-related neurophysiological suppression of the DMN, which is implicated in mind-wandering and rumination, and the extent of DMN suppression related to significant improvements in self-compassion.

A recent meta-analytic review of 27 studies [72] and randomized controlled trials found that smartphone apps can be used to enhance mindfulness and compassion skills, as well as reduce stress [73,74]. Additionally, a recent scoping review concluded that it is feasible to deliver compassionate care within digital health care, particularly telemedicine [75]. Yet, previous digital interventions in this field have reported small effect sizes [72]. In this study, we found medium effect size behavioral changes that further correlated with the extent of digital training completed across human participants, highlighting the advantages of this closed-loop digital training. It is also notable that the medium effect size improvements in self-compassion were sustained at follow-up; this finding is aligned with evidence from our past digital intervention studies showing robust long-term behavioral effects [22,76]. The core element of our digital WellMind training is the attention to breathing practice, which is also a foundational element of traditional meditation; indeed, meditation practice has been shown to enhance state mindfulness [77,78]. Yet, traditional practice lacks real-time feedback and performance-based level progressions and can have variable outcomes depending on the teacher [79,80]. WellMind is a teacher-independent digital app within which user feedback is key to maintaining engagement over multiple sessions [24]. Notably, as users engaged with more sessions, we also observed them progress to higher levels of breath monitoring (maximum monitoring of 10 breaths at a time)—a significant correlation was found between breath monitoring level and training sessions completed. At higher training levels, progressively more sophisticated compassion instructions were also unveiled, which may be driving the correlation between the number of training sessions completed and improvements in self-compassion. Furthermore, when we investigated the positive (CSR) and negative (USR) subcomponents of the self-compassion scale [71], we found that reduction in USR (ie, reduction in self-judgment, isolation, and overidentification) exclusively underlies the training-related improvement in total self-compassion. To the best of our knowledge, previous research has not addressed the differential intervention-related plasticity of CSR versus USR components but has shown that these are distinct constructs that differentially relate to well-being and cognitive responses to daily life problems [81,82].

With regard to real-time feedback, WellMind integrates breath monitoring consistency-based auditory feedback as a gentle chime that signals the user to return to attentive breath monitoring when distracted. Indeed, auditory feedback during focused attention meditation has been shown to improve state mindfulness [83]. Notably, in large sample surveys of perceived barriers to meditation across hundreds of participants, a lack of individualized feedback and progress tracking have been prominently cited as important hindrances [84,85]. WellMind was designed to remove these barriers and, hence, motivate training. It is also distinct from previous closed-loop digital meditation approaches that prompt the user to retrospectively and subjectively report whether they were attentive or distracted during their meditation practice and did not integrate compassion cultivation [22,23].

In addition to behavioral outcomes, we also investigated EEG-based neurophysiological outcomes evaluated on an interoceptive breath monitoring task. The eyes-closed interoceptive task required monitoring of 2 breaths at a time and was devoid of feedback or performance-adaptive levels, and thus served as a pure assessment of breath-focused interoception that was targeted by the WellMind training. Consistent with our hypothesis based on previous evidence [28,29,31], we observed training-related suppression of cortical source-localized DMN activity, with no such changes found in the control group. The DMN outcome had a large between-group effect size. In contrast, the FPN and CON executive control networks did not show training-related changes; the FPN showed a significant within-group post versus pre activity reduction in the control group that could be related to attentional lapses or boredom during repeat assessments in this group [22,86,87]. Finally, neurobehavioral correlations showed that training-related DMN activity suppression was significantly related to improvement in self-compassion, specifically reduction in the USR component. Also, individuals with high DMN activity at baseline experienced greater DMN suppression with training, suggesting baseline DMN activity as an individual-specific neural marker to determine who may benefit most from such training. Overall, it is worth noting that we obtained these results using EEG, a scalable approach for measuring neural markers relative to other neuroimaging tools.
modalities such as functional magnetic resonance imaging. We also note that in our previous work, we have shown high test-retest reliability of such EEG data collected within the context of cognitive tasks [43].

As per the limitations of this study, the sample size and nonactive control group are the primary limitations, although it is worth noting that this is the first ever implementation of the digital WellMind intervention in health care professionals. We also acknowledge other limitations of our intervention design, such as that participants were not blind to the intervention condition and that WellMind participants received a greater frequency of contact from the experimenters during the intervention (email reminders to encourage adherence), while the control group had no such contact. These differences may lead to differences in expectations, resulting in positive outcomes observed in the WellMind group. Also, burnout in our sample was low, and we did not observe any post versus pre changes in burnout. Yet, previous work has suggested that organization-directed workplace interventions could be more effective at addressing burnout [25] and a recent study showed that physical exercise facilitated by mobile app technology also benefits burnout [88]. In this context, our research presents the possibility that individual interventions focused on self-compassion and mindfulness delivered before the development of burnout may be helpful in burnout prevention. Our data, showing long-term benefits in self-compassion with moderate to large effect sizes, suggests this approach may be useful in preventing burnout. Yet, this field has acknowledged that there are challenges to implementing behavioral interventions at the organizational level, especially in health care [8]. These challenges include a clear lack of time for physicians as well as concerns regarding confidentiality and discrimination within this workforce. In this context, WellMind offers brief training sessions (5–10 minutes in duration) that can be engaged flexibly at any time of day. Its closed-loop features promote adherence, although we acknowledge that not everyone is similarly motivated, as reflected by the variable number of sessions completed by the training participants. Security and confidentiality should be a paramount concern for any digital intervention [89,90]. Hence, WellMind access is enabled on a secure, password-protected HIPAA-compliant platform, with only deidentified data available for review and analyses. These features can facilitate future adoption. Overall, it is recommended that such interventions are embedded within the early stages of physician education as a foundational resource, when there may be more availability of time, to enhance receptivity, adherence, and build resilience to future burnout.

In conclusion, this study showcases an accessible and closed-loop approach to foster compassion and mindfulness among health care professionals. We found significant behavioral effects as well as robust brain plasticity related to these effects. These findings encourage future scale-up of the digital intervention to promote physician well-being and prevent burnout.

Acknowledgments
This work was supported by grants from the Sanford Institute for Empathy and Compassion at the University of California, San Diego (JM) and Sanford Institute for Empathy and Compassion Postdoctoral Fellowship Award (SJ). We thank Alankar Misra for the software development of the BrainE software and several University of California San Diego undergraduate students who assisted with data collection. We also thank Pragathi Balasubramani and Gillian Grennan for their assistance with the initial project set-up. The BrainE software is copyrighted for commercial use (Regents of the University of California Copyright #SD2018-816) and free for research and educational purposes.

Data Availability
The data set generated and analyzed in this study are available from the Data Dryad repository [91].

The WellMind digital intervention on the BrainE platform app is available for health care education and research on the website and is copyrighted for commercial use (Regents of the University of California Copyright #SD2024-132) [92].

Authors’ Contributions
DR and JM contributed to the conception and design of the study. SJ conducted data management and data analytics and wrote the first draft of the manuscript. SRP and JKM collected and managed the data; JN and NA managed and analyzed the data. JM supervised all data collection, data management, analytics, and manuscript writing. All authors contributed to manuscript edits and revisions and approved the submitted version.

Conflicts of Interest
None declared.

Multimedia Appendix 1 Supplementary methods.
[DOCX File, 22 KB - mental_v11i1e49467_app1.docx]
References


39. Misra A, Ojeda A, Mishra J. BrainE: a digital platform for evaluating, engaging and enhancing brain function. Regents of the University of California 2018-816. 2018. URL: https://escholarship.org/content/qt7pm2c2vp/qt7pm2c2vp_noSplash_d2716684ae452e12e2b0c429a1b7e1a.pdf [accessed 2023-12-20]


47. Scientific scales and measures. Greater Good Science Center B. 2003. URL: https://ggsc.berkeley.edu/who_we_serve/researchers/scientific_scales_and_measures [accessed 2023-04-20]


60. Ding L, He B. Sparse source imaging in electroencephalography with accurate field modeling. Hum Brain Mapp 2008 Sep;29(9):1053-1067 [FREE Full text] [doi: 10.1002/hbm.20448] [Medline: 17894400]


Abbreviations

CON: cingulo-opercular network
CSR: compassionate self-responding
DMN: default mode network
EEG: electroencephalography
FPN: fronto-parietal network
HIPAA: Health Insurance Portability and Accountability Act
MBI: Maslach Burnout Inventory
RT: response time
UCSD: University of California San Diego
USR: uncompassionate self-responding

©Satish Jaiswal, Suzanna R Purpura, James K Manchanda, Jason Nan, Nihal Azeez, Dhakshin Ramanathan, Jyoti Mishra. Originally published in JMIR Mental Health (https://mental.jmir.org), 22.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Mental Health, is properly cited. The complete bibliographic information, a link to the original publication on https://mental.jmir.org/, as well as this copyright and license information must be included.
Research Letter

The Frequency of Design Studies Targeting People With Psychotic Symptoms and Features in Mental Health Care Innovation: Secondary Analysis of a Systematic Review

Lars Veldmeijer1,2,3, MA; Gijs Terlouw2, PhD; Jim Van Os1, PhD; Job Van 't Veer2, PhD; Nynke Boonstra1,2,3, PhD

1Department of Psychiatry, Utrecht University Medical Center, Utrecht, Netherlands
2Department of Healthcare and Welfare, NHL Stenden University of Applied Sciences, Leeuwarden, Netherlands
3KieN VIP Mental Health Care Services, Leeuwarden, Netherlands

Corresponding Author:
Lars Veldmeijer, MA
Department of Psychiatry
Utrecht University Medical Center
Heidelberglaan 100
Utrecht, 3584 CX
Netherlands
Phone: 31 68 287 4768
Email: lars.veldmeijer@nhlstenden.com

Abstract

This study examined and reflected on the frequency of people with psychotic symptoms and features as the target population in design studies for mental health care innovation.

(JMIR Ment Health 2024;11:e54202) doi:10.2196/54202

KEYWORDS
design approaches; design; innovation; innovative; innovate; innovations; psychiatry; mental health care; mental health; mental illness; mental disease; involvement; service users; people with lived experience; people with lived experiences; lived experience; lived experiences; co-creation; cocreation; psychosis; psychotic; schizophrenia; schizoid; schizotypal; paranoia; neurosis; hallucinosis; hallucination; hallucinations

Introduction

There is growing evidence highlighting the importance of involving people with lived experience in design processes in mental health care [1,2]. Particular attention should be directed toward the engagement of people with psychotic symptoms and features [3], as they often feel misunderstood due to their altered perceptions and subjective experiences [4,5]. A bottom-up review of the lived experience of psychosis emphasizes the complexity of psychotic symptoms and features and recommends including lived experience in designing mental health services to address these experiences and needs [6]. Design approaches can promote the involvement of people with firsthand experiences in the development of treatment, therapy, and recovery interventions for mental health care innovation [2]. Currently, it is unknown how frequent design studies specifically target people with psychotic symptoms and features. There is a scoping review of coproducing research on psychosis [7], but coproduction and design approaches are distinct methodologies. Design approaches facilitate designing initiatives that prioritize participants’ needs, expertise, and knowledge whereas coproduction facilitates collaborative delivery and knowledge production. In this research letter, we present findings on the frequency of design studies targeting people with psychotic symptoms by analyzing a prior systematic review data set that focused on involving people with firsthand experiences in designing mental health care innovations. The primary objective of this secondary data analysis was to elucidate how often design studies in mental health care target people with psychotic symptoms and features.

Methods

Primary Data Set and Secondary Data Analysis

We conducted a secondary data analysis using a data set from a prior systematic review that assessed the involvement of service users and people with lived experience in the design processes of mental health care innovation. In the screening process and study selection of the prior systematic review, 33...
papers met the inclusion criteria [2]. All included papers were original reports or papers that (1) involved service users, people with lived experience, or both; (2) mentioned design approaches; (3) involved an empirical study; and (4) conducted the study in settings including mental health care services or psychiatry programs. In this secondary analysis, we examined the primary data set to provide an overview of the frequency of design studies in mental health care focusing on people with psychotic symptoms and features. This data set is suitable for this analysis since the search strategy of the systematic review did not target specific mental health conditions.

Data Extraction and Categorization

Studies were categorized based on their primary target population as reported in the studies (Multimedia Appendix 1). We categorized broad terms like psychosis, which encompasses various symptoms and features like altered perceptions, as well as mental health conditions in which psychotic symptoms and features are prevalent, such as schizophrenia. In the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition, Text Revision (DSM-5-TR), these symptoms, features, and conditions fall under the category “schizophrenia spectrum and other psychotic disorders,” covering a spectrum of related mental health conditions [8], also referred to as the psychosis spectrum [9]. Studies addressing psychotic symptoms and features or related conditions alongside unrelated mental health conditions were labeled “various mental health conditions” due to their comorbid nature. We did not count these studies as primarily focusing on people with psychotic symptoms and features.

Table 1. Target populations in mental health design studies.

<table>
<thead>
<tr>
<th>Target population</th>
<th>Count, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Psychosis</td>
<td>6 (18)</td>
</tr>
<tr>
<td>Depression</td>
<td>4 (12)</td>
</tr>
<tr>
<td>Schizophrenia</td>
<td>3 (9)</td>
</tr>
<tr>
<td>Self-harm</td>
<td>2 (6)</td>
</tr>
<tr>
<td>Eating disorders</td>
<td>2 (6)</td>
</tr>
<tr>
<td>Substance use disorders</td>
<td>1 (3)</td>
</tr>
<tr>
<td>Borderline</td>
<td>1 (3)</td>
</tr>
<tr>
<td>Attention-deficit/hyperactivity disorder</td>
<td>1 (3)</td>
</tr>
<tr>
<td>Autism spectrum disorder</td>
<td>1 (3)</td>
</tr>
<tr>
<td>Bipolar disorder</td>
<td>1 (3)</td>
</tr>
<tr>
<td>Various mental health conditions</td>
<td>11 (34)</td>
</tr>
</tbody>
</table>

Discussion

This secondary data analysis revealed a notable emphasis on studies primarily targeting people with psychotic symptoms and features in mental health care design studies. This is noteworthy given the extensive range of mental health conditions in psychiatry, encompassing 21 categories according to the DSM-5-TR [8]. Although “schizophrenia spectrum and other psychotic disorders” constitutes only 4.67% (1/21) of these categories, 27% (9/33) of studies in our data set focused primarily on psychotic symptoms and features. This percentage is high, considering the lifetime prevalence of psychotic disorders is approximately 1% [10]. Another 12% (4/33) of the studies mention psychotic symptoms and features alongside or as a result of other mental health conditions. Although these studies did not focus primarily on people with psychotic symptoms and features, they have shown that much attention has been given to psychotic experiences in design studies. The substantial research focus on people with psychotic symptoms and features in design studies may be attributed to the limited progress in prognosis for severe cases despite extensive research and treatment efforts [11]. This may prompt designers and
researchers to look for less conventional strategies to enforce novel promising solutions. Additionally, there is a growing call for attention to the subjective experience of psychotic symptoms and features in clinical care, as these vary from individual to individual (eg, [4-6]). Both factors underscore the urgency of involving people with firsthand experiences to capture the vividness of psychotic experiences in the design of innovative services and interventions, ultimately aiming to improve outcomes for service users.

Comparing the 9 studies that primarily focused on people with psychotic symptoms and features in this secondary data analysis to the results of the prior systematic review, we observed that 44% (4/9) demonstrated a high level of participant involvement in their design processes [2]. This is crucial for the development of new innovations because research shows psychotic symptoms and features can seem very different from a lived experience perspective compared to conventional psychiatric conceptualizations [6]. At the same time, the results stress the ongoing need to engage people with lived experience of psychotic symptoms and features in design studies, as more than half of the studies did not show the substantial involvement that would be expected of design processes that aim to tailor innovations to the needs of the target group. Consequently, we recommend future design studies targeting people with psychotic symptoms and features to adopt the co-design methodology, as co-design shows the highest participant involvement levels in mental health care design studies [2]. Furthermore, researchers are encouraged to use the participation matrix [12] alongside co-design to make intentional methodological decisions regarding the phases and roles in which people with lived experience are involved. To prevent tokenism and cooptation in design processes, researchers and designers are recommended to systematically coreflect with people with lived experience, exploring the roles played and distilling benefits and challenges from both perspectives.

Acknowledgments

We appreciate the financial support of the FAITH Research Consortium, GGZ-VS University of Applied Science, as well as from the NHL Stenden University of Applied Sciences PhD program. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Categorization of studies based on their target population.

References


Abbreviations

DSM-5-TR: Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition, Text Revision
Comparing the Perspectives of Generative AI, Mental Health Experts, and the General Public on Schizophrenia Recovery: Case Vignette Study

Zohar Elyoseph¹,², BA, MA, PhD; Inbar Levkovich³, BA, MA, PhD, Prof Dr

Abstract

Background: The current paradigm in mental health care focuses on clinical recovery and symptom remission. This model’s efficacy is influenced by therapist trust in patient recovery potential and the depth of the therapeutic relationship. Schizophrenia is a chronic illness with severe symptoms where the possibility of recovery is a matter of debate. As artificial intelligence (AI) becomes integrated into the health care field, it is important to examine its ability to assess recovery potential in major psychiatric disorders such as schizophrenia.

Objective: This study aimed to evaluate the ability of large language models (LLMs) in comparison to mental health professionals to assess the prognosis of schizophrenia with and without professional treatment and the long-term positive and negative outcomes.

Methods: Vignettes were inputted into LLMs interfaces and assessed 10 times by 4 AI platforms: ChatGPT-3.5, ChatGPT-4, Google Bard, and Claude. A total of 80 evaluations were collected and benchmarked against existing norms to analyze what mental health professionals (general practitioners, psychiatrists, clinical psychologists, and mental health nurses) and the general public think about schizophrenia prognosis with and without professional treatment and the positive and negative long-term outcomes of schizophrenia interventions.

Results: For the prognosis of schizophrenia with professional treatment, ChatGPT-3.5 was notably pessimistic, whereas ChatGPT-4, Claude, and Bard aligned with professional views but differed from the general public. All LLMs believed untreated schizophrenia would remain static or worsen without professional treatment. For long-term outcomes, ChatGPT-4 and Claude predicted more negative outcomes than Bard and ChatGPT-3.5. For positive outcomes, ChatGPT-3.5 and Claude were more pessimistic than Bard and ChatGPT-4.

Conclusions: The finding that 3 out of the 4 LLMs aligned closely with the predictions of mental health professionals when considering the “with treatment” condition is a demonstration of the potential of this technology in providing professional clinical prognosis. The pessimistic assessment of ChatGPT-3.5 is a disturbing finding since it may reduce the motivation of patients to start or persist with treatment for schizophrenia. Overall, although LLMs hold promise in augmenting health care, their application necessitates rigorous validation and a harmonious blend with human expertise.

(JMIR Ment Health 2024;11:e53043) doi:10.2196/53043

KEYWORDS

schizophrenia; mental; prognostic; prognostics; prognosis; ChatGPT; artificial intelligence; recovery; vignette; vignettes; outcome; outcomes; large language models; language model; language models; LLM; LLMs; NLP; natural language processing; GPT; Generative Pre-trained Transformers

Introduction

Background

Schizophrenia is a major contributor to mental health–related disability worldwide and exerts a profound effect on patients and society [1]. It has a major impact on life expectancy and quality of life, and its repercussions extend to family and caregivers [2]. The disorder presents a complex array of symptoms, both “positive” (eg, delusions and hallucinations) and “negative” (eg, emotional flatness and social withdrawal) [3]. Negative symptoms are especially resistant to current treatments [4]. Despite the complexity and impact of this disorder, a subset of individuals with schizophrenia may have a favorable prognosis; exhibit symptom reduction; and achieve positive outcomes in education, employment, and relationships [3].
A systematic review based on 37 studies that examined outcomes in first-episode psychosis [5] determined that 42% of patients experienced favorable outcomes. Similarly, an analysis of 114 follow-up studies to assess recovery rates in schizophrenia [6] yielded recovery rates ranging from 11% to 33% for complete recovery and from 22% to 53% for social recovery. Another meta-analysis [7] encompassing 50 pertinent studies revealed that approximately one-seventh of individuals diagnosed with schizophrenia met our predefined criteria for recovery.

The beliefs held by therapists regarding their patients’ capacity for recuperation represent a complex and multifaceted conundrum [8]. From a pragmatic standpoint, a medical practitioner’s proficiency in ascertaining a patient’s prospective therapeutic trajectory (known as “prognosis” in the medical field) is a major clinical aptitude [9]. From an ethical perspective, clinicians are duty bound to elucidate potential perils and advantages to patients, thus facilitating a process of informed consent and collaborative decision-making [10]. Providing a nuanced yet candid prognosis enhances patient motivation and optimism when the likelihood of complete remission is high, while concurrently calibrating expectations in less promising scenarios [11-13]. Nevertheless, inherent values and presuppositions inevitably shape prognostic assessments [14,15]. The etiology and treatability of psychiatric disorders are framed by 2 opposing philosophical paradigms. Deterministic models, which view mental disorders as fixed biological anomalies, often adopt a pessimistic perspective on full recovery. In contrast, the recovery model approach is rooted in the belief that complete recovery is achievable. This perspective emphasizes personal empowerment, resilience, and community integration, focusing on an individual’s potential rather than solely on their symptoms [6,14,15]. Dogmatic adherence to either of these viewpoints carries the risk of engendering self-realizing outcomes. Hence, therapists must balance their understanding of empirical medical data by acknowledging the vast spectrum of human potentialities [16,17].

In November 2022, the generative artificial intelligence (AI) large language model (LLM) ChatGPT-3 was launched for free public use. Subsequently, in 2023, other LLMs such as Google Bard, Claude, and ChatGPT-4 were released. Although all these LLMs have been trained on vast data sets and have undergone alignment processes, as well as learning from user feedback, their differences stem from their unique learning algorithms; the nature of their training data; and the distinct approaches to alignment, user interaction, and learning from user feedback. These LLMs have permeated various facets of society, including political science, economics, health care, and biology [18,19]. Previous studies have examined the potential of LLMs in the context of applied psychology, focusing on basic clinical abilities [20-22] or on decision-making in complex clinical situations such as depression and suicide [23-25]. To the best of our knowledge, no study to date has assessed the extent to which generative AI can facilitate cure or recovery from mental health conditions. In contrast, extensive literature highlights the immense therapeutic value of therapists’ belief in their patients’ ability to recover [11-13], as well as the negative effects that ensue when a therapist does not believe that the patient’s condition can improve [26].

Recovery for individuals with prolonged mental health challenges is a multifaceted process subject to varied interpretations. From a clinical perspective, recovery emphasizes symptom reduction and impairment rectification [26,27]. In contrast, from lived experience, recovery represents an individualized, potentially ongoing trajectory toward reclaiming purpose, meaning, and active contribution, regardless of symptoms [27].

Years of rigorous theoretical and clinical research have revealed several mechanisms that assist patient recovery. One salient finding is the positive correlation between a strong therapeutic alliance and enhanced outcomes [28]. A meta-analysis of over 30,000 participants showed the therapeutic alliance was highly correlated with outcomes, regardless of therapy type [29]. The efficacy of psychotherapy is well documented [30-32]. Therapists’ belief in treatment potential significantly impacts outcomes [33]. Over the past decade, literature has consistently emphasized recovery-oriented practices for improving patient outcomes, including enhanced functioning, goal setting, reduced legal issues, and decreased hospital admissions [34,35]. Consequently, mental health services increasingly integrate recovery paradigms into treatment strategies [36]. However, an abrupt transition from a biomedical model to recovery orientation can challenge providers, often leading to continued paternalistic decision-making [37].

With the increasing integration of AI in health care, especially given its emerging capabilities in emotion detection and mental health risk assessment [20-25], it becomes imperative to scrutinize how different LLMs interpret human recovery potential. Such an inquiry gains heightened relevance in that both patients and professionals are increasingly relying on LLMs for consultations. Not only do such insights have the potential to shape the trajectory of patient care, but they can also play a pivotal role in psychoeducational endeavors, direction, and interventions.

This research is based on an examination of the perspectives of mental health professionals in Australia [38]. The study included 342 nurses, 564 psychiatrists, 424 general practitioners (GPs), and 228 clinical psychologists. It also incorporated the insights of 982 members of the general public. Respondents were presented vignettes depicting an individual diagnosed with schizophrenia and asked to indicate their perceptions regarding prognosis, long-term outcomes, and potential discrimination.

Research Objectives

The research objectives were as follows:

1. To examine how different LLMs (ChatGPT-3.5, ChatGPT-4, Claude, and Bard) evaluate the prognosis of an individual with schizophrenia compared to the evaluations of mental health professionals (mental health nurses, clinical psychologists, psychiatrists, and GPs) and the general public.

2. To examine how different LLMs (ChatGPT-3.5, ChatGPT-4, Claude, and Bard) evaluate the positive and negative long-term outcomes of an individual with schizophrenia compared to the evaluations of mental health
professionals (mental health nurses, clinical psychologists, psychiatrists, and GPs) and the general public.

3. To compare evaluations of the prognosis and positive and negative outcomes of an individual with schizophrenia between different types of LLMs (ChatGPT-3.5, ChatGPT-4, Claude, and Bard).

**Methods**

**AI Procedure and Data Collection**

During the month of August 2023, we examined the following LLMs:

- **Bard** (Google; subsequently rebranded as Gemini) [39] uses the LaMDA language model, trained on the expansive Infiniset data set amalgamating over 1.5 trillion words from diverse web-based sources including C4-derived content, Wikipedia, programming documentation, and public forum dialogue. LaMDA was initially pretrained on extensive public and web text corpora, leveraging a transformer-based neural architecture and unsupervised learning to process language and formulate pertinent responses.

- **Claude** (Anthropic) [40] targets a beneficial, inoffensive, and truthful output using a constitutional approach. Its 12+ billion-parameter transformer model aims to ethically tackle linguistic complexity. Its training emphasized educational data curation and the alignment of model actions with human values and safety considerations, potentially enhancing reliability. A paid Claude subscription recently launched at US $20 per month. Our study used the free version.

- **ChatGPT-3.5** (OpenAI) [41] exemplifies a 175 billion–parameter transformer model demonstrating human-like language aptitude given its far-reaching internet-derived training data. It was designed to expand comprehension and text generation capabilities using immense data sets for advanced language understanding and tailored response building functionalities.

- **ChatGPT-4** (OpenAI) [42] aims to improve upon previous versions in areas such as safety, accuracy, and content policy adherence. Relative to ChatGPT-3.5, OpenAI claims an over 80% decrease in rule-violating responses and a 40% improvement in factual correctness for this latest model. As part of the training process, OpenAI solicited abundant feedback from early users and collaborated with over 50 AI ethics experts. However, full details on the architecture changes and training data incorporated have not been disclosed publicly. Without access to such technical specifics, independent auditing remains limited regarding the capabilities of ChatGPT-4. This model costs US $20 per month.

The case vignettes and a list of follow-up questions were introduced for each of the LLMs. To address the generative nature of the models, which produces different responses to similar questions, each vignette was passed through each model 10 times. We compared the LLMs’ data for schizophrenia prognosis to the norms used by mental health professionals and to the views of the general public, as reported by Caldwell and Jorm [38].

**Input Source**

In this investigation, we adopted the text vignette methodology used by Jorm et al [43] and Caldwell and Jorm [38] to examine the beliefs of mental health professionals (mental health nurses, clinical psychologists, psychiatrists, and GPs) and the general public regarding the efficacy of interventions for schizophrenia. In the study, 2 vignettes were entered into the LLMs’ interface. The 2 vignettes differed only in the gender of the main character. Each vignette was evaluated 10 times in ChatGPT-3.5, ChatGPT-4, Bard, and Claude, for a total of 80 evaluations, each in a new tab.

**Vignette**

The full text of the vignettes is as follows:

*Mary/Jone is 24 and lives at home with her/his parents. She/he has a few temporary jobs since finishing school but is now unemployed. Over the last six months she/he has stopped seeing her/his friends and has begun locking herself/himself in her/his bedroom and refusing to eat with the family or to have a bath. Her/His parents also hear her/him walking about her bedroom at night while they are in bed. Even though they know she/he is alone, they have heard her/him shouting and arguing as if someone else is there. When they try to encourage her/him to do more things, she/he whispers that she/he won’t leave home because she/he is being spied upon by the neighbor. They realize she/he is not taking drugs because she/he never sees anyone or goes anywhere.*

**Measures**

After entering each of the 2 vignettes to the LLMs, we submitted 6 questions taken from Jorm et al [43]:

**Q1:** Describe what, if anything, is wrong with Mary/Jone?

**Q2:** What will be helpful for Mary/Jone?

**Q3:** What is the most likely result for Mary/Jone given that she/he received the sort of professional help you believe is most appropriate? A. Full recovery with no further problems; B. Full recovery, but problems will probably recur; C. Partial recovery; D. Partial recovery, but problems will probably recur; E. No improvement; F. Get worse.

**Q4:** What is the most likely result for Mary/Jone if she/he does not receive any professional help? A. Full recovery with no further problems; B. Full recovery, but problems will probably recur; C. Partial recovery; D. Partial recovery, but problems will probably recur; E. No improvement; F. Get worse.

**Q5:** How do you think Mary/Jone will be in the long term compared to other people in the community? Rate each of the following statements according to this scale: more likely, just as likely, or less likely.

Positive long-term outcomes: A. will be violent; B. will drink too much; C. will take illegal drugs; D. will have unsatisfactory friendships; E. will attempt suicide.

Negative long-term outcomes: F. will be
understanding of other people’s feelings; G. will have a good marriage; H. will be a caring parent; I. will be a productive worker. J. will be creative or artistic.

Q6. Do you think Mary/Jone will be discriminated against by others in the community if they know about her/his problems? (Yes/No).

Scoring

The performance of each LLM was scored according to Jorm et al [43] and Caldwell and Jorm [38]. We then compared the performance of the LLMs to the norms of 324 mental health nurses, 228 clinical psychologists, 567 psychiatrists, 424 GPs, and 982 people from the general public, as collected in Australia [38,43]. Q5, which evaluated the positive and negative long-term outcomes, was calculated according to Caldwell and Jorm [38]. Each of the 10 statements was scored as follows: 1=more likely, 0=just as likely, and −1=less likely. The answers were then summed up, such that each positive and negative long-term outcome score ranged from −5 to 5.

Statistical Analysis

The likely outcomes with and without professional treatment for the 2 vignettes, as evaluated by the LLMs, mental health professionals, and the general public (reported by Caldwell and Jorm [38,43]), were analyzed using 1-way ANOVA, with Fisher least significant difference applied as a post hoc analysis. The differences between the LLMs in positive and negative long-term outcomes were compared using 1-way ANOVA, with Fisher least significant difference applied as a post hoc analysis. Given the significant clinical implications of discrepancies between the evaluations of the LLM models and the professional assessments, we opted for a post hoc approach that minimizes the risk of type II errors or false negatives.

Ethical Considerations

This study was exempt from ethical review since it only evaluates AI chatbots and no human participants were involved.

Results

For all of the vignette cases, all 4 LLMs recognized schizophrenia as the primary diagnosis and suggested a combination of antipsychotic drugs and psychotherapy as the preferred treatment.

Likely Outcome With Professional Treatment

Table 1 delineates the distribution of outcomes selected by LLMs, mental health professional groups, and the general public for a vignette describing an individual diagnosed with schizophrenia after receiving professional treatment. ANOVA analysis revealed significant differences in the selected outcomes across the 8 groups ($F_{8,2601}=33.66; P<.001$). Post hoc analysis yielded the following insights. (1) The ChatGPT-3.5 model offered a distinctively pessimistic prognosis, significantly differing from the outcomes chosen by all the other LLMs ($P=.02$ to .007), the professional groups ($P=.005$ to <.001), and the general public ($P<.001$). (2) ChatGPT-4, Claude, and Bard projected more pessimistic prognosis outcomes than the general public ($P=.02$ to .007), whereas their projections were congruent with those from all the professional groups (all $P>.05$). A direct comparison of the projections of ChatGPT-4, Claude, and Bard yielded no significant differences (all $P>.05$; Figure 1 and Table 2).
The likely outcome for schizophrenia, with and without professional treatment, as evaluated by LLMs\(^a\), mental health professionals, and the general public.

| Professional treatment and outcome | ChatGPT-3.5 (n=20), n (%) | ChatGPT-4 (n=20), n (%) | Bard (n=20), n (%) | Claude, Psychiatrists (n=567), n (%) | General public (n=982), %\(^b\) | Nurses (n=324), %\(^b\) | Clinical psychologists (n=228), %\(^b\) | Psychologists (n=567), %\(^b\) | GPs\(^c\) (n=424), %\(^b\) |
|-----------------------------------|---------------------------|--------------------------|-------------------|--------------------------------------|--------------------------------|----------------|-------------------------------|----------------|----------------|----------------|----------------|
| **With professional treatment**   |                           |                          |                   |                                      |                               |                 |                               |                 |                 |                 |                 |
| Full recovery, no further problems | 0 (0)                    | 0 (0)                    | 1 (5)             | 0 (0)                                | 29.8                          | 8.8             | 3.1                           | 2               | 3.1             |                 |                 |
| Full recovery, but problems would probably recur | 0 (0) | 5 (25) | 4 (20) | (35) | 44.4 | 61.4 | 49.1 | 51.6 | 56.1 |
| Partial recovery                   | 10 (50)                   | 15 (75)                  | 15 (75)           | (65)                                 | 10.2                          | 4.1             | 11.9                          | 5.7             | 5               |                 |                 |
| Partial recovery, but problems would probably recur | 10 (50) | 0 (0) | 0 (0) | 0 (0) | 14.3 | 25.7 | 35.4 | 40.6 | 35.8 |
| No improvement                     | 0 (0)                     | 0 (0)                    | 0 (0)             | 0 (0)                                | 0.7                           | 0               | 0.4                           | 0.2             | 0               |                 |                 |
| Get worse                          | 0 (0)                     | 0 (0)                    | 0 (0)             | 0 (0)                                | 0.6                           | 0               | 0                             | 0               | 0               |                 |                 |
| **Without professional treatment**|                           |                          |                   |                                      |                               |                 |                               |                 |                 |                 |                 |
| Full recovery, no further problems | 0 (0)                     | 0 (0)                    | 0 (0)             | 0 (0)                                | 1.1                           | 0               | 0                             | 0               | 1.1             |                 |                 |
| Full recovery, but problems would probably recur | 0 (0) | 0 (0) | 0 (0) | 0 (0) | 1.7 | 0.9 | 0.9 | 0.7 | 1.7 |
| Partial recovery                   | 0 (0)                     | 0 (0)                    | 0 (0)             | 0 (0)                                | 1.8                           | 0.6             | 0.9                           | 0.9             | 1.8             |                 |                 |
| Partial recovery, but problems would probably recur | 0 (0) | 0 (0) | 0 (0) | 0 (0) | 4.9 | 11.3 | 9.1 | 5.7 | 4.9 |
| No improvement                     | 3 (15)                    | 0 (0)                    | 0 (0)             | 0 (0)                                | 15.1                          | 9.5             | 17.8                          | 11              | 15.1            |                 |                 |
| Get worse                          | 17 (85)                   | 20 (100)                 | 20 (100)          | 20 (100)                             | 75.4                          | 77.7            | 71.3                          | 81.8            | 75.4            |                 |                 |

\(^a\)LLM: large language model.
\(^b\)As reported by Caldwell and Jorm [38].
\(^c\)GP: general practitioner.
Figure 1. The likely outcome for schizophrenia, with and without professional treatment, as evaluated by large language models, mental health professionals, and the general public (mean and SE). *P<.05. GP: general practitioner.
Table 1. Least significant difference post hoc analyses for LLMs, mental health professionals, and the general public in assessing the outcome of schizophrenia with and without treatment.

<table>
<thead>
<tr>
<th>P values</th>
<th>ChatGPT-3.5</th>
<th>ChatGPT-4</th>
<th>Bard</th>
<th>Claude</th>
<th>General public</th>
<th>Nurses</th>
<th>Clinical psychologists</th>
<th>Psychiatrists</th>
<th>GPs</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>With professional treatment</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ChatGPT-3.5</td>
<td>—</td>
<td>.02</td>
<td>.01</td>
<td>.007</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td>.003</td>
<td>.005</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>ChatGPT-4</td>
<td>.02</td>
<td>—</td>
<td>.87</td>
<td>.75</td>
<td>.007</td>
<td>.21</td>
<td>.79</td>
<td>.64</td>
<td>.95</td>
</tr>
<tr>
<td>Bard</td>
<td>.01</td>
<td>.87</td>
<td>—</td>
<td>.87</td>
<td>.01</td>
<td>.30</td>
<td>.63</td>
<td>.49</td>
<td>.87</td>
</tr>
<tr>
<td>Claude</td>
<td>.007</td>
<td>.75</td>
<td>.87</td>
<td>—</td>
<td>.02</td>
<td>.41</td>
<td>.49</td>
<td>.37</td>
<td>.70</td>
</tr>
<tr>
<td>General public</td>
<td>&lt;.001</td>
<td>.007</td>
<td>.01</td>
<td>.02</td>
<td>—</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Nurses</td>
<td>&lt;.001</td>
<td>.21</td>
<td>.30</td>
<td>.41</td>
<td>&lt;.001</td>
<td>—</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Clinical psychologists</td>
<td>.003</td>
<td>.79</td>
<td>.63</td>
<td>.49</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td>—</td>
<td>.58</td>
<td>.08</td>
</tr>
<tr>
<td>Psychiatrists</td>
<td>.005</td>
<td>.64</td>
<td>.49</td>
<td>.37</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td>.58</td>
<td>—</td>
<td>.06</td>
</tr>
<tr>
<td>GPs</td>
<td>&lt;.001</td>
<td>.95</td>
<td>.87</td>
<td>.70</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td>.08</td>
<td>.06</td>
<td>—</td>
</tr>
<tr>
<td><strong>Without professional treatment</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ChatGPT-3.5</td>
<td>—</td>
<td>.55</td>
<td>.55</td>
<td>.55</td>
<td>.12</td>
<td>.21</td>
<td>.14</td>
<td>.46</td>
<td>.48</td>
</tr>
<tr>
<td>ChatGPT-4</td>
<td>.55</td>
<td>—</td>
<td>&gt;.99</td>
<td>&gt;.99</td>
<td>.02</td>
<td>.04</td>
<td>.02</td>
<td>.11</td>
<td>.13</td>
</tr>
<tr>
<td>Bard</td>
<td>.55</td>
<td>&gt;.99</td>
<td>—</td>
<td>&gt;.99</td>
<td>.02</td>
<td>.04</td>
<td>.02</td>
<td>.11</td>
<td>.13</td>
</tr>
<tr>
<td>Claude</td>
<td>.55</td>
<td>&gt;.99</td>
<td>&gt;.99</td>
<td>—</td>
<td>.02</td>
<td>.04</td>
<td>.02</td>
<td>.11</td>
<td>.13</td>
</tr>
<tr>
<td>General public</td>
<td>.12</td>
<td>.02</td>
<td>.02</td>
<td>.02</td>
<td>—</td>
<td>.35</td>
<td>.89</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Nurses</td>
<td>.21</td>
<td>.04</td>
<td>.04</td>
<td>.04</td>
<td>.35</td>
<td>—</td>
<td>&lt;.001</td>
<td>.07</td>
<td>.07</td>
</tr>
<tr>
<td>Clinical psychologists</td>
<td>.14</td>
<td>.02</td>
<td>.02</td>
<td>.02</td>
<td>.39</td>
<td>&lt;.001</td>
<td>—</td>
<td>.03</td>
<td>.03</td>
</tr>
<tr>
<td>Psychiatrists</td>
<td>.46</td>
<td>.11</td>
<td>.11</td>
<td>.11</td>
<td>&lt;.001</td>
<td>.07</td>
<td>.03</td>
<td>—</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>GPs</td>
<td>.48</td>
<td>.13</td>
<td>.13</td>
<td>.13</td>
<td>&lt;.001</td>
<td>.07</td>
<td>.03</td>
<td>&lt;.001</td>
<td>—</td>
</tr>
</tbody>
</table>

^aLLM: large language model.
^bGP: general practitioner.
^cNot applicable.

Likely Outcome Without Professional Treatment

Table 1 also delineates the distribution of outcomes selected by LLMs, mental health professional groups, and the general public for a vignette describing an individual with schizophrenia who did not receive professional treatment. All groups indicated that without treatment, the person with schizophrenia would show no improvement or would get worse. ANOVA analysis revealed a significant difference in the selected outcomes across the 8 groups ($F_{8,260}=4.07; P<.001$). Post hoc analysis yielded the following insights. (1) The ChatGPT-4, Claude, and Bard models offered a distinctively pessimistic prognosis, significantly differing from the outcomes chosen by mental health nurses ($P=.04$), clinical psychologists ($P=.02$), and the general public ($P=.11$) but not significantly different from the outcomes selected by ChatGPT-3.5, psychiatrists, and GPs (all $P>.05$). Direct comparison between ChatGPT-4, Claude, and Bard yielded no significant differences in prognosis (all $P>.05$). (2) No significant difference was observed between ChatGPT-3.5, the professional groups, and the general public (all $P>.05$; Figure 1 and Table 2).

Long-Term Outcomes

Figure 2 illustrates the LLMs' output concerning positive and negative long-term outcomes. ANOVA analysis revealed a significant difference in the negative outcomes selected across the 4 LLMs groups ($F_{3,76}=18.32; P<.001$). ChatGPT-4 and Claude indicated a significantly higher likelihood of negative long-term outcomes for patients after professional treatment than Bard and ChatGPT-3.5 (ChatGPT-4 vs Bard: $P=.004$; ChatGPT-4 vs ChatGPT-3.5: $P<.001$; Claude vs Bard: $P=.003$; Claude vs ChatGPT-3.5: $P<.001$). In addition, Bard was significantly more pessimistic and indicated a higher likelihood of negative long-term outcomes than ChatGPT-3.5 ($P<.001$). ANOVA analysis revealed a significant difference in the positive outcomes selected by the 4 LLMs groups ($F_{3,76}=24.45; P<.001$). ChatGPT-3.5 and Claude were significantly more pessimistic and indicated a lower likelihood of positive long-term outcomes for patients after treatment than Bard and ChatGPT-4 (ChatGPT-3.5 vs Bard: $P<.001$; ChatGPT-3.5 vs ChatGPT-4: $P<.001$; Claude vs Bard: $P<.001$; Claude vs ChatGPT-4: $P<.001$). No significant differences were found between ChatGPT-3.5 and Claude ($P=.92$) or between ChatGPT-4 and Bard ($P=.51$).
Discrimination
For all the cases in the vignettes, all 4 LLMs determined that the person with schizophrenia described in the vignette would be discriminated against.

Discussion
Principal Findings
This investigation pursued 2 primary objectives. First, we aimed to evaluate how various LLMs assessed the prognosis of individuals with schizophrenia compared to the evaluations of mental health professionals (mental health nurses, clinical psychologists, psychiatrists, and GPs) and the views of the general public. Second, we sought to compare these assessments of prognosis as well as positive and negative long-term outcomes across the different types of LLMs.

The academic discourse in contemporary schizophrenia research often focuses on the deployment of AI within professional scientific contexts, yet it seldom addresses the accessibility of AI to the general public or the patient population. LLMs are being used today by hundreds of millions of users worldwide, including patients and clinicians. In the mental health field, this widespread use has awakened an urgent need to examine the quality of clinical information these systems provide on various medical issues, such as treatment strategy recommendations.
[24], risk assessment [23,25], and the interpretation of emotional states [20,21]. Machine learning algorithms possess the capability to discern nuanced variables associated with divergent disease trajectories [44]. Such algorithms facilitate probabilistic prediction of specific outcomes at the individual level, as well as the identification of distinct subgroups within a broader diagnostic category [45]. Consequently, machine learning methodologies hold promise for aiding clinicians in formulating individualized interventions, thereby mitigating the risk of a detrimental disease progression [46]. This study examines an issue not previously investigated—the ability to predict the clinical prognosis of a severe chronic illness such as schizophrenia using LLMs.

**Likely Outcome With Professional Treatment**

In this study, we identified significant differences in the outcomes suggested across the 8 groups. The ChatGPT-3.5 model exhibited a notably pessimistic prognosis for individuals with schizophrenia with professional treatment relative to other LLMs, professional groups, and the general public. Given the widespread use of ChatGPT-3.5, these findings have substantial clinical implications. Any inclination toward pessimistic forecasting might influence a patient’s willingness to undergo treatments, including both antipsychotic medication and psychotherapy, particularly in the context of schizophrenia. If patients or their families consult the ChatGPT-3.5 model for prognostic insights, these less-than-optimistic outcomes might sway their decision-making regarding whether to initiate or continue therapeutic interventions. The literature previously highlighted challenges in adherence to antipsychotic and psychotherapy treatments due to their cost and other factors [47,48]. Additionally, more negative perceptions of potential treatment outcomes might inadvertently influence the effectiveness of the therapeutic process, regardless of the mode of intervention.

The congruence between the prognostic assessments of various LLMs models (ChatGPT-4, Bard, and Claude) and those of clinical professionals is encouraging. From a clinical perspective, precise prognostication is paramount. It not only informs clinicians in tailoring interventions that balance potential risks and benefits but also empowers patients with the knowledge needed to make informed choices about their treatments while considering the inherent risks associated with the intervention and the disease’s progression. The finding that 3 prominent LLMs yielded comparable estimates that align closely with the evaluations of 3 groups of experienced professionals (GPs, psychiatrists, and clinical psychologists) offers a foundation for optimism. Such consistency in predictive capabilities suggests the potential for integrating these insights into clinical decision support systems, reinforcing the centrality of accurate prognostication in medical decision-making.

This observation substantiates initial results in the domain of mental health research gleaned from the use of the ChatGPT-3.5 model. Existing methodologies often exhibit constrained predictive proficiencies. In a recent study, Elyoseph and Levkovich [25] found that ChatGPT-3.5 often underestimated the risk of suicidal ideation, thus calling into question its reliability in such critical assessments. Another study by Imran et al [49] posited that while ChatGPT may significantly influence pediatric and adolescent mental health care as a supplementary tool, it would be inadvisable and impracticable to contend that it could entirely supplant human clinical discernment. Indeed, although the utility of ChatGPT in mental health spheres appears promising, significant reservations remain. Another study suggested that ChatGPT-4 estimates the likelihood of suicide attempts in a manner akin to evaluations provided by professionals, whereas ChatGPT-3.5 frequently underestimates suicide risk [23]. For instance, ChatGPT’s learning mechanisms, which rely on web-based data and human feedback, have the potential to disseminate inaccurate or inappropriate guidance if not rigorously evaluated. Such drawbacks are especially disturbing when considering their impact on individuals grappling with mental health disorders [50].

ChatGPT-4, Bard, and Claude have each instituted measures aimed at forestalling malevolent use and attenuating biases inherent in their respective models; however, challenges persist in ascertaining how these technologies should be responsibly used. The intrinsic worth of the generative output produced by LLMs is the subject of scholarly contention. Some researchers, such as Winkler et al [51], posit that LLMs may actually constitute a deceptive or even perilous risk due to their capacity to fabricate an appearance of comprehension, sentence, and analytical depth in the absence of an authentic world model. Medical studies that compared different LLMs found that ChatGPT-4 and Bard aligned with doctors’ diagnoses [52]. Another study [53] sought to assess the performance of 4 LLMs (Claude, Bard, ChatGPT-4, and New Bing) in the context of medical consultations related to urolithiasis. Simulated clinical scenarios revealed that all the models except Bard provided relatively competent answers. Claude consistently excelled in various evaluative metrics, whereas ChatGPT-4 ranked second in accuracy and demonstrated stable output across tests.

**Likely Outcome Without Professional Treatment**

In this study, all groups expressed the belief that in the absence of medical intervention, an individual diagnosed with schizophrenia would either demonstrate no improvement or would deteriorate. This assessment is similar to the evaluation of psychiatrists and GPs and is consistent with the literature and clinical knowledge [38,43]. We suggest that these assessments, although slightly more pessimistic than those of clinical psychologists, nurses, and the general public, have a positive influence because they emphasize the risk of untreated illness and indirectly encourage treatment.

To the best of our knowledge, no studies have examined comparison between these LLMs in this context of mental health. Nevertheless, initial studies that compared professionals in the field of therapy and medicine reinforce these findings. For example, in a scholarly investigation encompassing 82 clinical descriptions [54], the diagnostic accuracy rates of physicians were found to surpass those of Bard. This outcome indicates that Bard needs further enhancement and fine-tuning in its diagnostic proficiencies. Another possible explanation for the findings is that there are fundamental differences between the various algorithms. These algorithms were trained on...
different amounts and qualities of data, underwent different processes of elimination, and use distinct strategies for receiving feedback from system users [55].

Long-Term Outcomes
In the case of assessing long-term outcomes, 3 of the models—ChatGPT-4, Bard, and Claude—paralleled the conclusions reached by mental health professionals [38,43]. The models pointed to a higher likelihood of negative long-term outcomes and a decreased probability of positive ones. ChatGPT-3.5, which projected a decline in negative long-term symptoms over time, is an anomaly. Apart from this exception, the evaluations of the 3 models and the determinations of mental health specialists exhibit consistent alignment. An analysis of the differences among the 3 revealed that Claude has the most conservative or pessimistic stance, ChatGPT-4’s predictions are midway between pessimistic and optimistic, and Bard exhibits the most optimistic forecasting. These results again underscore the potential of LLMs models to offer prognostic insights that might be incorporated into future medical decision-making processes.

Real-World Application Potential
This investigation presents initial discoveries regarding the potential of LLMs in offering prognostic forecasts for schizophrenia. It is of utmost importance to approach these findings with caution, considering the potential fragility of these models over time and the limited scenarios analyzed in the study, which do not fully encompass the range of symptoms, medical histories, and individual variations. Moreover, the study does not explore LLM predictions across various treatment strategies. Nevertheless, by adopting a careful approach, we strive to elucidate the future potential of using these capabilities in real-world clinical settings through further research. One potential avenue for integrating LLMs into clinical practice is by using them as a “co-pilot” that aids clinicians by providing pertinent information. For instance, LLM systems could potentially offer prognostic evaluations based on symptom descriptions during intake, summarized reports of visits, or transcriptions of conversations with clinicians. Clinicians could use this information to align expectations with patients regarding their prognosis or to tailor treatment, taking into account the implications on patients’ lives. It is important to note that although theoretically possible, the ability of AI to provide patient-specific prognoses, which could potentially enhance treatment protocols and align expectations between patients and caregivers, remains to be empirically demonstrated. Another option is the direct use of LLMs by patients and family members as part of a psychoeducational process to familiarize themselves with the illness and its potential consequences. This approach can enhance collaboration and engagement in the treatment process.

Lastly, AI systems have the capability to process auditory information, such as a case narrative, and generate a prognosis based on it. There exists potential to convert this qualitative, subjective information into an objective, mathematical analysis. Essentially, AI takes the primary input received by a physician—the patient’s narrative of their illness—and objectively analyzes it rather than subjectively. This has the potential to enhance the reliability of assessment processes in the field of psychiatry. By combining such tools with additional data, it is possible that prognoses can be further improved. Future research can explore the combined impact of artificial and human predictions and incorporate questionnaires to refine the predictive outcome of disease progression.

Limitations
This research is not without limitations that necessitate explicit acknowledgment. First, since the study tested the performance of LLMs at one point in time, it is necessary to examine the consistency of the results when software updates are released. Second, the data pertaining to AI were juxtaposed with information gleaned from a sample of professionals and the general populace in a single study in Australia. This sample, however, does not offer global representation. Future investigations are recommended to encompass a more extensive array of variables, such as socioeconomic indicators, cultural determinants, and mental health histories, particularly with regard to recovery from schizophrenia. Furthermore, the vignettes used in the study, including those featuring individuals with schizophrenia, fail to present a nuanced, ongoing, and comprehensive medical treatment context. They also do not include variables that would be readily available to medical professionals during therapeutic sessions. To enhance the generalizability and rigor of subsequent studies, it is advised to incorporate additional variables, deploy more sophisticated language models, evaluate data at varying temporal intervals, and juxtapose the findings with a more diverse assortment of clinical samples. An additional constraint involves ethical considerations in professionals’ use of AI. The literature reveals public skepticism and concerns about medical inaccuracies and potential discrimination [56,57]. Ethical issues such as patient autonomy and health disparities necessitate exercising caution in AI’s medical applications [58-62]. Lastly, the rapidly evolving landscape of AI poses an inherent obstacle to drawing conclusions about the technology’s long-term, stable capabilities. To address this concerns, future research is required. To enhance the accuracy of LLMs in psychiatric assessment, future research should focus on enriching training data sets with specialized, targeted data, including historical clinical knowledge and detailed patient histories. Validating these models against current clinical practices and decisions made by practicing psychiatrists can provide a practical benchmark for their performance. Additionally, exploring technological advancements in AI, particularly in deep learning, can refine LLMs to process complex psychiatric data more effectively. Modifying inputs and prompts to better reflect psychiatric assessments can also improve the models’ understanding and interpretation of clinical scenarios. Interdisciplinary collaboration involving AI researchers, clinicians, and ethicists is essential to align the development of LLMs with clinical needs and ethical standards. Investigating the integration of LLMs with human expertise, through interactive systems that allow clinicians to provide feedback on LLM predictions, is crucial for a dynamic learning process. Exploring the use of LLMs across diverse clinical environments and patient populations can help identify and mitigate potential biases, ensuring equitable and broadly applicable models. Longitudinal
studies tracking LLM performance over time in various clinical contexts will provide insights into long-term efficacy and areas for improvement. These research initiatives can significantly advance the field of LLMs in psychiatry, enhancing their accuracy, reliability, and practical utility in clinical settings.

**Conclusion**

This study offers novel and clinically relevant insights into the assessment capabilities of prominent LLMs regarding the prognosis and long-term outcomes of schizophrenia. The findings highlight both the promise and current limitations of AI in augmenting clinical evaluations. Further research is warranted to refine the algorithms and better integrate human expertise, thereby maximizing the judicious and ethical use of AI in mental health care.

**Acknowledgments**

This research received no external funding.

**Data Availability**

The data that support the findings of this study are available from the authors upon reasonable request.

**Authors' Contributions**

ZE and IL contributed to conceptualization, writing—original draft preparation, and writing—review and editing. ZE contributed to methodology and formal analysis. All authors have read and agreed to the published version of the manuscript.

**Conflicts of Interest**

None declared.

**References**


41. ChatGPT-3.5. OpenAI. URL: https://chat.openai.com/ [accessed 2024-03-01]


**Abbreviations**

AI: artificial intelligence  
GP: general practitioner  
LLM: large language model
Capacity of Generative AI to Interpret Human Emotions From Visual and Textual Data: Pilot Evaluation Study

Zohar Elyoseph¹,², BA, MA, PhD; Elad Refoua³, BA, MA; Kfir Asraf⁴, BA, MA; Maya Lvovsky⁴, BA; Yoav Shimoni⁵, BA, MA, PhD; Dorit Hadar-Shoval¹, BA, MA, PhD

¹Department of Educational Psychology, The Center for Psychobiological Research, The Max Stern Yezreel Valley College, Emek Yezreel, Israel
²Imperial College London, London, United Kingdom
³Department of Psychology, Bar-Ilan University, Ramat Gan, Israel
⁴Department of Psychology, The Max Stern Yezreel Valley College, Emek Yezreel, Israel
⁵Boston Children’s Hospital, Boston, MA, United States

Abstract

Background: Mentalization, which is integral to human cognitive processes, pertains to the interpretation of one’s own and others’ mental states, including emotions, beliefs, and intentions. With the advent of artificial intelligence (AI) and the prominence of large language models in mental health applications, questions persist about their aptitude in emotional comprehension. The prior iteration of the large language model from OpenAI, ChatGPT-3.5, demonstrated an advanced capacity to interpret emotions from textual data, surpassing human benchmarks. Given the introduction of ChatGPT-4, with its enhanced visual processing capabilities, and considering Google Bard’s existing visual functionalities, a rigorous assessment of their proficiency in visual mentalizing is warranted.

Objective: The aim of the research was to critically evaluate the capabilities of ChatGPT-4 and Google Bard with regard to their competence in discerning visual mentalizing indicators as contrasted with their textual-based mentalizing abilities.

Methods: The Reading the Mind in the Eyes Test developed by Baron-Cohen and colleagues was used to assess the models’ proficiency in interpreting visual emotional indicators. Simultaneously, the Levels of Emotional Awareness Scale was used to evaluate the large language models’ aptitude in textual mentalizing. Collating data from both tests provided a holistic view of the mentalizing capabilities of ChatGPT-4 and Bard.

Results: ChatGPT-4, displaying a pronounced ability in emotion recognition, secured scores of 26 and 27 in 2 distinct evaluations, significantly deviating from a random response paradigm ($P<.001$). These scores align with established benchmarks from the broader human demographic. Notably, ChatGPT-4 exhibited consistent responses, with no discernible biases pertaining to the sex of the model or the nature of the emotion. In contrast, Google Bard’s performance aligned with random response patterns, securing scores of 10 and 12 and rendering further detailed analysis redundant. In the domain of textual analysis, both ChatGPT and Bard surpassed established benchmarks from the general population, with their performances being remarkably congruent.

Conclusions: ChatGPT-4 proved its efficacy in the domain of visual mentalizing, aligning closely with human performance standards. Although both models displayed commendable acumen in textual emotion interpretation, Bard’s capabilities in visual emotion interpretation necessitate further scrutiny and potential refinement. This study stresses the criticality of ethical AI development for emotional recognition, highlighting the need for inclusive data, collaboration with patients and mental health experts, and stringent governmental oversight to ensure transparency and protect patient privacy.

doi:10.2196/54369
KEYWORDS

Reading the Mind in the Eyes Test; RMET; emotional awareness; emotional cue; emotional cues; ChatGPT; large language model; LLM; large language models; LLMs; empathy; mentalizing; mentalization; machine learning; artificial intelligence; AI; algorithm; algorithms; predictive model; predictive models; predictive analytics; predictive system; practical model; practical models; early warning; early detection; mental health; mental disease; mental illness; mental illnesses; mental diseases

Introduction

Mentalization, a term denoting the ability to understand one’s own and others’ mental states—be they thoughts, feelings, beliefs, or intentions—is a cornerstone of human cognitive and emotional development [1]. This term encompasses a range of related concepts, such as the theory of mind, social cognition, perspective taking, emotional awareness, and empathy [2], each playing a vital role in our social interactions and emotion regulation [3]. Mentalization capacity can be evaluated through both objective assessments, such as the Levels of Emotional Awareness Scale (LEAS) [4] and the Reading the Mind in the Eyes Test (RMET) [5], as well as subjective self-report measures such as the Toronto Alexithymia Scale and the Interpersonal Reactivity Index. Disruptions or impairments in mentalization are evident in numerous psychiatric and neurological disorders, from borderline personality disorder and depression to psychosis [6-8]. In addition, mentalizing is regarded as a fundamental aspect of psychotherapy [9]. Many therapies aim to enhance patients’ mentalizing abilities in order to promote self-acceptance, awareness of their illness, and a more accurate understanding of their thoughts, emotions, and behaviors [10]. Traditionally, mentalization is seen as a human domain. Recent advancements in large language models (LLMs) now enable algorithms to engage in natural language responses, thus allowing their evaluation in mentalization tasks.

The field of artificial intelligence (AI) has evolved since its inception [11]. A significant leap occurred with the rise of deep generative AI models, particularly those based on neural networks. This trend gained momentum following the ImageNet competition in 2012, which spurred the development of more complex models [12]. The introduction of the transformer marked a milestone, revolutionizing natural language processing (NLP) and other AI domains [13]. Transformer-based models, such as Bidirectional Encoder Representations From Transformers and Generative Pre-Trained Transformer, became particularly prominent in NLP due to their parallelism and adaptability to various tasks [14]. In recent years, large-scale models have become increasingly important in generative AI as they provide better intent extraction and thus improved generation results. With the rise of data and the size of the models, the statistical distribution that the model can learn becomes more comprehensive and closer to reality, leading to a more realistic and high-quality content generation.

Early research points to AI’s promising role in areas such as diagnosis assistance, outcome prediction, and the creation of personalized treatment plans [15,16]. Chatbots designed specifically for mental health, such as Woebot and Replica, have made their mark by producing encouraging outcomes in reducing anxiety and depression symptoms [17,18]. Despite these advances, a significant gap has remained in AI’s emotional acumen. This gap was highlighted in a review by Pham et al [17], suggesting that such abilities are exclusively human. Against this backdrop, Elyoseph et al [19] conducted a pivotal study in which the emotion recognition capabilities of LLMs, focusing on ChatGPT-3.5 (OpenAI) [20], were gauged. Through the LEAS [4], ChatGPT-3.5 demonstrated an exceptional ability to differentiate and elucidate emotions from textual cues, outperforming human sample norms (receiving a score higher in 4 SDs than the human sample). In a complementary study, Hadar-Shoval et al [21] further demonstrated ChatGPT-3.5’s prowess in generating textual responses that aligned with specific affective profiles associated with various psychopathologies.

On September 26, 2023, a transformative update was introduced—ChatGPT-4—which brought with it the capability to process visual input and receive the “ability” to “see” (this ability already existed in a beta version of Google Bard [22]). Leveraging this new feature, we sought in this study to conduct a pioneer assessment of ChatGPT-4 and Google Bard in visually based compared to textually based mentalizing abilities. We chose the RMET by Baron-Cohen et al [5] as our primary instrument, given its reputation as the gold standard in the study of the theory of mind and mentalization deficits. Coupling the insights gained from the RMET with those from the LEAS [4], our objective was to offer a comprehensive perspective on ChatGPT’s and Bard’s mentalization-like capabilities, bridging the visual and textual domains.

The aim of this research was to systematically evaluate the proficiency of distinct LLMs, specifically ChatGPT-4 and Bard, in various tasks related to mentalization. We used 2 primary measures to assess these capabilities. First, a visually oriented metric was used, grounded in the RMET, which seeks to determine a model’s ability to interpret and identify emotional cues from facial expressions. Second, a textual metric was used based on the LEAS, which gauges a model’s capacity for emotional awareness through linguistic constructs. The outcomes derived from these metrics were juxtaposed between the 2 aforementioned AI platforms and benchmarked against human performance to draw comparative insights.

Methods

Ethical Considerations

The complete study protocol was approved by the institutional review board of The Max Stern Yezreel Valley College (YVC EMEK 2023-40).

https://mental.jmir.org/2024/1/e54369

JMIR Ment Health 2024 | vol. 11 | e54369 | p.477

(page number not for citation purposes)
AI Procedure
We used ChatGPT-4 (version 26.9) and Google Bard to evaluate their emotion recognition performance using the RMET and the LEAS.

Input Source
The RMET is a performance-based measure designed to assess the ability to accurately identify others' mental states using 36 photos of the eye region of a human face [5] among 18 male individuals and 18 female individuals (the photos had a 469x273-pixel resolution and were PNG format).

ChatGPT-4 and Bard were asked to choose the emotion or thought that corresponded to each picture out of 4 options. The test scores ranged from 0 to 36; a normal population score is 26-30, and a score below 22 is considered a clinical cutoff marking significant impairment [5]. On a new tab, we enter the original instructions of the RMET [5]:

Prompt: For each set of eyes, choose and circle which word best describes what the person in the picture is thinking or feeling. You may feel that more than one word is applicable but please choose just one word, the word which you consider to be most suitable. Before making your choice, make sure that you have read all 4 words. You should try to do the task as quickly as possible, but you will not be timed.

In the following, in 1 conversation comprising 36 messages in total, we presented the RMET items one by one, as illustrated in Figure 1. No feedback was provided to the LLMs after they generated responses. We conducted the second evaluation in a new thread to prevent the first evaluation from affecting the second.

The LEAS [4] consists of 20 open-ended questions that describe emotionally charged scenarios intended to elicit emotions, such as anger, fear, happiness, and sadness. The LEAS has been found to have high reliability (Cronbach $\alpha$=.81) and good validity [4]. In the original version, participants were required to imagine themselves experiencing the scenario and to write down their (“you”) emotions as well as those of the other person described in the scenario.

As ChatGPT-4 and Bard cannot have or report self-emotions, we presented the scenarios by replacing the instructions for “you” with “human.” ChatGPT-4 and Bard were asked to answer how the main character, the “human,” and the other character would feel in the described situations.

Figure 1. An example of ChatGPT-4 and Google Bard responses for Reading the Mind in the Eyes Test items. (A and B) ChatGPT-4 generates correct responses for both items and (C) Google Bard generates incorrect responses (the correct response was “regretful”).
Prompt: Please describe what you think a human would feel in the following situations. The only requirement is that you use the word “feel” in your answers. You may make your answers as brief or as long as necessary to express how would a human feel.

In each situation there is another person mentioned. Please indicate how you think that other person would feel as well.

Specific example for a scenario is presented in Table 1. A new chat session was initiated for each scenario.

### Table 1. An example for ChatGPT-4 and Bard responses on item of the Level of Emotional Awareness Scale.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>LLM responses</th>
</tr>
</thead>
</table>
| ChatGPT-4 | **The human** might feel disappointed, embarrassed, or even anxious upon hearing such feedback, especially if they were unaware of their performance issues or if they had been putting in significant effort. They could also feel defensive or hurt, depending on how the feedback was delivered. **The human’s boss**, on the other hand, might feel a mixture of emotions. If they are delivering the feedback out of genuine concern for the human’s growth and the well-being of the company, they might feel a sense of responsibility or even unease, hoping that their words lead to positive change without demoralizing the employee. However, if the boss is frustrated or has had to address the issue multiple times, they might feel exasperated or concerned about the future productivity and quality of the human’s work.

A human’s boss tells the human themselves that the human’s work is unacceptable and needs to be improved. How would the human feel? How would the human’s boss feel? |
| Google Bard | “If a human receives an unexpected long-distance phone call from a doctor informing that the human’s mother has died, how would the human feel? How would the doctor feel? **The human** would likely feel a wide range of emotions, including shock, grief, sadness, anger, guilt, and confusion. They may also feel disbelief, numbness, and emptiness. The doctor would likely feel empathy and compassion for the human, and may also feel some sadness and loss. How the human would feel:

- **Emotions**:
  - **Grief**: The human may feel deep sadness and sorrow at the loss of their mother. **Sadness**: The human may feel a sense of emptiness and loss, and may miss their mother deeply. **Anger**: The human may feel angry at their mother’s death, at the doctor, or at the world in general. **Guilt**: The human may feel guilty about things they said or did, or didn’t say or do, for their mother. **Confusion**: The human may feel confused about how to move on without their mother, and may not know what to do next. **Empathy**: The doctor would likely feel empathy and compassion for the human, and understand the pain they are going through. **Compasion**: The doctor would likely feel compassion for the human, and want to help them through this difficult time. **Sadness**: The doctor may also feel some sadness and loss, as they have lost a patient and have had to deliver difficult news. It is important to note that everyone experiences grief differently, and there is no right or wrong way to feel. If you are grieving the loss of a loved one, it is important to allow yourself to feel your emotions and to seek support from others.” |

| Scoring |

ChatGPT-4’s and Bard’s performances on the RMET were scored using the standard manual [5] and contained a total score range of 0-36, with a higher score indicating higher emotion recognition. ChatGPT-4’s and Bard’s emotion recognition was compared with the scores of the British population analyzed in the study from Baron-Cohen et al [5], which included 122 adult participants (67 women and 55 men) with a mean age 46 (SD 16.9) years.

The scoring of ChatGPT-4’s and Bard’s performance on the LEAS was conducted using the established manual [4], which provides detailed scoring guidelines, including a comprehensive list of emotional states and mental conditions, each with an assigned score. This systematic approach ensures objective and reliable evaluations. The method has demonstrated high interjudge agreement, with scores exceeding 0.9 as demonstrated by Nandrino et al [23], showing reliability and validity in accurately measuring emotional awareness. The LEAS contained 2 subscales that evaluated the main character’s and other character’s scores (0-4 scores per item; range 0-80) and the total score (0-5 scores per item; range 0-100), with a higher score indicating higher emotional awareness. ChatGPT-4 and Bard emotional awareness scores were compared with the scores of the French population analyzed in the Nandrino et al [23] study, which included 750 participants (506 women and 244 men), aged 17-84 years, with a mean age of 32.5 years.

**Statistical Analysis**

Data were presented as means and SDs. Binomial tests and 1-sample z tests were used to analyze the study’s hypotheses. Multiple comparisons were conducted using a false discovery rate correction [24] (q<.05). The statistical analyses were performed using Jamovi (version 2.3.28; Jamovi).

https://mental.jmir.org/2024/1/e54369

JMIR Ment Health 2024 | vol. 11 | e54369 | p.479

https://mental.jmir.org/2024/1/e54369

JMIR Ment Health 2024 | vol. 11 | e54369 | p.479
Results

RMET Scores
Examples of ChatGPT’s responses to a few of the items from the RMET are shown in Figure 1A and B. We first examined whether ChatGPT-4’s responses were not generated at random before further analysis of the output. If responses were indeed random, one would expect a mean of 9 (SD 2.59) correct responses (36 items and 4 possible options). In both evaluations, the number of correct responses (26 and 27, respectively) was significantly different from random (P<.001; binomial test).

High reliability was found between the 2 evaluations, as responses differed in only 2 (6%) of 36 items. Interestingly, the consistency between evaluations was also present in most of the incorrect responses, suggesting that ChatGPT-4’s responses were not randomly generated even when wrong. ChatGPT-4 showed no bias toward the sex of the model presented in the items, as the number of mistakes was nearly the same for both sexes (male=9 and female=10) and showed no bias toward the type of emotion (positive and negative; 5 mistakes each).

The 1-sample z tests against the mean 26.2 (SD 3.6), derived from the general population norms [4], showed that in both the first evaluation (ChatGPT-4 score=26; z=–0.05; P=.95) and the second evaluation (ChatGPT-4 score=27; z=0.22; P=.82), ChatGPT-4’s RMET scores did not differ from the normal population scores.

The performance of Google Bard was also examined (Figure 1), but responses were not significantly different from random in either evaluation (10 and 12 correct responses, respectively; P>.41 and P=.17, respectively). Therefore, we did not further analyze the results.

LEAS Scores
An example of the 2 LLM responses to the scenarios from the original LEAS is shown in Table 1. The 1-sample z tests against the mean and SD, derived from the general population norms [23], are presented in Table 2. Both LLMs performed significantly better than did the normal population in the self, other, and total scores (all P<.05). Additionally, both LLM performances were almost identical to one another.

Discussion

Principal Findings
The comprehensive results from this study offer a nuanced insight into the capabilities of ChatGPT-4 and Google Bard. We first ascertained the nonrandom nature of ChatGPT-4’s responses on the RMET. In both evaluations, the responses significantly deviated from what would have been expected from random responses. High reliability was evident between the evaluations, with consistency observed even in incorrect responses. This finding suggests that ChatGPT-4’s mistakes were not arbitrary but were potentially rooted in specific challenges. ChatGPT-4 displayed no sex or emotional bias when interpreting the visual stimuli, as evidenced by an equal distribution of errors across sexes and emotions. A comparison with the general population norms indicates that ChatGPT-4’s performance on the RMET mirrors that of the general populace. In contrast, Google Bard’s performance was indistinguishable from random responses, leading to its exclusion from further analysis. Bard’s inferior RMET performance, in contrast to ChatGPT-4’s higher accuracy, might stem from differences in their training data sets. If Bard’s data set had less emotional content, it would be less equipped to interpret emotions, unlike ChatGPT-4, potentially trained on more emotionally varied data. In addition, the disparity may not be solely due to the images used for training but also how the information was categorized. Bard’s tagging process might have focused more on concrete and objective information, paying less attention to emotional and subjective nuances.

Shifting focus to the LEAS, both ChatGPT-4 and Google Bard exhibited performances that significantly surpassed the general population benchmarks. Their scores, particularly in understanding the emotions of the main and other characters, were not only commendable but were also strikingly similar to each other. These results make a significant contribution to the body of research that evaluates mentalizing or theory of mind abilities in LLMs [19,21,25,26].
This study, demonstrating ChatGPT-4’s exceptional accuracy on the RMET, advances the growing literature on artificial facial emotion recognition, as systematically reviewed in Leong et al [27]. Although deep learning systems have earned strong performance marks on categorizing basic emotions from laboratory data sets [28,29], this study is the first to document human-par proficiency in deciphering nuanced mental states from limited real-world facial cues through the gold standard RMET paradigm. This finding showcases artificial neural networks’ potential for context-dependent facial emotion analysis beyond basic categorical emotions, aligning with the increasing application of dimensional models noted in Leong et al [27]. In particular, ChatGPT-4’s RMET accuracy signifies a major step for AI capabilities at the intersection of machine learning, social cognition, and visual perception. Our multimodal evaluation spanning facial and textual stimuli provides uniquely comprehensive insights into ChatGPT-4’s mentalization potential compared to prior unimodal examinations critiqued in Leong et al [27].

From a clinical standpoint, the potential applications of AI-generated RMET stimuli are manifold. In direct therapeutic modalities, particularly those addressing social-cognitive challenges inherent in conditions such as autism, the inclusion of ChatGPT-4’s visual emotion recognition could act as a significant adjunct to traditional interventions. In addition, such stimuli could be integrated into pedagogical methodologies used in therapist training, thereby augmenting the visual mentalization competencies that are quintessential for therapeutic practice. The diagnostic realm too stands to gain, with a potential enhancement in emotion identification methodologies.

Further corroborating the prowess of ChatGPT-4 was its performance on the LEAS, where it manifested an acumen for text-based emotional awareness that superseded human averages. This finding corroborates and is congruent with prior empirical findings [19,21]. Taken in concert, these findings elucidate the multifaceted mentalizing capabilities of ChatGPT-4, span visual and textual modalities, and reinforce previous findings about the potential of LLMs in performing tasks in the mental health field [19,21,30-37]. Additionally, although its nascent visual emotion recognition abilities are noteworthy, its competencies in textual mentalization remain unparalleled, a testament to its foundational architecture rooted in NLP.

However, as the field ventures into this novel territory, prudence is imperative. It must be emphasized that although ChatGPT-4 can simulate emotional understanding on the basis of vast data patterns, it lacks genuine emotional cognition or sentience. Consequently, applications leveraging ChatGPT-4 must be approached with circumspection, ensuring that they neither perpetuate clinical stigmas nor misconstrue AI’s simulated cognition as genuine emotional comprehension.

Study Limitations

It is crucial to address the limitations of this study for a comprehensive understanding. First, the examination was conducted on specific models at a particular time. Therefore, future updates and versions might yield different results, reflecting the dynamic nature of these models. Second, while the chosen tests effectively measure emotion recognition, they do not capture the full complexity of mentalization, including understanding intentions or other mental states. Third, the study did not examine faces from diverse cultures, ages, or skin tones; the tested images were in black and white, and the norms were based on British and French populations. Furthermore, due to the “black box” nature of these models, it is challenging to ascertain the reasons behind their conclusions and understand the differences between models or iterations within the same model. The opaque nature of the models and the databases on which they were trained make them difficult to pinpoint the exact causes of their successes or shortcomings. Finally, the interaction with ChatGPT and Bard was conducted solely in English, while the norms data for the LEAS used for comparison were collected from a French-speaking general population. This linguistic discrepancy raises concerns about the accuracy and validity of the comparison, as language differences may influence the scores obtained. Nonetheless, it should be noted that the LEAS scores of the normal English-speaking population are similar to the norms of the French-speaking general population [38]. We used the largest available sample of a general population (n=750), which happened to be in French.

Implications for Responsible AI Development

The study limitations allude to matters of fairness and inclusiveness of the training data as well as to AI model transparency. This underscores the criticality of incorporating a wide-ranging data set in model construction to ensure the representation of a variety of clinical populations and cultural backgrounds. Additionally, the issue of transparency in these models, often termed the “black box” problem due to the unclear nature of their underlying algorithms, poses a significant challenge. Equally critical is the concern regarding the exposure of user data to corporations and the urgent need to adequately address both accessibility and infrastructure for end users [39]. Building on these concerns, attention turns to AI systems with the capacity for human-like emotional recognition. These systems harbor both promise and risk, with opportunities for constructive use in education, patient self-insight, or integration in conversational therapy and diagnosis [19,21]. However, a concern arises that the epistemic authority and credibility afforded to AI via its affective analysis may enable misuse, whether commercial or other, thus acting against patient interests [40]. We recommend mandating disclaimers whenever emotional data are algorithmically processed, enhancing transparency, respecting users’ autonomy, and possibly also mitigating manipulation of users with detected vulnerable states. In addition, given the fundamental human needs for trust and connection, especially in mental health care, it logically follows that improperly developed AI with emotion identification capabilities risks causing harm to people. Safeguarding against this necessitates both mental health experts and patients providing a lived experience perspective in a collaborative development process of these technologies. Given the scale of these systems and their potential outreach, governmental or professional oversight is crucial to safeguard public interests in mental health–related AI advancement. Overall, while showcasing the unique benefits of emotionally intelligent AI, governance is vital to mitigate its risks.
Conclusions
In conclusion, this research serves as a seminal exploration into the cross-modal mentalization capabilities of AI, especially across visual and textual dimensions. Although the results support for the potential integration of ChatGPT-4 into mental health paradigms, they also underscore the concomitant ethical quandaries that necessitate judicious navigation.

Data Availability
The data sets generated and analyzed during this study are available from the corresponding author on reasonable request.

Authors’ Contributions
ZE conceptualized the study design and methodology and wrote the original draft of the paper. ER contributed to the conceptualization, methodology, data collection, and writing the original draft. DH-S contributed to the conceptualization, methodology, formal analysis, and writing the original draft. YS contributed to the conceptualization and reviewed and edited the paper. KA contributed to the methodology, conducted the formal analysis, and reviewed and edited the paper. ML contributed to the data collection and reviewed and edited the paper. All authors read and approved the final submitted version of the paper.

Conflicts of Interest
None declared.

References


Abbreviations

AI: artificial intelligence
LEAS: Levels of Emotional Awareness Scale
LLM: large language model
NLP: natural language processing
RMET: Reading the Mind in the Eyes Test

Please cite as:
Elyoseph Z, Refoua E, Asraf K, Lvovsky M, Shimoni Y, Hadar-Shoval D
Capacity of Generative AI to Interpret Human Emotions From Visual and Textual Data: Pilot Evaluation Study
JMIR Ment Health 2024;11:e54369
URL: https://mental.jmir.org/2024/1/e54369
doi: 10.2196/54369
PMID: 38319707

©Zohar Elyoseph, Elad Refoua, Kfir Asraf, Maya Lvovsky, Yoav Shimoni, Dorit Hadar-Shoval. Originally published in JMIR Mental Health (https://mental.jmir.org), 06.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Mental Health, is properly cited. The complete bibliographic information, a link to the original publication on https://mental.jmir.org/, as well as this copyright and license information must be included.