JMIR Mental Health

Impact Factor (2021): 6.33
Volume 9 (2022), Issue 9 ISSN: 2368-7959 Editor in Chief: John Torous, MD

Contents

Reviews

Original Papers

Review

Effectiveness and Minimum Effective Dose of App-Based Mobile Health Interventions for Anxiety and Depression Symptom Reduction: Systematic Review and Meta-Analysis

Sheng-Chieh Lu[1], PhD; Mindy Xu[2], BS; Mei Wang[3], MS; Angela Hardi[4], MLIS; Abby L Cheng[3,5], MD; Su-Hsin Chang[3], PhD; Po-Yin Yen[6,7], RN, PhD

[1]Department of Symptom Research, University of Texas MD Anderson Cancer Center, Houston, TX, United States

[2]Keck School of Medicine, University of Southern California, Los Angeles, CA, United States

[3]Division of Public Health Sciences, Department of Surgery, Washington University in St Louis, St Louis, MO, United States

[4]Becker Medical Library, Washington University in St Louis, St Louis, MO, United States

[5]Division of Physical Medicine and Rehabilitation, Department of Orthopaedic Surgery, Washington University in St Louis, St Louis, MO, United States

[6]Institute for Informatics, Washington University in St Louis, St Louis, MO, United States

[7]Goldfarb School of Nursing, Barnes Jewish College, BJC HealthCare, St Louis, MO, United States

Corresponding Author:
Sheng-Chieh Lu, PhD
Department of Symptom Research
University of Texas MD Anderson Cancer Center
6565 MD Anderson Blvd
Houston, TX, 77030
United States
Phone: 1 7137944453
Fax: 1 7137453475
Email: Slu4@mdanderson.org

Abstract

Background: Mobile health (mHealth) apps offer new opportunities to deliver psychological treatments for mental illness in an accessible, private format. The results of several previous systematic reviews support the use of app-based mHealth interventions for anxiety and depression symptom management. However, it remains unclear how much or how long the minimum treatment "dose" is for an mHealth intervention to be effective. Just-in-time adaptive intervention (JITAI) has been introduced in the mHealth domain to facilitate behavior changes and is positioned to guide the design of mHealth interventions with enhanced adherence and effectiveness.

Objective: Inspired by the JITAI framework, we conducted a systematic review and meta-analysis to evaluate the dose effectiveness of app-based mHealth interventions for anxiety and depression symptom reduction.

Methods: We conducted a literature search on 7 databases (ie, Ovid MEDLINE, Embase, PsycInfo, Scopus, Cochrane Library (eg, CENTRAL), ScienceDirect, and ClinicalTrials, for publications from January 2012 to April 2020. We included randomized controlled trials (RCTs) evaluating app-based mHealth interventions for anxiety and depression. The study selection and data extraction process followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines. We estimated the pooled effect size using Hedge g and appraised study quality using the revised Cochrane risk-of-bias tool for RCTs.

Results: We included 15 studies involving 2627 participants for 18 app-based mHealth interventions. Participants in the intervention groups showed a significant effect on anxiety (Hedge g=–.10, 95% CI –0.14 to –0.06, I2=0%) but not on depression (Hedge g=–.08, 95% CI –0.23 to 0.07, I2=4%). Interventions of at least 7 weeks' duration had larger effect sizes on anxiety symptom reduction.

Conclusions: There is inconclusive evidence for clinical use of app-based mHealth interventions for anxiety and depression at the current stage due to the small to nonsignificant effects of the interventions and study quality concerns. The recommended dose of mHealth interventions and the sustainability of intervention effectiveness remain unclear and require further investigation.

(JMIR Ment Health 2022;9(9):e39454) doi:10.2196/39454

KEYWORDS

mental health; mobile health; smartphone apps; intervention dose effectiveness; systematic review and meta-analysis

Introduction

More than 250 million people worldwide have depression or anxiety, which are the 2 most common mental illnesses that contribute to the global burden of disease [1]. The recent coronavirus disease pandemic has further increased the numbers of people reporting symptoms of anxiety and depression [2]. Both psychological and pharmacological therapies have been reported to effectively reduce the symptoms of mental illness. Yet, depression and anxiety disorders are notably undertreated due to a variety of barriers, such as lack of access to treatments and reluctance to get treatments because of social stigma and unawareness of symptoms [3]. The ongoing pandemic resulting in restrictions on social and physical distancing has posed additional challenges to these individuals, worsening undertreatment [2].

Mobile health (mHealth) apps leverage the ubiquity of mobile devices and the mobile-cellular telecommunication infrastructure and offer new opportunities to deliver psychological treatments for mental illness in an accessible, private format [4]. As the affordability and accessibility of smartphones are increasing, mobile apps are becoming the main component of many interventions promoting mental wellness and thus could be an exceptional tool to support mental health care delivery [5,6]. Research effort has been made to develop and examine mobile app-based interventions to improve patient engagement in symptom management and reduce mental illness symptoms. For instance, several smartphone apps are available for delivering self-directed cognitive behavioral therapy (CBT) for those with depression [7]. Other psychotherapies that are feasible to be facilitated by apps include acceptance and commitment therapy (ACT), problem-solving therapy (PST), and psychoeducation [8,9].

Several previous systematic reviews and meta-analyses have supported the use of app-based mHealth interventions for anxiety and depression symptom management. Firth et al [10,11] have reported a small-to-moderate effect size for both anxiety and depression symptom reduction following interventions delivered fully or partially by smartphone compared to control groups (anxiety: Hedge g=.33, 95% CI 0.17-0.48, P<.01; depression: Hedge g=.38, 95% CI 0.24-0.52, P<.001). Another recent systematic review reported similar results supporting the use of stand-alone smartphone apps for depression (Hedge g=.34, 95% CI 0.18-0.49, P<.001) and anxiety (Hedge g=.43, 95% CI 0.19-0.66, P≤.001) symptom reduction [12] Nevertheless, although previous studies have examined intervention features and components to identify the most effective design for app-based mHealth interventions [10,12], due to the various study lengths (ie, 4 weeks, 6 weeks, 3 months, 6 months), it remains unclear how much or how long the minimum treatment "dose" is for an mHealth intervention to be effective.

Just-in-time adaptive intervention (JITAI) has been introduced in the mHealth domain to facilitate behavior changes; it proposes the use of ongoing information (individuals' changing status) to adapt the delivery of the intervention in its type, timing, or amount (intensity) [13]. The goal of JITAI is to increase an individual's acceptance of the intervention as the intervention is delivered "at the moment and in the context that the person needs it most and is most likely to be receptive" [14]. Smartphones are an ideal platform to deliver JITAIs because individuals' responses and their location can reveal whether the intervention is delivered and received at its maximum capacity. JITAI has been used to support health behaviors changes, such as physical activity [15,16], healthy diet [17,18], weight loss [19], and addiction [20-22]. A recent meta-analysis of 31 JITAI studies found significant effects of JITAI on improving health outcomes and enhancing study retention and intervention adherence [13,23,24]. JITAI emphasizes intervention tailoring to meet individual needs to achieve the best outcomes; thus, JITAI strategies regarding intervention dose (ie, type, amount, and timing of delivery) are positioned to guide the design of mHealth interventions with enhanced adherence and effectiveness.

In this study, our primary goal was to evaluate and update the evidence of app-based mHealth interventions for anxiety and depression symptom reduction through a systematic review and meta-analysis. In addition, inspired by the JITAI framework, we examined the effective mHealth dose for anxiety and depression symptoms where information was available. In other words, what is the minimum amount of usage or exposure to an mHealth app to effectively reduce anxiety and depression symptoms? To the best of our knowledge, this is the first systematic review and meta-analysis to examine the effectiveness of mHealth in anxiety and depression from a dose perspective.

Methods

Design

We conducted this systematic review and meta-analysis and reported the results following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines [25].

Search Strategy

We searched the published literature using keywords and strategies designed by the team with the assistance of a medical librarian. These strategies were created using a combination of controlled vocabulary terms and plain keywords (Multimedia Appendix 1). Databases that were searched were Ovid MEDLINE, Embase, PsycInfo, Scopus, Cochrane Library (eg, CENTRAL), ScienceDirect, and ClinicalTrials. We limited the search to studies published from January 2012 to April 2020. All searches were completed on April 30, 2020.

Study Selection

Studies were included if they (1) evaluated an app-based mHealth intervention designed to treat anxiety or depression or both, (2) measured symptoms of anxiety or depression, (3) were published as original research/trials in peer-reviewed journals, and (4) were written in English. We included studies that examined interventions delivered in part via mobile apps (ie, smartphone + regular phone call). We excluded studies if they (1) evaluated interventions not delivered in real-world settings (eg, only delivered within a laboratory or clinical setting), (2) evaluated interventions not delivered through a hand-held/mobile device, (3) only measured intervention usability or adherence but not the intervention effect on anxiety or depression symptoms or outcomes, (4) only measured physical stress responses but not any psychological anxiety-related symptoms, (5) did not include a control group and an experimental/comparison group with a random allocation process, or (6) used a quasi-experimental or other study design without a random allocation process.

Quality Appraisal

We used the revised Cochrane risk-of-bias tool for randomized trials (RoB 2) [26] to assess each included study in 5 domains: (1) risk of bias arising from the randomization process, (2) risk of bias due to deviations from the intended interventions (effect of assignment to intervention), (3) risk of bias due to missing outcome data, (4) risk of bias in outcome measurement, and (5) risk of bias in selection of the reported result. The assessor rated each domain as "low risk of bias," "some concerns," or "high risk of bias," which constituted an overall risk-of-bias judgment for the study. Every included study was assessed by at least 2 assessors; any discrepancies were resolved through a consensus discussion during our team meeting.

Data Extraction

We developed a Microsoft Excel spreadsheet to facilitate systematic data extraction through iterative discussions. We extracted the following data from the included studies: study details (authors, journal, year of publication, study purposes), study design (sample size, participant eligibility criteria, control type), interventions (theoretical foundations and app components), and outcomes, including data for calculating the effect size at study endpoints and follow-ups. In addition, we obtained intervention dose design information, if available, including frequency, duration, length, and timing of delivery, to examine the minimum effective intervention dose. For outcomes, we extracted primary outcomes relevant to anxiety and depression from the included studies. If a study did not indicate the primary outcome or had multiple primary outcome measures, we used data from the most used clinically validated instruments (ie, the State-Trait Anxiety Inventory [STAI] for anxiety and the Patient Health Questionnaire-9 [PHQ-9] for depression).

Data Synthesis

To pool the effect size of the interventions for each of the depression and anxiety measurements from the included studies with various measures, we computed Hedge g by taking the difference in the mean scores (1) between the intervention and control groups at each reported time point (between-group comparison) as well as (2) between the different time points following the interventions and the preintervention (ie, baseline) for the intervention groups (pre-post comparison). These time points included any reported time points during the interventions and during the follow-up after the conclusion of the interventions. For each comparison, we pooled and analyzed these Hedge g values for the target time point using both random-effect and fixed-effect models. The between-group and pre-post comparisons were also analyzed at the conclusion of the designed study intervention for depression and anxiety, respectively. Further, we used line graphs to visualize the pooled Hedge g values by time point, including follow-ups, to facilitate the analyses of dose-dependent effects and substantiality of the interventions.

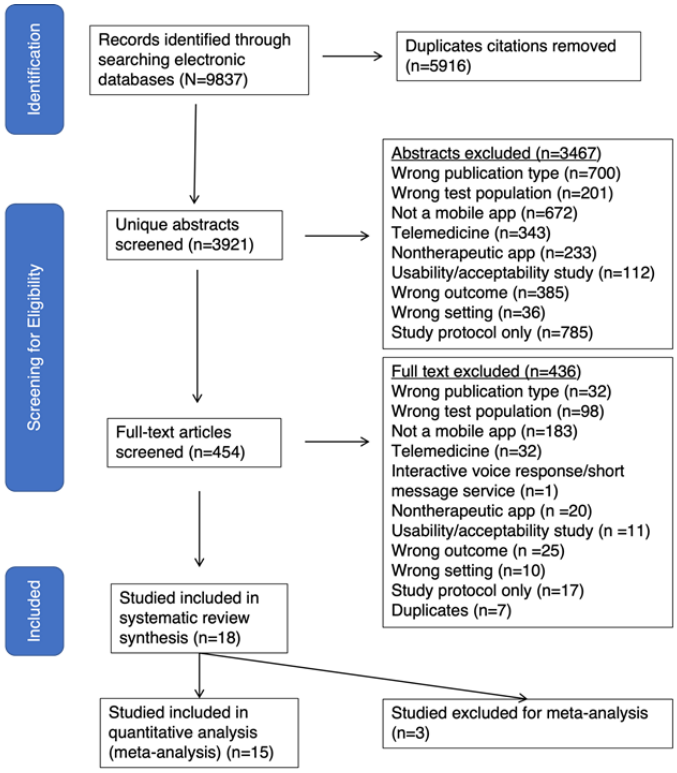
We evaluated heterogeneity between studies using I^2, which measures the percentage of total variance that can be explained. Study heterogeneity is considered low when I^2<25%, moderate when I^2 ranges from 25%-75%, and high when I^2>75% [27]. We also visually and statistically evaluated publication bias using funnel plots and the Egger test [28]. The pooled effect accounting for missing studies was assessed using the Duval and Tweedie trim-and-fill analysis [29]. In addition, we conducted subanalyses to compare the effect sizes generated from studies that targeted both depression and anxiety symptom reduction by pooling and analyzing Hedge g values at the end of the study.

Results

Study Selection and Characteristics

Our search strategy yielded a total of 9837 citations from the 7 databases, including ClinicalTrials. After removing duplicates, we screened 3921 (39.9%) abstracts and excluded 3467 (88.4%) citations that did not meet our inclusion criteria. We then reviewed 454 (11.6%) full-text articles and further excluded 436 (96%) studies based on our exclusion criteria (Figure 1). Of the remaining 18 (4%) studies, 15 (83%) were included in the meta-analysis; 3 (17%) studies did not report the data for meta-analysis [30-44].

Figure 1. PRISMA flowchart for study selection. PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses.



We summarize the characteristics, interventions, and primary outcomes of the studies included in our meta-analysis (N=15) in Tables 1 and 2. A total of 1942 participants were included in the 15 studies. These studies were conducted in the United States (n=4, 26%) [32,34,38,40], Germany (n=2, 13%) [36,39], Sweden (n=2, 13%) [31,43], Australia (n=1, 7%) [35], Japan (n=1, 7%) [37], Korea (n=1, 7%) [44], Switzerland (n=1, 7%) [41], Taiwan (n=1, 7%) [42], and the United Kingdom (n=1, 7%) [33]; in addition, 1 (7%) study recruited participants worldwide (the total percentage is more than 100% due to rounding) [30].

The most frequently targeted population was adults (age≥18 years) self-reporting anxiety or depression symptoms (n=6, 40%) [30,34,36,38,40,41]. Other examined populations included university students (n=2, 13%) [33,39]. Australian indigenous youth (n=1, 7%) [35], and people with a diagnosis of cancer (n=2, 13%) [32,44], social anxiety disorder (n=1, 7%) [31], major depressive disorder (n=2, 13%) [37,43], and general anxiety disorder (GAD; n=1, 7%) [31].

A total of 18 mobile apps were examined in the studies, with 8 (44%) targeting depression symptom management, 4 (22%) targeting anxiety reduction, and 6 (34%) targeting both anxiety and depression (Table 2). The majority of the mHealth apps facilitated various CBTs (n=12, 67%) [31-33,35-39,41-44]. Other therapies included ACT (n=1, 6%) [35], mindfulness and breathing relaxation techniques (n=1, 6%) [30], self-esteem and acceptance of the present (n=1, 6%) [40], and attentional bias modification (n=1, 6%) [42]. The length of intervention ranged from 4 to 12 weeks, with 4 weeks being the most commonly used length (n=5, 28%) [30,33,36,40,42]. Most apps were designed to be used on a daily basis.

Various instruments were used as primary outcome measurements. For depression, most studies used the Beck Depression Inventory-II (BDI-II) as their primary outcome measure (n=5, 33%) [38,41-44]. Other depression assessment tools included the PHQ-9 (n=4, 27%) [34-37] and the Center for Epidemiologic Studies Depression Scale (CES-D; n=2, 13%) [39,40]. There were no common anxiety assessment tools across the studies. There were a total of 8 different measurements used in the studies, including the STAI (n=2, 13%) [42,44], the 6-item short-form of the STAI (n=2, 13%) [33,39], GAD-7 (n=1, 7%) [30], the Hamilton Anxiety Rating Scale (HAM-A; n=1, 7%) [32], the Beck Anxiety Inventory (BAI; n=1, 7%) [43], the Liebowitz Social Anxiety Scale-Self Report (LSAS-SR; n=1, 7%) [31], and the Social Interaction Anxiety Scale (SIAS; n=1, 7%) [41].

Table 1. Characteristics of the included studies (N=15).

Author (year), country	Study populations/eligibility criteria	Sample size		Age (years), mean (SD)	Assessment time points	Outcome measures
		Intervention, n	Control, n			
Anxiety						
Pham (2016), global	Anxiety Sensitivity Index (ASI)-3≥16, Overall Anxiety Severity and Impairment Scale (OASIS)≥8, GAD-7[a]≥6	31	Waitlist: 32	18-34 (51)	Baseline, week 2, week 4 end point (EP)	GAD-7, ASI, OASIS
Boettcher (2018), Sweden	Diagnosis of social anxiety disorder (SAD), LSAS-SR[b]≥30	70	Bibliotherapy: 70; waitlist: 69	Intervention group (Txt): 35.4 (11.0); bibliotherapy: 35.9 (14.1); control group (Ctrl): 35.0 (11.6)	Baseline, week 3, week 7 (EP), follow-up (FU) week 3, FU week 7, FU week 9, FU week 41	LSAS-SR, PHQ[c]-9, GAD-7
Greer (2019), United States[d]	Age≥18 years, diagnosis of incurable solid tumor, Hospital Anxiety and Depression Scale (HADS) anxiety subscale>7, Eastern Cooperative Oncology Group (ECOG)=0-2	72	Education control: 73	Txt: 55.9 (12.4); Ctrl: 57.0 (10.1)	Baseline, week 12 (EP)	HAM-A[e], HADS, PHQ-9
Ponzo (2020), United Kingdom[d]	University students, Depression Anxiety Stress Scales (DASS)-21 stress subscale>14 or DASS-21 anxiety subscale>7	72	Waitlist: 74	Txt: 19.9 (1.83); Ctrl: 19.8 (1.8)	Baseline, week 2, week 4 (EP), FU week 2	STAI[f]-S-6, PHQ-9, DASS-21
Depression						
Stile-Shields (2019), United States[d]	Age≥18 years, PHQ-9>10, Quick Inventory of Depressive Symptoms (QIDS)>11	Boost me=10; thought challenger=10	Waitlist: 10	N/R[g]	Baseline, week 3, week 6 (EP), FU week 4	PHQ-9
Tighe (2017), Australia	Australian indigenous youth (age 18-35 years), PHQ-9>10 or 10-item Kessler Psychological Distress Scale (K10)>25	31	Waitlist: 30	Txt: 27.5 (9.5); Ctrl: 25.0 (6.3)	Baseline, week 6 (EP)	PHQ-9
Ludtke (2018), Germany	Subjective need for a depression symptom reduction intervention	44	Waitlist: 44	Txt: 41.2 (11.9); Ctrl: 44.6 (10.7)	Baseline, week 4 (EP)	PHQ-9
Mantani (2017), Japan	Age 25-59 years, diagnosis of major depressive disorder, BDI[h]-II≥10, currently taking and resistant to 1 antidepressant	81	Medication change only: 83	Txt: 40.2 (8.8); Ctrl: 41.6 (8.9)	Baseline, week 5, week 9 (EP), FU week 8	PHQ-9, BDI-II
Dahne (2019), United States[d]	Age 18-65 years, PHQ-8>10	Moodivate: 24; MoodKit: 19	Treatment as usual (TAU): 9	Moodivate: 43.8 (13.3); MoodKit: 44.7 (14.0); Ctrl: 43.1 (11.9)	Week 2, week 3, week 4, week 5, week 6, week 7, week 8 (EP)	BDI-II
Both anxiety and depression						
Harrer (2018), Germany	University students, perceived stress posttreatment (PSS)-4≥8	75	Waitlist: 75	Txt: 24.0 (4.6); Ctrl: 24.2 (3.6)	Baseline, week 7 (EP), FU week 5	STAI-6, CES-D[i]
Roepke (2015), United States	Age≥18 years, CES-D≥16	General SB: 97; CBT[j]/positive psychotherapy SuperBetter (PPT SB): 93	Waitlist: 93	CBT/PPT SB: 42.3 (12.6); general SB: 38.0 (11.3); Ctrl: 40.3 (13.1)	Baseline, week 2, week 4 (EP), FU week 2	CES-D, GAD-7

Author (year), country	Study populations/eligibility criteria	Sample size		Age (years), mean (SD)	Assessment time points	Outcome measures
		Intervention, n	Control, n			
Stolz (2018), Switzerland	Age≥18 years, ≥cut-off score on SIAS[k] or Social Phobia Scale (SPS), *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition* (DSM-IV) diagnosis of SAD	60	Waitlist: 30	Txt: 34.7 (9.9); Ctrl: 35.2 (12.1)	Baseline, week 12 (EP), FU week 12	SIAS, LSAS-SR, BDI-II
Teng (2019), Taiwan	Age 25-35 years, PSWQ>60, DMS-IV diagnosis of GAD subscale	30	Placebo: 30; waitlist: 22	Txt: 21.5 (2.2); placebo: 21.5 (1.6); waitlist: 21.5 (1.6)	Baseline, week 2, week 3, week 4 (EP), FU week 4	STAI-S, STAI-T, BDI-II, BAI[l]
Ly (2015), Sweden	Age≥18 years, PHQ-9≥5, DMS-IV diagnosis of major depression	46	Face-to-face behavior activation therapy: 47	Txt: 30.2 (11.9); Ctrl: 31.0 (11.0)	Baseline, week 9 (EP), FU week 24	BDI-II, PHQ-9, BAI
Ham (2019), Korea[d]	Age 16-65 years, diagnosis of cancer, BDI-II≥16 or STAI>39	28	Waitlist: 26; attention control: 26	Txt: 41.9 (11.3); attention control: 43.5 (10.4); waitlist control: 47.1 (11.2)	Baseline, week 10 (EP)	BDI-II, STAI-T, STAI-S

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

[b]LSAS-SR: Liebowitz Social Anxiety Scale-Self Report.

[c]PHQ: Patient Health Questionnaire.

[d]Studies were not included in the previous meta-analyses we identified.

[e]HAM-A: Hamilton Anxiety Rating Scale.

[f]STAI: State-Trait Anxiety Inventory.

[g]N/R: not reported.

[h]BDI: Beck Depression Inventory.

[i]CES-D: Center for Epidemiological Studies Depression Scale.

[j]CBT: cognitive behavioral therapy.

[k]SIAS: Social Interaction Anxiety Scale.

[l]BAI: Beck Anxiety Inventory.

Table 2. Intervention characteristics of the included studies (N=15).

Author (year), country	App contents	Intended dose	Length	Additional components
Anxiety				
Pham (2016), global	• Flowy app: minigames for breathing retraining with reward feedback	N/R[a]	4 weeks	N/A[b]
Boettcher (2018), Sweden	• CBT[c] with gamification and life skill challenges	Daily use	6 weeks	Internet-based CBT with 9 modules
Greer (2019), United States[d]	• CBT with psychoeducation, activity planning, problem solving, staying present, thought creation, and summary/review	6 sessions (20-30 minutes each) with homework (10 -15 minutes each)	10-12 weeks	N/A
Ponzo (2020), United Kingdom[d]	• BioBase: CBT and self-compassion-based psychoeducational content, mood tracking, and relaxation exercises	Daily use	4 weeks	"Biobeam" wristband for passive data collection (physical activity, sleep pattern, and heart rate)
Depression				
Stile-Shields (2019), United States[d]	• Boost Me: behavioral activation (BA) with activity scheduling, aiming to increase rewarding activities and monitoring of mood • Thought Challenger: CBT involving identifying and apprising maladaptive thoughts and creating adaptive counter thoughts	N/R	6 weeks	Weekly coaching via phone or email to enhance intervention adherence
Tighe (2017), Australia	• iBobbly: ACT[e] with identifying thoughts, feelings, and behaviors; learning distancing techniques; regulating emotions through mindfulness, acceptance, and self-soothing activities; and identifying values, goals, personalized action plans	N/R	6 weeks	N/A
Ludtke (2018), Germany	• Good to Yourself: CBT with cognitive strategies, mindfulness, social competence skills, activating exercises	A few minutes per day	4 weeks	N/A
Mantani (2017), Japan	• Kokoro: CBT, mood monitoring, BA, and homework	1 session/week with 20 minutes/session (not including homework)	8 weeks	Antidepressant switch to escitalopram (5-10 mg/day) or to sertraline (25-100 mg/day)
Dahne (2019), United States[d]	• Moodivate: BA (psychoeducation, value identification, activity planning based on values, completion badges) • MoodKit: CBT (thought identification/modification, mood tracking, journaling, activity scheduling)	At least once per day	8 weeks	N/A
Both anxiety and depression				
Harrer (2018), Germany	• CBT with social support, rumination, time management, procrastination, text anxiety, sleep, motivation, nutrition, exercise, mood diary, motivational messages, and online eCoach	30-90 minutes/module with 1-2 modules/week for 8 modules total	7 weeks	N/A
Roepke (2015), United States	• SuperBetter: gamified app to increase drive to accomplish goals and build social support • SuperBetter" version with CBT/positive psychotherapy (PPT): same app with additional CBT content adapted from PPT and 2 classic CBT (cognitive restructuring and behavioral activation)	10 minutes/day	4 weeks	N/A

Author (year), country	App contents	Intended dose	Length	Additional components
Stolz (2018), Switzerland	• CBT with motivational enhancement, psychoeducation, cognitive restructuring, self-focused attention, behavioral experiments, summary and repetition, healthy lifestyle and problem solving, and relapse prevention	1 module/week	12 weeks	Weekly feedback from a coach
Teng (2019), Taiwan[d]	• Home-delivered attentional bias modification (HD-ABM): administers attention training for which disgusted and neutral facial expressions are used as stimuli; target "probe" replacing only the neutral face	3 times/day	4 weeks	N/A
Ly (2015), Sweden	• CBT with recall (statistics and summaries) and save important nondepressed behavior, a behavior activity database for providing suggestions, support, and inspiration; a bake-end system for therapists monitoring participants' activities; and a messaging system for communication between participants and therapists	N/R	9 weeks	Face-to-face behavior activation therapy
Ham (2019), Korea[d]	• HARUToday: CBT with psychoeducation, BA, relaxation training, cognitive restructuring, problem solving, and point reward system	10-15 minutes/session with a quiz for 48 sessions	10 weeks	N/A

[a]N/R: not reported.

[b]N/A: not applicable.

[c]CBT: cognitive behavioral therapy.

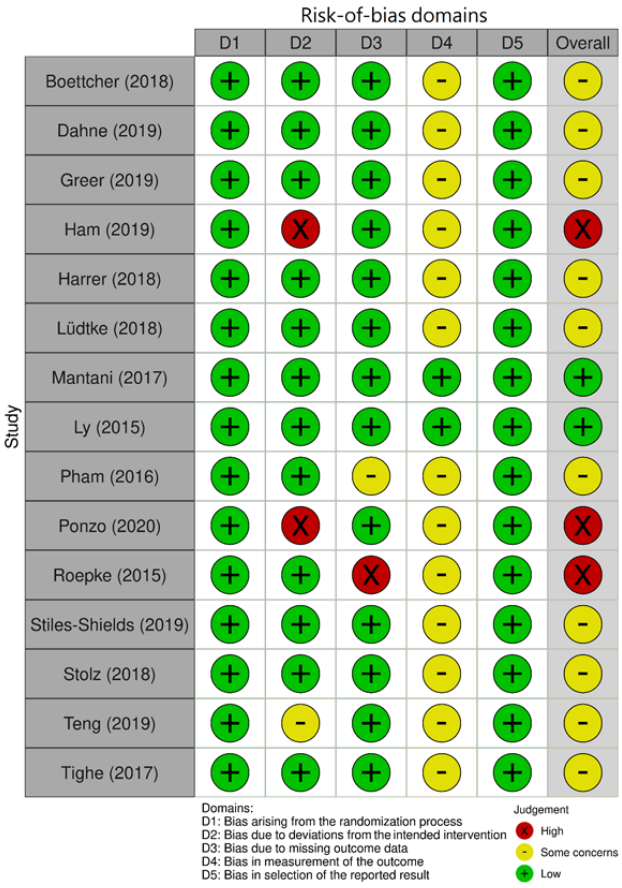
[d]Studies were not included in the previous meta-analyses we identified.

[e]ACT: acceptance and commitment therapy

Risk-of-Bias Assessment

Most studies (n=10, 67%) [30-32,34-36,38,39,41,42] were rated as "some concerns" for bias, and 3 (20%) [33,40,44] were rated as "high risk of bias" (Figure 2). All studies reported adequate randomization sequence generation and allocation concealment. In addition, 3 (20%) studies [33,42,44] reported unclear information concerning their approaches adjusting the effects of intervention nonadherence on outcomes, and 2 (13%) studies had a high attrition rate and provided no information about their approaches addressing missing data. Blinding of outcome assessment was not possible for most included studies due to the use of self-reported outcome assessments; thus, the results of most studies (n=13, 87%) [30-34,36-41,43,44], although unlikely, may be influenced by the awareness of the intervention received. Concerning outcome reporting, we found no evidence to suspect selective reporting for all studies. There was no evidence of publication bias according to the funnel plots and Egger test (Multimedia Appendix 2).

Figure 2. Diagram summarizing the result of our risk-of-bias evaluation among the 15 included studies using the Cochrane risk-of-bias tool for RCTs. RCT: randomized control trial.



Effectiveness of mHealth Apps in Anxiety and Depression

Of the included 15 studies, 10 (67%) [30-33,40-42] examined the effectiveness of app-based mHealth interventions in anxiety management. When compared to the preintervention, at the conclusion of the interventions, participants receiving the interventions showed a statistically significant effect on anxiety symptoms (Hedge g=–.20, 95% CI –0.31 to –0.09, heterogeneity I^2=0%, *P*=.79); see Figure 3a. Similarly, when compared to the control groups, at the conclusion of the interventions, participants receiving the interventions showed a statistically significant effect on anxiety symptoms (Hedge g=–.10, 95% CI –0.14 to –0.05, heterogeneity I^2=0%, *P*>.99); see Figure 3b.

Of the included 15 studies, 11 (73%) [32-36,39-44] evaluated the effectiveness of app-based mHealth interventions in depression management. When compared to the preintervention, at the conclusion of the interventions, participants receiving the interventions showed a statistically significant effect on depression symptoms (Hedge g=–.25, 95% CI –0.39 to –0.11, heterogeneity I^2=3%, *P*=.42); see Figure 4a. However, when compared to the control groups, at the conclusion of the interventions, participants receiving the interventions did not show a statistically significant effect on depression symptoms (Hedge g=–.08, 95% CI –0.23 to 0.07, heterogeneity I^2=4%, *P*=.41); see Figure 4b.

Figure 3. Pooled effect size of mHealth apps on anxiety symptom management at the conclusion of the intervention: (a) before-after comparison for the intervention groups and (b) comparison between intervention and control groups. BAI: Beck Anxiety Inventory; CBT: cognitive behavioral therapy; GAD: generalized anxiety disorder; HAM-A: Hamilton Anxiety Rating Scale; HD-ABM: home-delivered attentional bias modification; LSAS-SR: Liebowitz Social Anxiety Scale-Self Report; mHealth: mobile health; PPT: positive psychotherapy; STAI: State-Trait Anxiety Inventory; WL: waitlist.

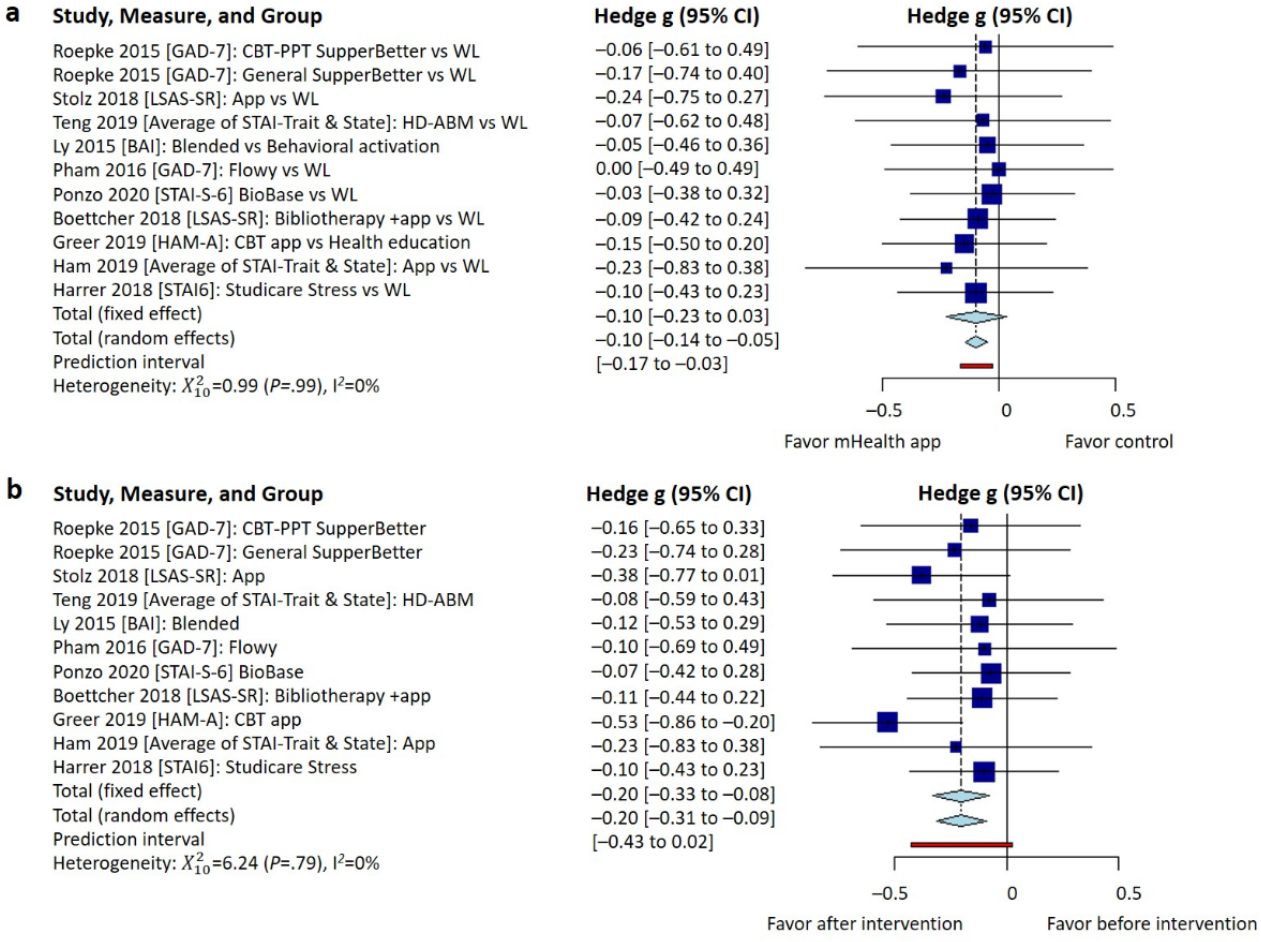
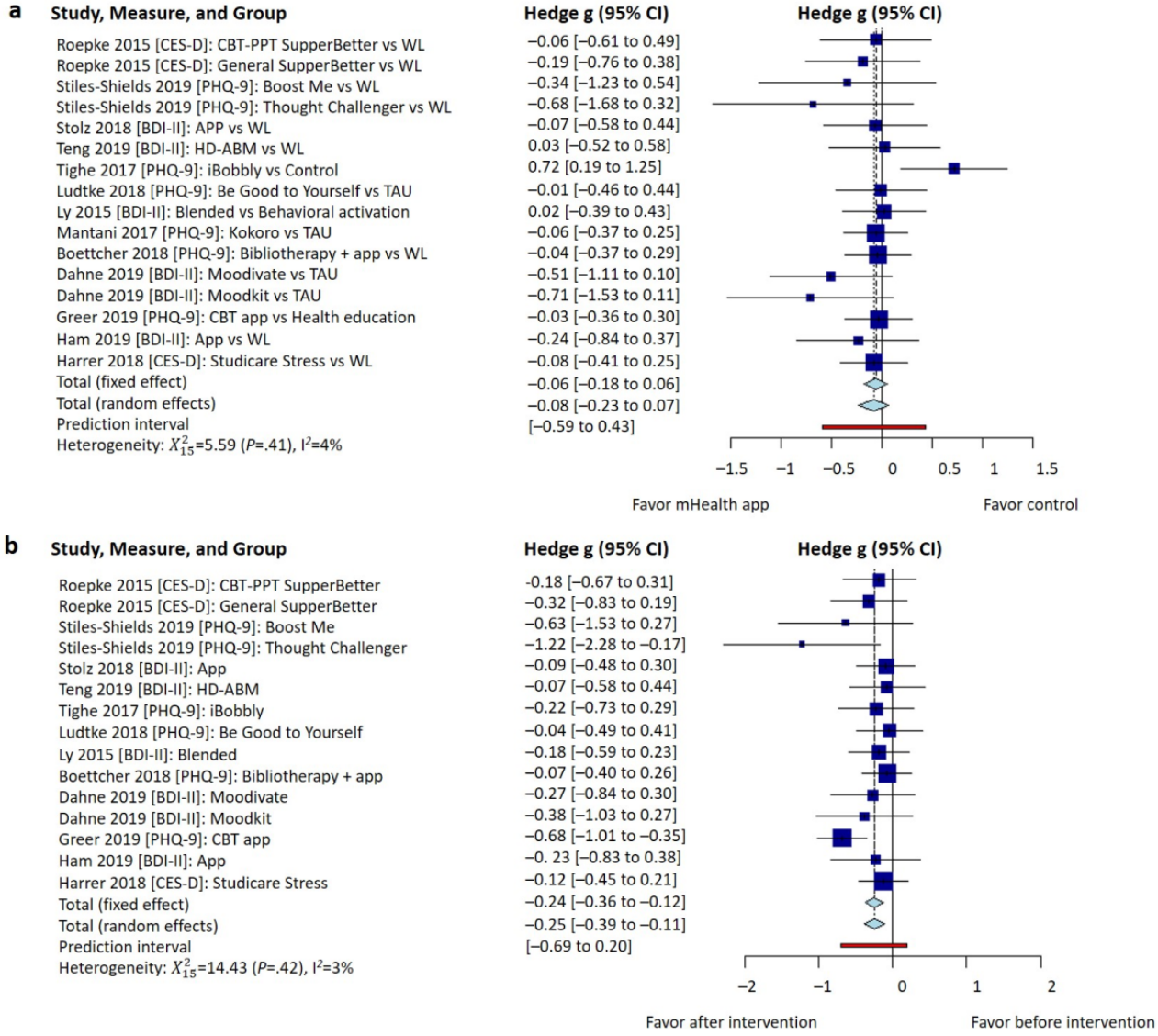


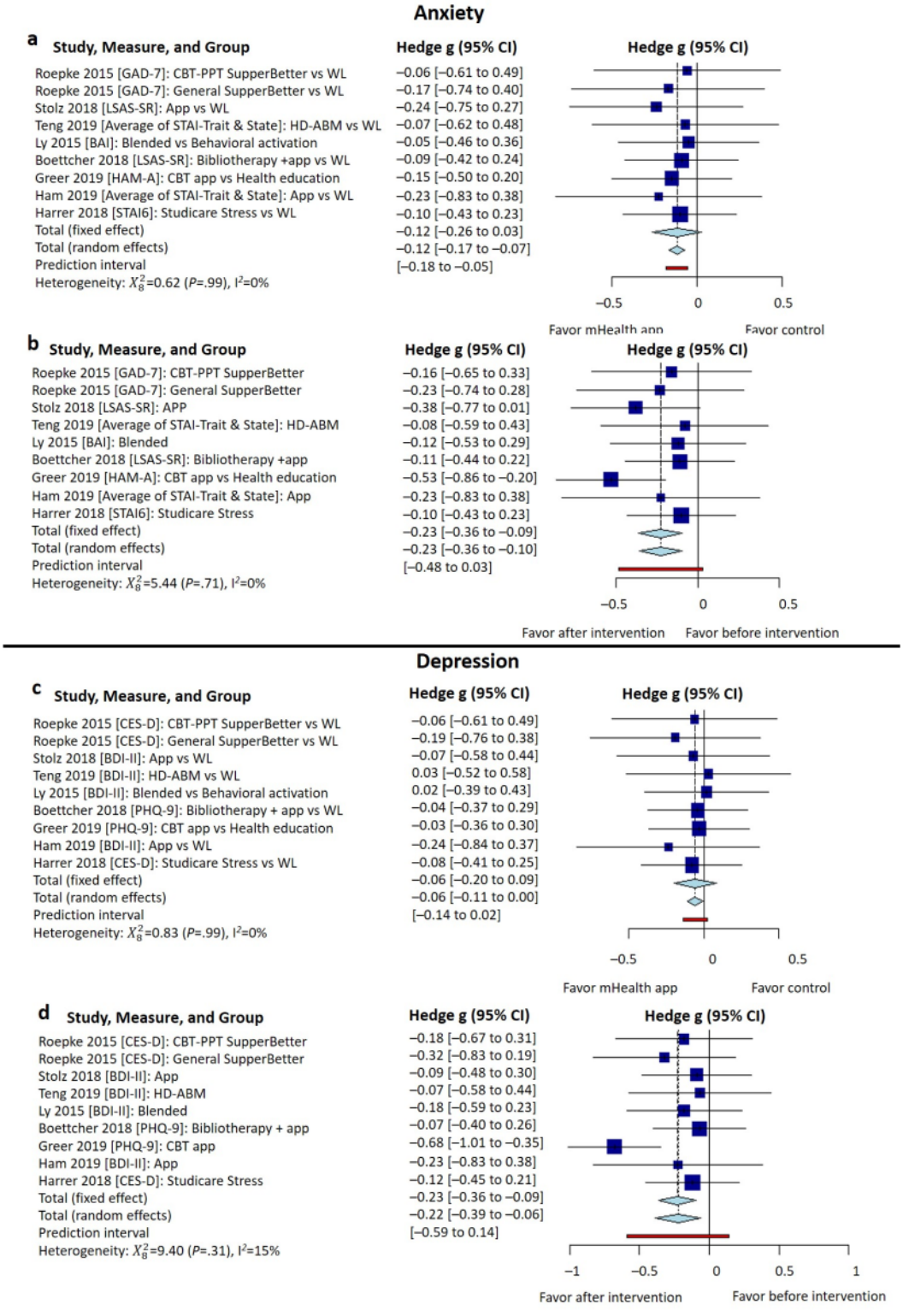
Figure 4. Pooled between-group effectiveness of mHealth apps on depressive symptom management: (a) before-after comparison for the intervention groups and (b) comparison between intervention and control groups. BDI: Beck Anxiety Inventory; CBT: cognitive behavioral therapy; CES-D: Center for Epidemiological Studies Depression questionnaire; HD-ABM: home-delivered attentional bias modification; mHealth: mobile health; PHQ: Patient Health Questionnaire; PPT: positive psychotherapy; TAU: treatment-as-usual; WL: waitlist.



Effects of mHealth Interventions on Depression vs Anxiety

Our subgroup analysis included 8 (53%) studies [31,32,39-44] evaluating the effectiveness of their interventions in both depression and anxiety. The results indicated that the intervention groups showed a significant effect on both anxiety (Hedge g=–.23, 95% CI –0.36 to –0.10, heterogeneity I^2=0%, *P*>.99) and depression (Hedge g=–.22, 95% CI –0.39 to –0.06, heterogeneity I^2=15%, *P*=.31) compared to baseline (Figures 5a and 5b). However, compared to the control groups (waiting list), mHealth interventions showed a significant effect only on anxiety at the conclusion of the interventions (Figure 5c) but not on depression (Figure 5d). This shows that mHealth interventions are more likely to improve anxiety but not depression.

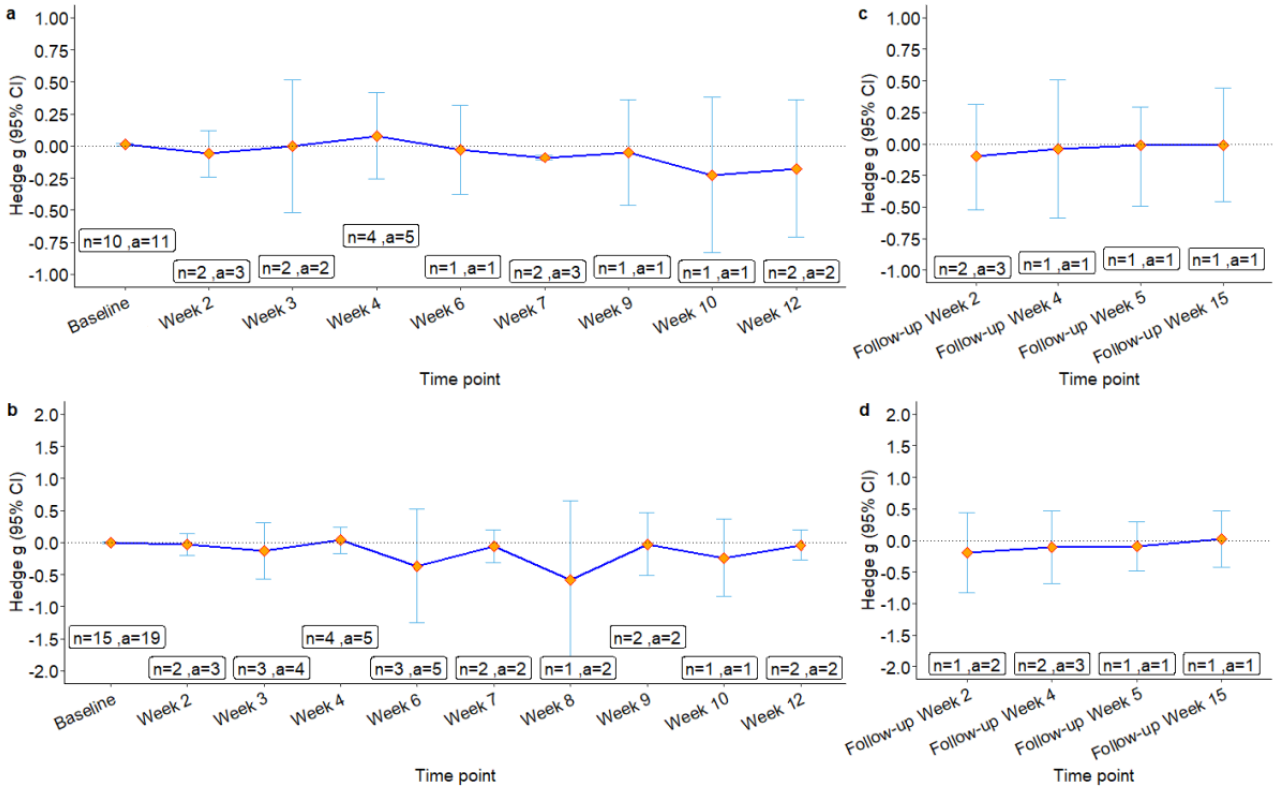
Figure 5. Subanalysis of pooled within-group and between-group effects of mHealth interventions on anxiety (upper panel) and depression (lower panel) from studies evaluating intervention effects on both anxiety and depression (n=8): (a) within-group comparison for the intervention groups for anxiety, (b) comparison between intervention and control groups for anxiety, (c) within-group comparison for the intervention groups for depression, and (d) comparison between intervention and control groups for depression. BAI: Beck Anxiety Inventory; BDI: Beck Anxiety Inventory; CBT: cognitive behavioral therapy; CES-D: Center for Epidemiological Studies Depression questionnaire; GAD: generalized anxiety disorder; HAM-A: Hamilton Anxiety Rating Scale; HD-ABM: home-delivered attentional bias modification; LSAS-SR: Liebowitz Social Anxiety Scale-Self Report; mHealth: mobile health; PHQ: Patient Health Questionnaire; PPT: positive psychotherapy; STAI: State-Trait Anxiety Inventory; TAU: treatment-as-usual; WL: waitlist.



Dose-Dependent Effects of the mHealth Interventions

When examining the dose-dependent effects of the mHealth interventions, interventions longer than 7 weeks had larger effect sizes on anxiety reduction, with a statistically significant effect size at week 7 (Figure 6a). In contrast, the pooled effects on depression fluctuated without a clear trend of dose-dependent effects (Figure 6b). Regarding the sustainability of intervention effects, the pooled effect sizes were not significant and reduced over time during follow-ups for both anxiety and depression (Figures 6c and 6d).

Figure 6. Pooled effects of the app-based mHealth intervention on anxiety (upper panel) and depression (lower panel) at different time points: (a) during the designed study intervention length and (b) during the follow-up after the designed study intervention. a: number of study arms; mHealth: mobile health.



Discussion

Principal Findings

We conducted a systematic review and meta-analysis to examine the existing evidence on the effectiveness of app-based mHealth interventions for anxiety and depression symptom reduction. We included a total of 15 randomized controlled trials (RCTs), with many studies [32-34,38,44] published after previous reviews on a similar topic, providing an update to the current evidence. Our meta-analysis shows that app-based mHealth interventions have a modest but significant effect on anxiety reduction, consistent with previous reviews [4,11,12]. This finding adds confidence to the further development and implementation of smartphone apps to facilitate psychological treatments for anxiety symptom management [11]. In addition, our results suggest that a longer intervention (ie, 7 weeks or longer) is more likely to result in significant anxiety reduction. This finding may explain the restricted effects in studies with less than 7 weeks of app-based mHealth interventions. To the best of our knowledge, this is the first meta-analysis to assess the relationship between the app-based mHealth intervention length and the effect of the intervention. We encourage researchers to design a longer app-based mHealth intervention for anxiety symptom control and to verify our findings regarding the length of the intervention.

With regard to depression, we found that participants receiving interventions for depression experienced little symptom reduction that was not statistically significant. The finding is inconsistent with other systematic reviews reporting that smartphone apps have small-to-moderate effect sizes on depression symptom reduction [4,10,12]. The inconsistency could result from the fact that we included 1 measure per outcome per study instead of averaging the data from studies using multiple measurements for an outcome. The inconsistency may also be because previous studies included both native smartphone and web-based apps [4,10]. Web-based apps have better accessibility by allowing participants to access the interventions via various platforms [45]. In addition, the long history of web app development led to optimal user interface design, contributing to better usability and usefulness. Usability, usefulness, and accessibility have been documented as the key factors leading to successful and effective apps for mental illness management [46]. Nevertheless, we decided to exclude web-based apps because most studies reported no information about the tools their participants used to access their apps, which diminishes the purpose of our analysis on mHealth apps.

Another possible explanation of the consistency between the results of this and previous studies can be that we included 5 [32-34,38,44] studies published after previous reviews and all of them had insignificant effect sizes in our analysis. The effect

sizes from the new evidence may neutralize the effect sizes from the studies included in the previous reviews. The intervention effect heterogeneity indicated that the optimal intervention content, format, and dose designs remain unclear. This is further supported by our dose-dependent analysis revealing that there is no clear relationship between the intervention length and the effect on depression, similar to a previous study [10].

There were 3 RCTs that met our eligibility criteria but were excluded from our meta-analysis due to insufficient data reported for the analysis [47-49]. All 3 studies reported positive results toward the effects of smartphone apps facilitating CBT on mental illness. Li et al [47] conducted a 12-week RCT and reported that a CBT-based smartphone chatbot intervention is efficacious for depression symptom reduction for patients with HIV and depression at both 3 and 6 months [47]. Morbeg et al [48] conducted a 4-week RCT and found that adult people receiving a CBT-based smartphone app had significantly lower anxiety and depression symptoms. Lastly, Arean et al [49] examined 2 smartphone apps in a 4-week RCT for depression and found that both apps generated a greater reduction, although not significant, in the depression symptom score compared to the control. However, we excluded these 3 studies because they either reported statistics that cannot be used to compute Hedge g without transformation based on assumptions or did not report enough data for Hedge g calculation. Studies by Li et al [47] and Morbeg et al [48] were also not included in other previous systematic reviews. The study by Arean et al [49], after data transformation with assumptions, was included in previous reviews but showed inconsistent effects on depression symptom reduction. Therefore, it was unclear whether the inclusion of these studies would alter our results for depression. Further researchers and reviewers should emphasize the gold standard of reporting to enable better study comparison and synthesis [25,50].

Consistent with previous reviews (eg, Lui et al [9]), the majority of the included studies used mobile apps to deliver CBT for anxiety or depression or both. Cognitive behavioral therapy has been delivered by computer or web apps for the treatment of various mental illnesses [51]. Our results did not suggest that smartphone apps are not useful for facilitating CBT. Rather, our results suggest that current evidence may be insufficient to guide the app-based mHealth intervention design for effective CBT-based mental illness intervention facilitation, thus requiring more research engagement. In addition, other psychotherapies, such as ACT, may also be effective in mental illness control but received relatively less attention. More studies are needed to uncover whether smartphone apps can facilitate other psychotherapies and how effective they are.

One objective of our study was to evaluate the current dose design of existing app-based mHealth interventions for anxiety and depression for an understanding of the optimal mHealth treatment length. We found that most interventions were designed to be used on a daily basis and completed within 1.5 months [52]. However, most studies provided a paucity of information about how much time their participants were asked to spend on the interventions per day or per module/session of the interventions; in addition, most studies reported no data on how much time their participants actually spent on the interventions (the actual intervention exposure). As a result, we were only able to summarize the intervention effect by the designed intervention length and dose reported in the included studies.

Limitations and Strengths

Our review has several limitations that should be considered when interpreting the results. First, our literature search was restricted to English publications and resulted in a small amount of research available compared to other meta-analyses examining the evidence of smartphone-based interventions for mental illness. Second, the included studies used various outcome measures, and we extracted only the primary or secondary measures for anxiety and depression. Although this strategy was used in previous systematic reviews and meta-analyses on similar topics, we might have missed the effects detected by other measures. Both limitations might result in our findings of limited or nonexistent efficacy of the interventions and confidence reduction in our dose analysis results. Finally, we included 6 studies that delivered their interventions in part by smartphone. Although app components were the main parts of their interventions, our results may not represent the effects of stand-alone smartphone apps due to the inclusion of the studies. Nevertheless, we decided to include these studies because we considered these interventions were still app-based mHealth interventions. In addition, small effect sizes for 4 of the 6 studies suggest that the nonapp components do not seem to contribute to the primary effect. Further studies, including more studies for blended interventions (smartphone app + other intervention components), are needed to compare the effects of stand-alone smartphone apps and blended interventions on mental illness management.

Despite the limitations, this review has many strengths. First, our included studies covered several publications that were published after 2019 [32-34,38,44] to reflect updated evidence, which can support future development and use of app-based mHealth interventions for anxiety and depression. Second, we conducted several analyses assessing pooled intervention effects at various study time points to understand the effective length of app-based mHealth interventions. Finally, we computed the pooled effect size of the mHealth interventions during the follow-up period to uncover the sustainability of the intervention effects on anxiety and depression reduction, which was not revealed in previous systematic reviews focusing on a similar topic [10-12]. These analyses provide innovative insights informing the future study design of app-based mHealth interventions assessing for anxiety and depression symptom reduction.

Implications for Future Studies

The dose design of app-based interventions has been suggested as an important aspect that profoundly influences intervention effects [13,24,53]. However, incomplete and inconsistent reporting of the intervention dose design and exposure in the existing studies impeded our quantitative analysis exploring the optimal intervention dose design for anxiety and depression. Future studies should explore the effect of app-based mHealth interventions with various dose designs and exposures for anxiety and depression symptom management. In addition,

research efforts are needed to improve the reporting of intervention doses to enable comparable data for evidence evaluation and synthesis. The use of the JITAI framework to inform intervention design, evaluation, and reporting has potential to enable high-quality evidence for future app-based mHealth interventions for mental illness [13,24]. Finally, although most studies reported that their interventions sustained over follow-up compared to baseline, our analysis indicated that the pooled between-group effects of the interventions were not significant and rapidly reduced over time for both anxiety and depression. We recommend future studies to further explore the sustainability of symptom improvements from app-based mHealth interventions for anxiety and depression at various time points, including both during the study and after study completion (follow-up).

Conclusion

In summary, although there is some evidence in using app-based mHealth interventions for anxiety and depression symptom reduction, clinical use cannot be recommended based on this systematic review and meta-analysis due to the small to nonexistent pooled effects found in existing studies, not to mention concerns regarding study quality/reporting of the existing studies. The effects of app-based mHealth interventions may not yet be realized, as the optimal intervention dose is still unclear. Future research should consider (1) adopting a theoretical framework, such as JITAI, to inform intervention design, evaluation, and reporting to enable high-quality evidence for app-based mHealth interventions for anxiety and depression care; (2) improving the reporting of data to enable comparable data for evidence evaluation and synthesis; and (3) exploring the sustainability of treatment benefit from the mHealth interventions.

Authors' Contributions

SCL, MX, ALC, and PYY conceptualized and designed the study; AH conducted the literature search and reference management: SCL, MX, and PYY selected relevant studies and performed data collection and assembly; SCL, MW, PYY, and S-HC conducted data analysis and result interpretation; SCL, PYY, and S-HC drafted the manuscript; and PYY and S-HC jointly provided supervision, timeline control, and resource management of the study as senior authors. All authors participated in the revision of the manuscript and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategy for all reference databases used.
[DOCX File , 18 KB - mental_v9i9e39454_app1.docx]

Multimedia Appendix 2
Results of the funnel plots and Egger test.
[DOCX File , 158 KB - mental_v9i9e39454_app2.docx]

References

1. GBD 2017 Disease and Injury Incidence and Prevalence Collaborators. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet 2018 Nov 10;392(10159):1789-1858 [FREE Full text] [doi: 10.1016/S0140-6736(18)32279-7] [Medline: 30496104]
2. Mental Health America (MHA). The State of Mental Health in America. 2021. URL: https://mhanational.org/issues/state-mental-health-america [accessed 2022-08-26]
3. McNair BG, Highet NJ, Hickie IB, Davenport TA. Exploring the perspectives of people whose lives have been affected by depression. Med J Aust 2002 May 20;176(10):S69-S76. [doi: 10.5694/j.1326-5377.2002.tb04507.x] [Medline: 12065001]
4. Linardon J, Cuijpers P, Carlbring P, Messer M, Fuller-Tyszkiewicz M. The efficacy of app-supported smartphone interventions for mental health problems: a meta-analysis of randomized controlled trials. World Psychiatry 2019 Oct;18(3):325-336 [FREE Full text] [doi: 10.1002/wps.20673] [Medline: 31496095]
5. Torous J, Roberts LW. Needed Innovation in Digital Health and Smartphone Applications for Mental Health: Transparency and Trust. JAMA Psychiatry 2017 May 01;74(5):437-438. [doi: 10.1001/jamapsychiatry.2017.0262] [Medline: 28384700]
6. Miralles I, Granell C, Díaz-Sanahuja L, Van Woensel W, Bretón-López J, Mira A, et al. Smartphone apps for the treatment of mental disorders: systematic review. JMIR Mhealth Uhealth 2020 Apr 02;8(4):e14897 [FREE Full text] [doi: 10.2196/14897] [Medline: 32238332]
7. Byambasuren O, Sanders S, Beller E, Glasziou P. Prescribable mHealth apps identified from an overview of systematic reviews. NPJ Digit Med 2018;1(1):1-12 [FREE Full text] [doi: 10.1038/s41746-018-0021-9] [Medline: 31304297]

8. Gratzer D, Strudwick G, Yeung A. Mental illness: is there an app for that? Fam Syst Health 2019 Dec;37(4):336-339. [doi: 10.1037/fsh0000451] [Medline: 31815514]
9. Lui JHL, Marcus DK, Barry CT. Evidence-based apps? A review of mental health mobile applications in a psychotherapy context. Professional Psychology: Research and Practice 2017 Jun;48(3):199-210. [doi: 10.1037/pro0000122]
10. Firth J, Torous J, Nicholas J, Carney R, Pratap A, Rosenbaum S, et al. The efficacy of smartphone-based mental health interventions for depressive symptoms: a meta-analysis of randomized controlled trials. World Psychiatry 2017 Oct;16(3):287-298 [FREE Full text] [doi: 10.1002/wps.20472] [Medline: 28941113]
11. Firth J, Torous J, Nicholas J, Carney R, Rosenbaum S, Sarris J. Can smartphone mental health interventions reduce symptoms of anxiety? A meta-analysis of randomized controlled trials. J Affect Disord 2017 Aug 15;218:15-22 [FREE Full text] [doi: 10.1016/j.jad.2017.04.046] [Medline: 28456072]
12. Weisel KK, Fuhrmann LM, Berking M, Baumeister H, Cuijpers P, Ebert DD. Standalone smartphone apps for mental health-a systematic review and meta-analysis. NPJ Digit Med 2019;2(1):1-10 [FREE Full text] [doi: 10.1038/s41746-019-0188-8] [Medline: 31815193]
13. Nahum-Shani I, Smith SN, Spring BJ, Collins LM, Witkiewitz K, Tewari A, et al. Just-in-Time Adaptive Interventions (JITAIs) in Mobile Health: Key Components and Design Principles for Ongoing Health Behavior Support. Ann Behav Med 2018 May 18;52(6):446-462 [FREE Full text] [doi: 10.1007/s12160-016-9830-8] [Medline: 27663578]
14. Spruijt-Metz D, Wen CKF, O'Reilly G, Li M, Lee S, Emken BA, et al. Innovations in the Use of Interactive Technology to Support Weight Management. Curr Obes Rep 2015 Dec;4(4):510-519 [FREE Full text] [doi: 10.1007/s13679-015-0183-6] [Medline: 26364308]
15. King AC, Hekler EB, Grieco LA, Winter SJ, Sheats JL, Buman MP, et al. Harnessing different motivational frames via mobile phones to promote daily physical activity and reduce sedentary behavior in aging adults. PLoS One 2013;8(4):e62613 [FREE Full text] [doi: 10.1371/journal.pone.0062613] [Medline: 23638127]
16. Hardeman W, Houghton J, Lane K, Jones A, Naughton F. A systematic review of just-in-time adaptive interventions (JITAIs) to promote physical activity. Int J Behav Nutr Phys Act 2019 Apr 03;16(1):31 [FREE Full text] [doi: 10.1186/s12966-019-0792-7] [Medline: 30943983]
17. Brookie KL, Mainvil LA, Carr AC, Vissers MCM, Conner TS. The development and effectiveness of an ecological momentary intervention to increase daily fruit and vegetable consumption in low-consuming young adults. Appetite 2017 Jan 01;108:32-41. [doi: 10.1016/j.appet.2016.09.015] [Medline: 27642037]
18. Heron KE, Smyth JM. Ecological momentary interventions: incorporating mobile technology into psychosocial and health behaviour treatments. Br J Health Psychol 2010 Mar;15(1):1-39 [FREE Full text] [doi: 10.1348/135910709X466063] [Medline: 19646331]
19. Svetkey LP, Stevens VJ, Brantley PJ, Appel LJ, Hollis JF, Loria CM, Weight Loss Maintenance Collaborative Research Group. Comparison of strategies for sustaining weight loss: the weight loss maintenance randomized controlled trial. JAMA 2008 Mar 12;299(10):1139-1148. [doi: 10.1001/jama.299.10.1139] [Medline: 18334689]
20. Rodgers A, Corbett T, Bramley D, Riddell T, Wills M, Lin R, et al. Do u smoke after txt? Results of a randomised trial of smoking cessation using mobile phone text messaging. Tob Control 2005 Aug;14(4):255-261 [FREE Full text] [doi: 10.1136/tc.2005.011577] [Medline: 16046689]
21. Suffoletto B, Callaway C, Kristan J, Kraemer K, Clark DB. Text-message-based drinking assessments and brief interventions for young adults discharged from the emergency department. Alcohol Clin Exp Res 2012 Mar;36(3):552-560. [doi: 10.1111/j.1530-0277.2011.01646.x] [Medline: 22168137]
22. Witkiewitz K, Desai SA, Bowen S, Leigh BC, Kirouac M, Larimer ME. Development and evaluation of a mobile intervention for heavy drinking and smoking among college students. Psychol Addict Behav 2014 Sep;28(3):639-650 [FREE Full text] [doi: 10.1037/a0034747] [Medline: 25000269]
23. Wang L, Miller LC. Just-in-the-moment adaptive interventions (JITAI): a meta-analytical review. Health Commun 2020 Nov 05;35(12):1531-1544. [doi: 10.1080/10410236.2019.1652388] [Medline: 31488002]
24. Goldstein SP, Evans BC, Flack D, Juarascio A, Manasse S, Zhang F, et al. Return of the JITAI: applying a just-in-time adaptive intervention framework to the development of m-Health solutions for addictive behaviors. Int J Behav Med 2017 Oct;24(5):673-682 [FREE Full text] [doi: 10.1007/s12529-016-9627-y] [Medline: 28083725]
25. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
26. Sterne JAC, Savović J, Page MJ, Elbers RG, Blencowe NS, Boutron I, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ 2019 Aug 28;366:l4898. [doi: 10.1136/bmj.l4898] [Medline: 31462531]
27. Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ 2003 Oct 06;327(7414):557-560 [FREE Full text] [doi: 10.1136/bmj.327.7414.557] [Medline: 12958120]
28. Egger ST, Vetter S, Weniger G, Vandeleur C, Seifritz E, Müller M. The Use of the Health of the Nation Outcome Scales for Assessing Functional Change in Treatment Outcome Monitoring of Patients with Chronic Schizophrenia. Front Public Health 2016;4:220 [FREE Full text] [doi: 10.3389/fpubh.2016.00220] [Medline: 27790607]

29. Duval S, Tweedie R. Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. Biometrics 2000 Jul;56(2):455-463. [doi: 10.1111/j.0006-341x.2000.00455.x] [Medline: 10877304]
30. Pham Q, Khatib Y, Stansfeld S, Fox S, Green T. Feasibility and Efficacy of an mHealth Game for Managing Anxiety: "Flowy" Randomized Controlled Pilot Trial and Design Evaluation. Games Health J 2016 Mar;5(1):50-67. [doi: 10.1089/g4h.2015.0033] [Medline: 26536488]
31. Boettcher J, Magnusson K, Marklund A, Berglund E, Blomdahl R, Braun U, et al. Adding a smartphone app to internet-based self-help for social anxiety: A randomized controlled trial. Computers in Human Behavior 2018 Oct;87(6):98-108. [doi: 10.1016/j.chb.2018.04.052] [Medline: 2018]
32. Greer JA, Jacobs J, Pensak N, MacDonald JJ, Fuh C, Perez GK, et al. Randomized Trial of a Tailored Cognitive-Behavioral Therapy Mobile Application for Anxiety in Patients with Incurable Cancer. Oncologist 2019 Aug;24(8):1111-1120 [FREE Full text] [doi: 10.1634/theoncologist.2018-0536] [Medline: 30683710]
33. Ponzo S, Morelli D, Kawadler JM, Hemmings NR, Bird G, Plans D. Efficacy of the Digital Therapeutic Mobile App BioBase to Reduce Stress and Improve Mental Well-Being Among University Students: Randomized Controlled Trial. JMIR Mhealth Uhealth 2020 Apr 06;8(4):e17767 [FREE Full text] [doi: 10.2196/17767] [Medline: 31926063]
34. Stiles-Shields C, Montague E, Kwasny MJ, Mohr DC. Behavioral and cognitive intervention strategies delivered via coached apps for depression: Pilot trial. Psychol Serv 2019 May;16(2):233-238 [FREE Full text] [doi: 10.1037/ser0000261] [Medline: 30407055]
35. Tighe J, Shand F, Ridani R, Mackinnon A, De La Mata N, Christensen H. Ibobbly mobile health intervention for suicide prevention in Australian Indigenous youth: a pilot randomised controlled trial. BMJ Open 2017 Jan 27;7(1):e013518 [FREE Full text] [doi: 10.1136/bmjopen-2016-013518] [Medline: 28132007]
36. Lüdtke T, Westermann S, Pult LK, Schneider BC, Pfuhl G, Moritz S. Evaluation of a brief unguided psychological online intervention for depression: A controlled trial including exploratory moderator analyses. Internet Interventions 2018 Sep;13:73-81. [doi: 10.1016/J.INVENT.2018.06.004]
37. Mantani A, Kato T, Furukawa TA, Horikoshi M, Imai H, Hiroe T, et al. Smartphone Cognitive Behavioral Therapy as an Adjunct to Pharmacotherapy for Refractory Depression: Randomized Controlled Trial. J Med Internet Res 2017 Nov 03;19(11):e373 [FREE Full text] [doi: 10.2196/jmir.8602] [Medline: 29101095]
38. Dahne J, Lejuez CW, Diaz VA, Player MS, Kustanowitz J, Felton JW, et al. Pilot Randomized Trial of a Self-Help Behavioral Activation Mobile App for Utilization in Primary Care. Behav Ther 2019 Jul;50(4):817-827 [FREE Full text] [doi: 10.1016/j.beth.2018.12.003] [Medline: 31208690]
39. Harrer M, Adam SH, Fleischmann RJ, Baumeister H, Auerbach R, Bruffaerts R, et al. Effectiveness of an Internet- and App-Based Intervention for College Students With Elevated Stress: Randomized Controlled Trial. J Med Internet Res 2018 Apr 23;20(4):e136 [FREE Full text] [doi: 10.2196/jmir.9293] [Medline: 29685870]
40. Roepke AM, Jaffee SR, Riffle OM, McGonigal J, Broome R, Maxwell B. Randomized Controlled Trial of SuperBetter, a Smartphone-Based/Internet-Based Self-Help Tool to Reduce Depressive Symptoms. Games Health J 2015 Jul;4(3):235-246. [doi: 10.1089/g4h.2014.0046] [Medline: 26182069]
41. Stolz T, Schulz A, Krieger T, Vincent A, Urech A, Moser C, et al. A mobile app for social anxiety disorder: A three-arm randomized controlled trial comparing mobile and PC-based guided self-help interventions. J Consult Clin Psychol 2018 Jun;86(6):493-504. [doi: 10.1037/ccp0000301] [Medline: 29781648]
42. Teng M, Hou Y, Chang S, Cheng H. Home-delivered attention bias modification training via smartphone to improve attention control in sub-clinical generalized anxiety disorder: A randomized, controlled multi-session experiment. J Affect Disord 2019 Mar 01;246:444-451. [doi: 10.1016/j.jad.2018.12.118] [Medline: 30599367]
43. Ly KH, Topooco N, Cederlund H, Wallin A, Bergström J, Molander O, et al. Smartphone-Supported versus Full Behavioural Activation for Depression: A Randomised Controlled Trial. PLoS One 2015;10(5):e0126559 [FREE Full text] [doi: 10.1371/journal.pone.0126559] [Medline: 26010890]
44. Ham K, Chin S, Suh YJ, Rhee M, Yu E, Lee HJ, et al. Preliminary Results From a Randomized Controlled Study for an App-Based Cognitive Behavioral Therapy Program for Depression and Anxiety in Cancer Patients. Front Psychol 2019;10:1592 [FREE Full text] [doi: 10.3389/fpsyg.2019.01592] [Medline: 31402881]
45. United States Census Bureau. Computer and Internet Use in the United States: 2018. 2021. URL: https://www.census.gov/newsroom/press-releases/2021/computer-internet-use.html [accessed 2022-07-05]
46. Chan S, Torous J, Hinton L, Yellowlees P. Towards a Framework for Evaluating Mobile Mental Health Apps. Telemed J E Health 2015 Dec;21(12):1038-1041. [doi: 10.1089/tmj.2015.0002] [Medline: 26171663]
47. Li Y, Guo Y, Hong YA, Zhu M, Zeng C, Qiao J, et al. Mechanisms and Effects of a WeChat-Based Intervention on Suicide Among People Living With HIV and Depression: Path Model Analysis of a Randomized Controlled Trial. J Med Internet Res 2019 Nov 27;21(11):e14729 [FREE Full text] [doi: 10.2196/14729] [Medline: 31774411]
48. Moberg C, Niles A, Beermann D. Guided Self-Help Works: Randomized Waitlist Controlled Trial of Pacifica, a Mobile App Integrating Cognitive Behavioral Therapy and Mindfulness for Stress, Anxiety, and Depression. J Med Internet Res 2019 Jun 08;21(6):e12556 [FREE Full text] [doi: 10.2196/12556] [Medline: 31199319]

49. Arean PA, Hallgren KA, Jordan JT, Gazzaley A, Atkins DC, Heagerty PJ, et al. The Use and Effectiveness of Mobile Apps for Depression: Results From a Fully Remote Clinical Trial. J Med Internet Res 2016 Dec 20;18(12):e330 [FREE Full text] [doi: 10.2196/jmir.6482] [Medline: 27998876]
50. Agarwal S, LeFevre AE, Lee J, L'Engle K, Mehl G, Sinha C, WHO mHealth Technical Evidence Review Group. Guidelines for reporting of health interventions using mobile phones: mobile health (mHealth) evidence reporting and assessment (mERA) checklist. BMJ 2016 Mar 17;352:i1174. [doi: 10.1136/bmj.i1174] [Medline: 26988021]
51. Musiat P, Tarrier N. Collateral outcomes in e-mental health: a systematic review of the evidence for added benefits of computerized cognitive behavior therapy interventions for mental health. Psychol. Med 2014 Feb 19;44(15):3137-3150. [doi: 10.1017/s0033291714000245]
52. Xiong S, Berkhouse H, Schooler M, Pu W, Sun A, Gong E, et al. Effectiveness of mHealth Interventions in Improving Medication Adherence Among People with Hypertension: a Systematic Review. Curr Hypertens Rep 2018 Aug 07;20(10):86. [doi: 10.1007/s11906-018-0886-7] [Medline: 30088110]
53. Evans W, Nielsen PE, Szekely DR, Bihm JW, Murray EA, Snider J, et al. Dose-response effects of the text4baby mobile health program: randomized controlled trial. JMIR Mhealth Uhealth 2015 Jan 28;3(1):e12 [FREE Full text] [doi: 10.2196/mhealth.3909] [Medline: 25630361]

Abbreviations

ACT: acceptance and commitment therapy
BAI: Beck Anxiety Inventory
BDI: Beck Depression Inventory
CBT: cognitive behavioral therapy
CES-D: Center for Epidemiological Studies Depression questionnaire
GAD-7: Generalized Anxiety Disorder-7
HAM-A: Hamilton Anxiety Rating Scale
JITAI: just-in-time adaptive intervention
LSAS-SR: Liebowitz Social Anxiety Scale-Self Report
mHealth: mobile health
PHQ-9: Patient Health Questionnaire-9
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RCT: randomized controlled trial
SIAS: Social Interaction Anxiety Scale
STAI: State-Trait Anxiety Inventory

Edited by J Torous; submitted 10.05.22; peer-reviewed by E Motrico, S Stewart-Brown, K Baker; comments to author 29.06.22; revised version received 07.08.22; accepted 11.08.22; published 07.09.22.

Please cite as:
Lu SC, Xu M, Wang M, Hardi A, Cheng AL, Chang SH, Yen PY
Effectiveness and Minimum Effective Dose of App-Based Mobile Health Interventions for Anxiety and Depression Symptom Reduction: Systematic Review and Meta-Analysis
JMIR Ment Health 2022;9(9):e39454
URL: https://mental.jmir.org/2022/9/e39454
doi:10.2196/39454
PMID:36069841

Review

The Apple Watch for Monitoring Mental Health–Related Physiological Symptoms: Literature Review

Gough Yumu Lui[1], BEng, PhD; Dervla Loughnane[2], BSc, MSc; Caitlin Polley[3], BEng; Titus Jayarathna[1], BSc, PhD; Paul P Breen[1,4], BEng, PhD

[1]The MARCS Institute for Brain, Behaviour and Development, Western Sydney University, Penrith, NSW, Australia

[2]Virtual Psychologist, Southport Park, QLD, Australia

[3]Electrical and Electronic Engineering, School of Engineering, Design and Built Environment, Western Sydney University, Penrith, NSW, Australia

[4]Translational Health Research Institute, Western Sydney University, Penrith, NSW, Australia

Corresponding Author:
Gough Yumu Lui, BEng, PhD
The MARCS Institute for Brain, Behaviour and Development
Western Sydney University
Locked Bag 1797
Penrith, NSW, 2751
Australia
Phone: 61 298525222
Email: G.Lui@westernsydney.edu.au

Abstract

Background: An anticipated surge in mental health service demand related to COVID-19 has motivated the use of novel methods of care to meet demand, given workforce limitations. Digital health technologies in the form of self-tracking technology have been identified as a potential avenue, provided sufficient evidence exists to support their effectiveness in mental health contexts.

Objective: This literature review aims to identify current and potential physiological or physiologically related monitoring capabilities of the Apple Watch relevant to mental health monitoring and examine the accuracy and validation status of these measures and their implications for mental health treatment.

Methods: A literature review was conducted from June 2021 to July 2021 of both published and gray literature pertaining to the Apple Watch, mental health, and physiology. The literature review identified studies validating the sensor capabilities of the Apple Watch.

Results: A total of 5583 paper titles were identified, with 115 (2.06%) reviewed in full. Of these 115 papers, 19 (16.5%) were related to Apple Watch validation or comparison studies. Most studies showed that the Apple Watch could measure heart rate acceptably with increased errors in case of movement. Accurate energy expenditure measurements are difficult for most wearables, with the Apple Watch generally providing the best results compared with peers, despite overestimation. Heart rate variability measurements were found to have gaps in data but were able to detect mild mental stress. Activity monitoring with step counting showed good agreement, although wheelchair use was found to be prone to overestimation and poor performance on overground tasks. Atrial fibrillation detection showed mixed results, in part because of a high inconclusive result rate, but may be useful for ongoing monitoring. No studies recorded validation of the Sleep app feature; however, accelerometer-based sleep monitoring showed high accuracy and sensitivity in detecting sleep.

Conclusions: The results are encouraging regarding the application of the Apple Watch in mental health, particularly as heart rate variability is a key indicator of changes in both physical and emotional states. Particular benefits may be derived through avoidance of recall bias and collection of supporting ecological context data. However, a lack of methodologically robust and replicated evidence of user benefit, a supportive health economic analysis, and concerns about personal health information remain key factors that must be addressed to enable broader uptake.

(JMIR Ment Health 2022;9(9):e37354) doi:10.2196/37354

KEYWORDS

Apple Watch; data; validation; mental health; psychology; precision medicine; heart rate variability; energy expenditure; sleep tracking; digital health; mobile phone

Introduction

Background

The COVID-19 pandemic has caused disruptions to the way people go about their daily lives. From the changing nature of work and employment, economic factors, the isolation brought about by stay-at-home orders, and the uncertainty of ever-changing health advice and medical directives, it is anticipated that these stresses will lead to an increase in mental health service demand beyond the current capacity [1]. The adoption of digital health technologies can potentially alleviate this burden.

Wearable devices are electronic sensors that are designed to be placed onto, or near to, the skin to measure signals from the body. Such devices can include wrist-worn devices similar to a watch or wristband, which can pair wirelessly with a mobile phone. Such devices have become a popular behavioral intervention for monitoring physiological activity to promote a healthy lifestyle [2]. Early forms of health monitoring include pedometers that would track daily steps and derive basic energy expenditure (EE) [3]. The potential of wearable devices for the monitoring of health has become particularly attractive to health care innovators seeking to enable new models of telehealth. However, these devices monitor physiological signals or physiologically related proxies (such as physical activity) of the user rather than mental health. Such devices may take the form of fitness trackers, which are typically simpler, lower-cost, and fixed-function devices with limited capabilities. Such devices often cannot support third-party apps, have limited user interactivity, and focus on fitness monitoring as their primary goal. By contrast, smartwatches are usually higher-end devices with a richer mix of sensors and user interfaces and a flexible, extensible software architecture permitting third-party software access and extended features such as voice calling, media control, and messaging. As the market matures, there are some products that may blur the lines; however, it is the richer suite of sensors, user interfaces, and support for third-party apps and data access, which makes these devices attractive for mental health research and monitoring purposes.

Mental health can be defined as “a state of wellbeing in which the individual realizes his or her own abilities, can cope with the normal stresses of life, can work productively and fruitfully, and is able to make a contribution to his or her community” [4]. This state is intimately connected with physical health and forms an integral part of general or overall health [5]. A mediation study examined the effects of physical health on mental health and vice versa, finding significant direct and indirect effects and cross-effects [6]. Studies have also indicated the effectiveness of physical activity in improving anxiety and depressive symptoms [7]. The measurement of signals from wearable devices that allow for an understanding of physical activity may also allow mental health status to be inferred.

Motivation

Apple Inc has emerged as an industry leader in health technology and wellness tracking devices [8]. The Apple Watch, first introduced in 2015, has retained the largest market share since its introduction and has continually advanced the capabilities of smartwatches [9]. These devices are primarily intended as wellness tools, garnering additional personal health monitoring for the wearer, typically for physiological activities such as heart rate (HR), HR variability (HRV), respiration rate, and physiologically related measures such as EE and fall detection. Some capabilities of these devices, such as the electrocardiogram (ECG) function, including a supporting app, have received Food and Drug Administration (FDA) clearance [10], whereas other aspects of their sensors and app capabilities have not yet been independently validated or received regulatory clearances. Monitoring of stress using these devices has been less studied but appears to be a promising avenue for application, particularly in the mental health sphere.

As digital health provides a novel model of care through the use of intelligent data, computing, and telecommunications, it holds promise for meeting the challenges of increased mental health demands. It can also enable *precision medicine*, which provides treatments bespoke to the patient’s needs [11]. There is interest in digital health across a number of industry sectors, including health care providers, insurers, and businesses [12-17], that may desire access to information on personal health through wearable devices such as the Apple Watch.

Wider adoption of devices for mental health monitoring is, in part, hampered by a lack of clarity regarding the devices’ capabilities, the accuracy and validity of the data that are collected, and their applicability to mental health monitoring and diagnosis [18-20]. This research aimed to fill this knowledge gap by examining the embedded sensor capabilities within the Apple Watch range, the physiological and physiologically related metrics recorded and made available for analysis, the validation status of these metrics within the literature, the connections (where they exist) between relevant health conditions associated with each metric, and implications for treatment. This analysis was performed both in a “top-down” approach focusing on reviewing published literature regarding the Apple Watch and a “bottom-up” approach focusing on the hardware and software capabilities of the Apple Watch to identify both currently available features and potential features that could be operationalized through the creation of customized apps using the Apple WatchKit, CareKit, and ResearchKit frameworks.

Methods

The literature review was conducted from June 2021 to July 2021.

Types of Studies and Materials

Various types of published studies and editorials were included. The types of studies were extended to some unpublished (gray) literature that was evaluated and reviewed for its suitability to close gaps in knowledge. Other gray literature sources included developer documentation for the HealthKit application programming interface for storing and managing data collected on the devices. Several opinion pieces were reviewed contextually to further provide a professionally informed perspective or illustrate further points of consideration. This literature review was structured to include the literature

concerning the monitoring of physical conditions that may present with psychological stressors and the implementation of the Apple Watch for such monitoring.

Search Strategy

The electronic databases selected for this literature review were PubMed, Scopus, and Google Scholar. A list of secondary keywords (Textbox 1) was developed with an emphasis on "Apple Watch" and truncated keywords combined using Boolean operators. Publication dates were restricted to 2015 onward, coinciding with the announcement of the first Apple Watch. Other recent literature that included wearable devices and novel developments to monitor or detect depression, anxiety, or stress was also included in the search process, in addition to reviews and systematic reviews.

Textbox 1. Literature review secondary search terms.

Secondary search terms

- "anxiety"
- "atrial fibrillation"
- "collection"
- "data"
- "depression"
- "digital health"
- "heart rate*"
- "insomnia"
- "mHealth"
- "monitor*"
- "oximet*"
- "physiology*"
- "psychology*"
- "remote"
- "respiration rate"
- "sens*"
- "sleep"
- "sleep apn*"
- "stress"
- "telehealth"
- "validat*"
- "wearable"

Selection Process

Published literature was included based on its use of the Apple Watch for either physiological data validation or psychology or mental health studies. Areas of interest for applications in monitoring physiological stress and mental health included HR monitoring, sleep tracking, respiration monitoring, and EE. Other inclusion criteria included studies performed on the suitability of wearable devices for monitoring physiological stress and their impacts on mental health. Only publications in English were included in the review. Screening was performed by a primary researcher and reviewed by other authors. Duplicate studies were removed.

Data Collection Process

Data extraction was performed using a spreadsheet that synthesized the findings and grouped the studies. Data management was achieved using EndNote (Clarivate Analytics) as the bibliographic management software. Where studies did not specify the Apple Watch Series, it was inferred by comparing the date of publication with the Apple Watch Series release dates.

Results

Literature Review

The literature search strategy resulted in 5583 paper titles being identified. Screening of titles and abstracts resulted in 2.06% (115/5583) of papers being selected and reviewed in full. Of these 115 papers, 19 (16.5%) were identified as related to Apple Watch validation or comparison studies, which are summarized in Table 1.

Table 1. Summary of Apple Watch validation studies (N=19).

Study	Study focus	Outcome
Binsch et al [21], 2016	Resilience and workload monitoring	• PPG[a] reliable in the at-rest condition; wide-ranging outcomes during movement • Apple Watch showed the most variance in steps and distances compared with ground truth measurements, followed by the comparison, Fitbit Surge and Microsoft Band • Such variances are surmised to be because of differences in data resolution and access and underlying algorithms using accelerometer and GPS data for step count estimation
Shcherbina et al [22], 2017	HR[b] and EE[c]	• Lowest error in HR and EE for cycling; highest error for walking • Apple Watch achieved the lowest overall error in HR and EE of the tested devices (Basis Peak, Fitbit Surge, Microsoft Band, Mio Alpha 2, PulseOn, and Samsung Gear S2)
Dooley et al [23], 2017	HR and EE	• Apple Watch HR mean absolute percentage error was between 1.14% and 6.70%, not significantly different during baseline and vigorous-intensity treadmill exercise; lower HR in light- or moderate-intensity treadmill exercise and recovery • EE mean absolute percentage error was between 14.07% and 210.84%, measuring higher EE in all states compared with the criterion measure (Parvo Medics TrueOne 2400), with greater errors for higher BMI and the male population • HR and EE results were mostly better than other tested devices (Fitbit Charge HR and Garmin Forerunner 225)
Wang et al [24], 2017	HR	• Apple Watch had 95% differences between −27 bpm[d] and +29 bpm; concordance correlation coefficient was 0.91; accuracy diminished with exercise.
Hernando et al [25], 2018	HRV[e]	• Apple Watch RR interval data were found to contain gaps lasting 6.5 seconds per gap, averaging 5 gaps per recording, not correlated with stress or relaxation case • The cause is surmised to be because of failure to detect reliable pulses from PPG data • Temporal HRV indices were not significantly affected, but frequency-based LF[f] and HF[g] power showed a significant decrease • Apple Watch was able to successfully detect mild mental stress
Abt et al [26], 2018	Moderate-intensity exercise	• Apple Watch threshold for moderate-intensity exercise was lower than the defined criterion of 40% to 59% VO2R[h], leading to overestimation of moderate-intensity exercise minutes
Abt et al [27], 2018	Maximal HR	• Apple Watch had good to very good criterion validity for measuring maximal HR with no substantial under- or overestimation • Moderate and small errors were found for simultaneous recording from left versus right watches
Roomkham et al [28], 2019	Sleep monitoring	• Apple Watch raw acceleration data were used to compute ENMO[i] for classification • Apple Watch had high accuracy (97.3%) and sensitivity (99.1%) in detecting sleep and adequate specificity (75.8%) in detecting wakefulness
Perez et al [29], 2019	AF[j]	• Apple Watch irregular rhythm notification was triggered on 0.52% of 419,297 participants • Of those who returned an ECG[k] patch, 84% of subsequent notifications were confirmed to be AF • A total of 34% of ECG patches returned identified AF in part because of the transient nature, suggesting that Apple Watch may be useful for ongoing monitoring
Nuss et al [30], 2019	EE	• Apple Watch overestimated EE in women and underestimated EE in men • Pooled relative error was 24.3%, 18.6% for men, and 19.9% for women • Neither device showed accurate results compared with EE measured with a MetCart
Thomson et al [31], 2019	HR	• ECG correlation was strongest for very light intensity with a >0.90 concordance correlation coefficient • Most relative error rates were <5% with a maximum of 5.73% • Apple Watch was more accurate in recording HR than the Fitbit Charge HR 2

Study	Study focus	Outcome
Nelson and Allen [32], 2019	HR and passive monitoring	• Apple Watch 3 was generally accurate across a 24-hour period compared with ECG; the mean difference was −1.8 bpm, the mean absolute error was 5.86%, and the mean agreement was 95% • Apple Watch was more accurate than Fitbit Charge 2
Falter et al [33], 2019	HR and EE in patients with cardiovascular disease	• Apple Watch showed good correlation without systematic error comparing Apple Watch PPG HR with ECG ground truth • Apple Watch showed a systematic overestimation of EE compared with indirect calorimetry • Apple Watch HR accuracy was clinically acceptable
Düking et al [34], 2020	HR and EE	• Apple Watch 4 showed the highest validity in measuring HR, followed by Polar Vantage V, Garmin Fenix 5, and Fitbit Versa • The coefficient of variation for HR was 0.9% to 4.3% and, for EE, it was 13.5% to 27.1%
Espinosa et al [35], 2020	Step counting and HR	• The walking error was 2.6%; jogging error was 5.1% • HR limit of agreement was −2.2 to 1.8 bpm for walking and −3.5 to 4.3 bpm for jogging • Apple Watch displayed a high level of agreement and was highly accurate
Seshadri et al [36], 2020	HR in patients with AF	• Patients with AF showed a correlation coefficient of 0.7 between Apple Watch 4 and telemetry • Apple Watch 4 HR was more accurate for patients in the AF condition than for those not in the AF condition • Caution suggested in Apple Watch HR monitoring in patients with arrhythmia
Seshadri et al [37], 2020	AF	• Apple Watch 4 notification correctly identified AF in 34 of 90 instances (41% sensitivity), with no false positives and 31% inconclusive • The agreement between Apple Watch 4 and telemetry was 61% • Apple Watch–exported ECG PDF files showed AF in 84 of 90 instances (96% sensitivity), no false positives, and 2 failures to generate PDFs • Agreement between Apple Watch 4 ECG PDFs and telemetry was 98.9% • Further validation is required because of the high inconclusive result rate
Glasheen et al [38], 2021	Wheelchair use	• Apple Watch 1 only showed good agreement on higher-rate fixed-frequency tasks, with significant overestimation at low frequency • Arm ergometry showed good agreement across all cadences • Overground tasks showed poor agreement, with significant differences found
Huynh et al [39], 2021	HR in patients with obstructive sleep apnea and AF	• Apple Watch 1 variability increased as the magnitude of the HR measurement increased • The Lin concordance correlation coefficient was 0.88, suggesting acceptable agreement between Apple Watch 1 and telemetry

[a]PPG: photoplethysmography.

[b]HR: heart rate.

[c]EE: energy expenditure.

[d]bpm: beats per minute.

[e]HRV: heart rate variability.

[f]LF: low-frequency.

[g]HF: high-frequency.

[h]VO2R: reserve oxygen consumption.

[i]ENMO: Euclidean norm minus one.

[j]AF: atrial fibrillation.

[k]ECG: electrocardiogram.

Several published reviews focusing on wearable devices, smartwatches, and associated physiological measurements were also identified as part of this search (Textbox 2). These reviews provide a contextual background in a number of areas; however, this review was focused on Apple Watch–specific research.

Textbox 2. Wearable device reviews identified.

Authors and review focus

- Lu et al [40], 2016: health care applications
- Reeder and David [41], 2016: health and wellness
- Kim et al [42], 2018: stress and heart rate variability
- Jo et al [2], 2019: patient benefits from wearable devices
- Shin et al [43], 2019: accuracy, adoption, acceptance, and health impact
- Attig and Franke [44], 2020: reasons for abandonment of personal tracking
- Guillodo et al [45], 2020: clinical applications of wearable-based sleep monitoring
- O'Driscoll et al [46], 2020: accuracy of energy expenditure monitoring
- Hickey et al [47], 2021: detect and monitor mental health conditions and stress

HR and HRV

Overview

Across the Apple Watch Series, there are several mechanisms for detecting and monitoring HR metrics. At a minimum, all Apple Watch Series use photoplethysmography (PPG) optical HR sensors to detect either low or high HR and irregular rhythm. In the newer model Apple Watch, there is the option for additional sensors to record ECG. Therefore, Apple Watch users have access to 2 independent measurements of HR through separate apps that can serve similar functions to medical devices [48].

Traditionally, clinical HR and cardiac assessments are performed with 12-lead ECG recordings; however, this is unsuitable for continuous monitoring applications. Wearable devices generally use PPG- and ECG-based sensors, which can be more easily integrated but provide less information. Irregular HR notifications check for events that show irregular rhythm that "may be suggestive of AF" [49]. In Apple Watch Series 1 onward, notifications can be derived from PPG-based tachograms captured opportunistically at irregular times during the day and subsequently classified using an algorithm [50]. In the event that irregular heart activity is detected within the ECG version 2 app, the Apple Watch (Series 4 onward) classifies the ECG recorded event as either atrial fibrillation (AF), sinus rhythm, high or low HR, or inconclusive or declares a poor reading.

The Apple Heart Study, conducted from November 2017 to August 2018, assessed 419,093 enrolled participants via PPG recordings to determine the presence of previously undiagnosed AF [29,50,51]. If an AF event was detected with a duration of >30 seconds, the patient was offered a telemedicine consultation and ePatch ambulatory ECG patch for confirmatory monitoring over a period of up to 7 days. The study noted that of the participants who had been notified by the Apple Watch of the presence of AF, only 34% had subsequent ECG recordings conducted via mailed ECG patches [29]. However, 84% of the app-detected AF notifications were concordant with subsequent clinical AF diagnoses [29].

A pilot validation study monitoring HR via PPG to detect the presence of AF in patients with obstructive sleep apnea found an agreement between the Apple Watch HR–declared events and GE Healthcare CARESCAPE Monitor B650 telemetry [39]. The findings concluded that 95% of the HR readings made by the Apple Watch Series 1 measured within 19 beats per minute (bpm) of telemetry with a Lin concordance correlation coefficient of 0.88 and a mean bias of 0.26 bpm. These values were considered acceptable but relatively wide. Another study used the Apple Watch Series 1 to detect clinical correlations between HR during subacute periods in patients recovering from acute myocardial infarction [52]. HR recordings were taken 4 times per day during a 30-day postdischarge period. Healthy patients showed a decline in average daily HR of 0.2 bpm per day compared with patients with prior coronary artery bypass surgery showing an increasing HR trend of 0.1 bpm per day and those with hypertension and type 2 diabetes mellitus showing a slower HR decline.

A study by Shcherbina et al [22] compared the Apple Watch (presumed to be Series 1) with other commercially available wrist-worn devices. It found that the Apple Watch using the Apple Health app was able to provide HR, EE, and step counts sampled at 1-minute intervals or more frequently if higher-intensity exercise was detected or declared by a workout routine [22]. All other commercially available wrist-worn devices in this study, including the Basis Peak, Fitbit Surge, Microsoft Band, PulseOn, and Samsung Gear S2, only had granularity down to 1 minute. Across all modes of activities, the Apple Watch achieved the lowest error of all tested devices, averaging a 2% error in HR. This was echoed in another 11% (2/19) of studies comparing the accuracy of HR within Apple Watch devices with other commercially available devices relative to traditional ECG [23,33].

Derived from HR is HRV, another measurement of cardiac performance indicating the variation in time between heartbeats (NN or RR interval) in either the time or frequency domain. It is a method for monitoring cardiac health, sleep quality, mental stress, chronic pain, posttraumatic stress disorder, bipolar disorder, and traumatic brain injury [53,54]. There are a number of statistical methods to calculate HRV, including the SD of NN intervals (SDNN), the HRV triangular index, the SD of the average NN intervals, and the root mean square of successive differences [55,56]. The Apple Watch provides HRV as the SD of the beat-to-beat intervals (SDNN) [57]. Although HRV can

be calculated from ECG, in the case of the Apple Watch, it is calculated using the optical HR sensors and can be accessed within HealthKit on a paired iPhone device.

Dalmeida et al [58] looked at HRV features in the time domain and the high- and low-frequency domains to determine the most ideal metric by implementing a machine learning algorithm. They concluded that SDNN, as used by Apple Watch, was acceptable among other methods for calculating HRV [58]. The Apple Watch data used with the developed web application for this study predicted stress states with 71% probability and relaxation states with 79% probability. Another validation study by Hernando et al [25] investigated the impacts of various HRV statistical models on both the time and frequency domains in both relaxed and stressed states and compared the various statistical methods for their accuracy. Approximately 10% of beats were missed, usually consecutively, with a greater number of missing beats in the stressed state and at the beginning of recordings. This is speculated to be because of poor skin contact or sudden movement; however, no empirical evidence is available because of the proprietary nature of the algorithms within the Apple Watch. Computed time domain HRV metrics were comparable with data from a Polar H7 chest belt, with frequency domain metrics showing differences because of the missed beats [25]. It was found that there was no significant difference in the effectiveness of time domain HRV methods and that SDNN was just as effective as other methods.

Applications in Mental Health

The potential of wearable devices for monitoring mental health and related physiological stressors lies in the prospective ability of users to interpret and understand their emotional awareness and emotional regulation or of this information to be collected and relayed to a caregiver or clinician for follow-up action.

Panic disorders commonly present with other mental health issues, for which monitoring can prove to be valuable. Panic attacks are specified as sudden or abrupt surges of involuntary arousal, increasing HR rapidly and subsiding within minutes, and are commonly preceded by cardiorespiratory instabilities [59]. These involuntary movements are controlled by the autonomic nervous system, which is part of the peripheral nervous system. The autonomic system comprises sympathetic and parasympathetic systems that have significant control over HR, HRV, blood pressure, respiration rate, and temperature [60]. In simple terms, sympathetic activity leads to arousal or "fight or flight" responses, whereas parasympathetic activation leads to more recovery activity. Research on the psychological significance of the imbalance between these 2 systems suggests that HRV could be used as a more ideal physiological measurement of stress compared with HR. Reduced HRV is seen in individuals with psychiatric disorders [61]. This is because low-frequency components of HRV indicate increased sympathetic activity, whereas high-frequency components are generated within the parasympathetic system. An imbalanced ratio between low- and high-frequency components suggests a greater presence of stressing stimuli [42,58]. These findings were also encouraged by a systematic review of wearable devices, which determined that HRV was "the most useful metric for detection of stress and anxiety" and that devices that combined accelerometers, ECG, and subjective questionnaires could assist in the diagnosis of depression [47].

Physiological data accuracy with regard to HR and HRV is generally viewed as favorable compared with other devices, especially in the at-rest condition, and is likely to provide valuable data for the needs of mental health monitoring applications.

EE Measurement

Overview

Another key tracking feature is step counting and the average or total calories burned through EE. A key feature of EE and movement tracking is the motivation provided by setting personal activity goals. The Workout app used for the Apple Watch assists in tracking progress updates and setting activity goals. Motivation goal setting can assist in weight management and overall health tracking and can be programmed within the Apple Watch [62]. Apple provides several apps that can be used with the Apple Watch to assist in health tracking and statistical data collection with the Workout and Activity apps. The Workout app includes a list of activities (Table 2), an automatic workout detection feature, a record of workout sessions (including start and end times), progress update tracking, and reminders to start routines. The Activity app is used to monitor general activity and movement throughout the day and is intended to encourage users to move, stand up, and exercise. Activity targets are displayed using dynamically closing rings, illustrating a clear overall goal [63]. Passive data such as HR, steps, distance, active minutes, and stand reminders are collected. The total EE calculated from the Apple Watch accelerometer was noted to improve with the inclusion of HR in the calculation algorithm [46,64]. As such, the Apple Watch continuously measures HR in the Workout app during exercise and for 3 minutes afterward to calculate a "recovery rate," which is further used to enhance the estimate of how many calories have been burned during the workout routine [48].

Wearable devices are typically able to determine the difference between low- and high-intensity activity but require improvement in resilience to changes in setting, particularly with an increase in exercise intensity, if more accurate absolute EE is to be extracted. Most validation studies that included the Apple Watch indicated an overestimation of total EE at different activity intensity levels [26,33,38,65]. However, 11% (2/19) of the studies noted an underestimation of total EE in the study group, and 5% (1/19) of the studies noted that the Apple Watch overestimated EE in female participants but underestimated it in male participants [30,64]. Despite the variation in the accuracy of EE estimation, the device could successfully distinguish activity intensity. This is summarized in a systematic review of activity trackers and total EE proficiency by O'Driscoll et al [46], which noted that devices exhibiting the largest EE error relied exclusively on accelerometer data.

At present, a range of activity types and intensities can be defined by the wearer (Table 2) [66]. This would enable the Apple Watch to generate an improved EE estimate [52]. Additional data, such as altimeter data to indicate changes in elevation, could further improve this estimate. Modifications

to the accuracy of algorithms for activity tracking and calorie counting can be improved with software updates and more nuanced user input; for example, watchOS 8 (released in September 2021) adds outdoor cycling detection, e-bike pairing for improved calorie calculations, and Pilates and tai chi workout types [66,67].

Table 2. Workout types for Apple Watch within the Workout app.

Activity type	Subtype	Notes
Walking	Indoor or outdoor	• Apple Watch Series 1 requires iPhone to calibrate pace and distance calculated from GPS (Apple Watch Series 2 onward) • Elevation from altimeter (Apple Watch Series 3 onward)
Running	Indoor or outdoor	• Option to use Bluetooth chest strap instead of integrated PPG[a] heart sensor to reduce motion artifacts
Cycling	Indoor or outdoor; e-bike or manual (watchOS 8)	• Speed and distance (Apple Watch Series 2 onward) and map elevation (Apple Watch Series 3 onward) • Automatic detection for start and stop (from watchOS 8)
Elliptical	Elliptical machine	N/A[b]
Rower	Rower machine	N/A
Stair stepper	Stepping machine	N/A
HIIT[c]	Intense exercise followed by short periods of rest (30-45 seconds)	• May affect HR[d] sensors • Calories tracked with accelerometer
Hiking	Tracks pace, distance, elevation gain, and calories burned	• Requires altimeter (Apple Watch Series 3 onward) or paired the phone with an altimeter
Yoga	All types of yoga	N/A
Functional strength training	Dynamic strength training with dumbbells, resistance bands, and medicine balls	N/A
Dance	All types of dance	N/A
Cooldown	Easy moves and stretches	N/A
Core training	Strength-building of abdominals and back	N/A
Swimming	Pool or open swim	• Set pool length; GPS is not used to conserve battery • Open swim requires GPS; may affect HR sensors
Wheelchair	Outdoor wheel-walk pace and outdoor wheel-run pace	• Apple Watch Series 2 onward uses GPS or paired iPhone with GPS for Apple Watch Series 1 • Measures time, pace, distance, calories, HR, and pushes
Other	Add a workout type	• HR and motion sensors work together to provide an accurate reading • Will display popular workouts from users

[a]PPG: photoplethysmography.

[b]N/A: not applicable.

[c]HIIT: high-intensity interval training.

[d]HR: heart rate.

Applications in Mental Health

Personal activity tracking and goal setting can lead to increased exercise, with physical and mental health benefits. The key components of mental health benefits can be seen in individualized means of self-reflectivity and mindfulness [15]. Tracking changes in activity and movement can be used as an indicator of health management, such as weight loss, but also as a key indicator of changes in mood stages (eg, low activity could indicate the presence of a depressive episode). A cross-sectional study investigated the effects of wearable trackers and how they make users feel and concluded that most users felt positive about tracking technology and that negative experiences were mostly confined to individuals with low conscientiousness or openness to experience [68]. Further investigation of wearable trackers and their psychological effects in younger demographics is recommended, as well as an examination of the effects in those who exhibit neuroticism and obsessive-compulsive traits [68].

There is some ambiguity regarding the level of accuracy that is acceptable for EE, as it depends on the context of the

application. For wellness applications, the absolute accuracy of EE may not be critical or align with the primary goal of the intervention. In this case, small inaccuracies may not be particularly significant for the user. Tracking of general movement patterns in combination with measures of HRV and respiratory rate variability may be sufficient for monitoring work-related stress, detecting episodes of mania, anxiety or depression, or sleep-related disorders (insomnia) [69,70]. Similarly, the detection of psychological distress through activity metrics appears viable [71]. However, more research is required to validate the capability of the Apple Watch to detect such episodes.

Sleep Monitoring

Overview

The introduction of watchOS 7 in June 2020 brought about integrated sleep monitoring to track the quality and duration of wearers' sleep for Apple Watch Series 3 and above. The watchOS 8 release in September 2021 improved this by also reporting sleeping respiratory rate [72]. As this is a relatively recently introduced feature, which is primarily promoted as a "wellness monitoring" feature, no literature was identified that tested or validated it. Sleep tracking through third-party apps is also available, some of which are more sophisticated and integrate HR measurements from PPG [73].

Roomkham et al [28] performed a 27-night sleep study with the Apple Watch Series 1 using raw data from its accelerometers at 50 Hz through Apple's Core Motion framework (independent from the watchOS Sleep app, which did not exist at the time) and compared the results with the Philips Actiwatch Spectrum PRO [28]. The overall patterns between the 2 devices demonstrated correlations of key movement events with 97.3% accuracy and 99.1% sensitivity in detecting sleep and a specificity of 75.8% for detecting wakefulness.

However, wrist-worn sleep monitors based on accelerometry are not without criticism, and there is some skepticism about the reliability of using wrist-worn devices for monitoring sleep to identify the depth of sleep and wake periods. Approximately 5% (1/19) of the studies looked into 3 devices—the Mi Band activity tracker, the MotionWatch 8, and the Sleep Cycle mobile phone app—to monitor sleep [74]. All devices reported high accuracy of time in bed but were incapable of accurately detecting sleep and wake periods and sleep efficiency. This study also found that each of the devices had unacceptable levels of agreement with polysomnography. This view was echoed in a systematic review of wearable devices for sleep monitoring, which stated that wearables generally have "acceptable sleep monitoring but with poor reliability" [45]. It is evident from these studies that using wrist-worn accelerometers as the sole sleep-monitoring sensor severely limits the ability to contextualize sleep patterns and behavior. As such, they are not capable of full-spectrum sleep monitoring but remain promising.

Applications in Mental Health

It is recognized that low quality of sleep may exacerbate physical and mental health problems and that sleep tracking can be used to improve user awareness of possible sleep problems [75]. The prevalence of insomnia and chronic sleep issues such as sleep apnea is increasing, with an estimate that 1 in 2 people experience bouts of sleep disturbances during their life, with negative impacts [39,45]. Sleep monitoring is also valuable for mental health monitoring, as a lack of sleep can be the cause of impaired performance, low energy levels, and problems with mood.

The literature indicates that most wearable devices with accelerometers have high sensitivity but low specificity for sleep detection [45]. Specific information about the quality of sleep would require other sensor data or could be inferred through patient-practitioner communication. However, there are practical concerns regarding battery use and when the device can be charged, as many users may prefer to charge their Apple Watch devices overnight [76]. Charging creates interruptions in monitoring, which could pose a challenge in accurately monitoring panic attacks, which usually occur unexpectedly [28,59,77]. Improvements in charging times have occurred with the announcement of Series 7, which includes the Apple Watch Magnetic Fast Charging USB-C cable that can charge to 80% battery capacity within 45 minutes, which may serve to minimize such interruptions [78]. Limitations in the accuracy and detail of sleep quality restrict clinical utility in cases of mood disorders, mania, anxiety or panic attacks, and sleep-wake disorders, which may require investigation in a specific sleep cycle. The interpretation of sleep data can be complicated by incorrect sleep detection (eg, while being still or watching television) [75]. However, in combination with other tools and strategies, general sleep monitoring and tracking can assist in developing and implementing behavior change techniques.

Discussion

Apple Watch Sensors

The Apple Watch is a sensor-rich, well-constructed, and connected device. It uses a large range of apps and has significant potential for applications in mental health (Figure 1).

Apple Watch sensors typically include a 3-axis accelerometer, a gyroscope and magnetometer, optical PPG-based HR sensors, altimeters, ambient light sensors, temperature sensors, ECG, and capacitive (touch) sensors [3]. Across each iteration of the Apple Watch, sensor inclusions and capabilities have increased, matched with software updates aimed at increasing the overall accuracy of the collected data. Figure 2 presents a timeline of the development of the Apple Watch, summarizing the changes in sensor inclusions over time. The latest version of watchOS (version 8.0.0) is supported by Series 3 to Series 7 models. The models currently available for purchase include Series 3, SE, and Series 7. The Apple Watch Series 3 does not include fall detection as the 6-axis inertial measurement unit containing the gyroscope and accelerometer was modified for later-generation Apple Watches [49].

Figure 1. Summary of Apple Watch sensors, apps, and potential mental health applications. ECG: electrocardiogram; GNSS: global navigation satellite system; LTE: Long-Term Evolution; PPG: photoplethysmography.

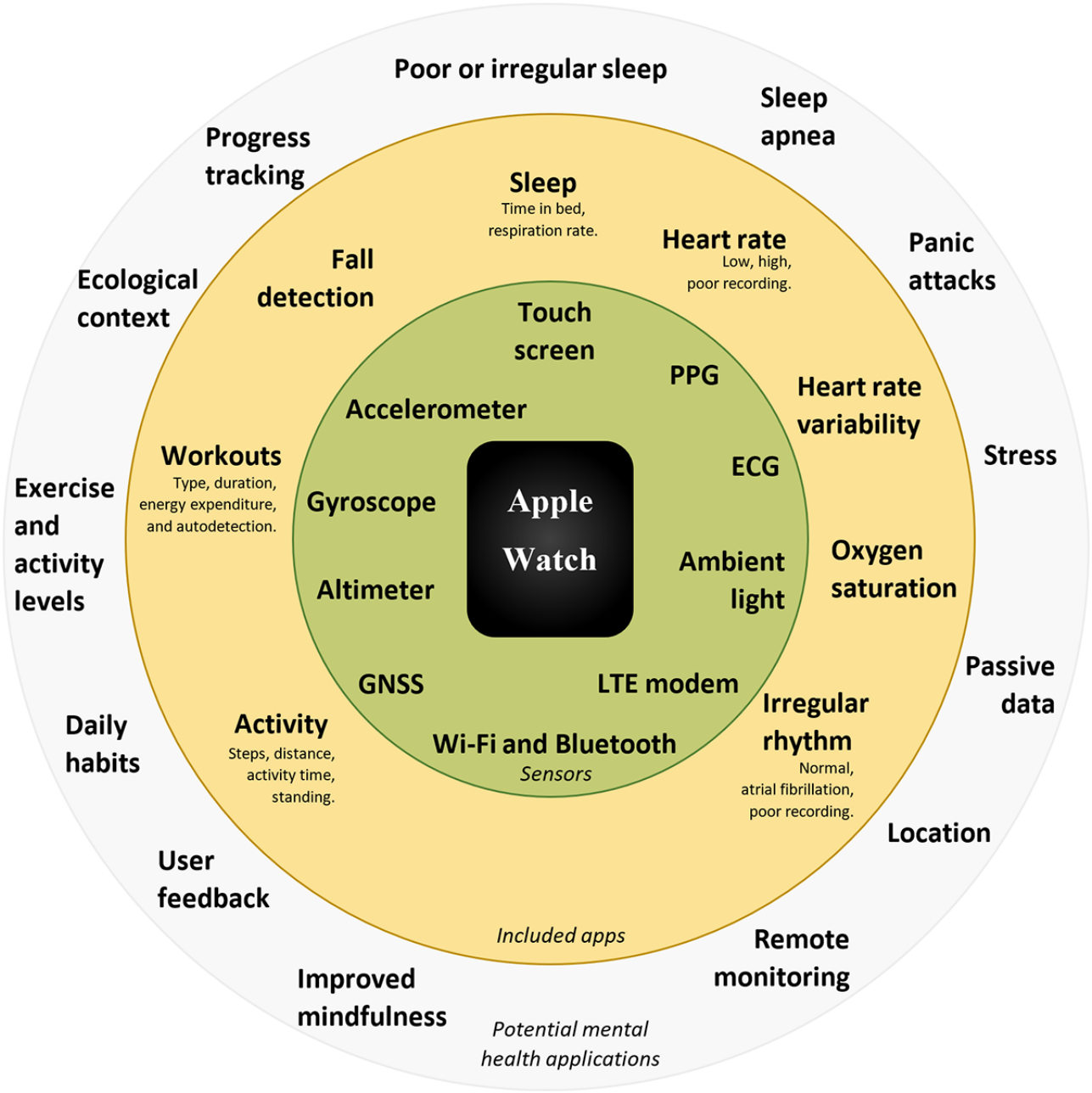


Figure 2. Evolution of Apple Watch Series features. Feature upgrades (↑) and new feature additions (+) are indicated. ECG: electrocardiogram; GNSS: global navigation satellite system; LTE: Long-Term Evolution; NFC: near-field communication; OLED: organic light-emitting diode; UWB: ultrawide band.

One of the primary sensors in all generations of the Apple Watch is the optical HR sensor, which is used to collect HR data. The scientific principle that these sensors rely on is PPG to detect the amount of blood that is flowing through the wearer's wrist at any given moment. The reflection of green and infrared light-emitting diode (LED) light is measured with photodiodes that allow for the determination of HR as a periodic variation in the signal. By flashing hundreds of times per second, the optical HR sensor can measure HR across a range of 30 to 210 bpm [48]. Infrared light is used to measure HR in the background and for HR notification systems as infrared light can penetrate the skin better; however, this makes it more susceptible to motion artifacts. Green LEDs are used for workouts and to calculate HRV [48]. The Apple Watch will

automatically detect when there is an increase or decrease in motion from the inertial measurement unit and change the LED light color accordingly. Variations have been made in the design and layout of the LED and photodiode arrays with each iteration of the Apple Watch to improve accuracy [79]. These optical HR sensors are used by the Irregular Rhythm Notification Feature (IRNF), which can assist in the detection of AF [80,81]. A red LED was added in Series 6, enabling blood oxygen saturation calculation by comparing the ratio of infrared light and red light. Reflectance oximetry is noted as being less accurate than clinically used transmittance oximetry [79], and we did not identify any literature validating the accuracy of the Apple Watch blood oximetry.

In addition to the optical HR sensors, from Series 4 onward (not including the SE model), an ECG electrode was integrated into the back face of the watch and the digital crown. When engaged by the user's finger, a closed circuit is created to measure the electrical potential across the heart, similar to a 1-lead ECG. An ECG measurement takes 30 seconds. The ECG sensor is exclusively used with the ECG classifier to categorize heart events as AF, normal sinus rhythm, high or low HR, or inconclusive [48,82]. Version 2 of the ECG app also includes additional classifications of AF, high HR, and poor recording. For the earlier Apple Watch Series, a third-party accessory (Kardia Band) could be used to provide a 1-lead ECG that achieved a sensitivity of 93% and a specificity of 84% when compared with a standard tachograph [83].

A clinical study compared the ECG app developed by Apple Inc with an FDA-cleared clinical ECG device (GE Healthcare CardioSoft ECG device), with recordings verified by 3 independent board-certified American cardiologists in each of the ECG app categories [50]. The app received clearance by the FDA as a De Novo class II device as it was proven to perform similarly to the comparator device [82]. The same approval was also given to the optical HR sensor IRNF software in 2018 [80]. Some limitations exist in the use of both apps, which are not intended to be used on persons aged <22 years. Depending on the country in which the Apple Watch user resides, they may not have access to the software and, as such, may not be able to use these notification features. In Australia, both the ECG app (version 2.0) and the IRNF software were approved by the Therapeutic Goods Administration of the Australian Government in early 2021 [84,85].

Further Considerations

Health data collected from the Apple Watch could complement smartphone data collection and self-reported measures to provide additional context and assist in determining and tracking a user's affective and emotional health. Advancements in the sensing technologies available within wearable devices and enhanced user interfaces have removed some of the previously limiting factors of monitoring mental health using wearable technology. However, the current general consensus for using wearable device sensors is that they should be paired with traditional screening and diagnostic tools and not be considered as a replacement [33,83]. Wearable devices can assist in clinical diagnosis and application of therapy if the findings are consistent with the patient's complaints or concerns or if the patient is unsure of their physiological level of stress [86]. Indeed, a systematic review of digital health interventions for depression and anxiety in young people has shown that such interventions may only be of clinical significance when their use is highly supervised [87].

An article compared several wearable devices, including the Apple Watch (series unspecified), and their applications for "advancing resilience and mental health of employees that experience high workload" [21]. The study noted that an increase in psychological disabilities in the modern workplace requires the development of new and emerging technologies to measure and monitor physical or mental status. As such, these tools are being implemented to assist in the diagnosis and treatment of stress within professional workplaces and in a performance review. A potential issue with workplace inclusion for monitoring mental health and wellness is regulations and access to technology.

The use of the Apple Watch as a source of data may address problems with patient recall bias as most assessments are reliant on patient self-reporting. This could reduce the reliance on patient memory and continued questioning to ensure consistency. In addition, it could be a relatively low-cost method for better long-term tracking of symptoms and trends in the data [69]. The use of these data permits the construction of an ecological context that could empower a more cohesive diagnosis and application of therapy or assist in refining threshold values used in algorithms toward a validated measure.

Although there are potentially great benefits of wearable devices in improving mental health, there are some potential drawbacks, including concerns about abandonment rates. Approximately 11% (2/19) of individual studies commented on the long-term use of electronic wearables, one noting that 20% of consumers stop using their wearables after 3 months, and <50% continue to use them after 1.5 years [83,88]. This is compounded by the need to provide enough contextual information regarding the data collected, which requires some level of active user participation. For a clinical diagnosis of a mental disorder, clinicians must make a decision based on weighing the mix of potentially contradictory evidence according to their expert judgment, which could require symptom tracking over a period of months to come to a clear conclusion. Symptom tracking for the validation of several mental health diagnoses against the Diagnostic and Statistical Manual of Mental Disorders can require the presence of symptoms over a period of weeks, months, or even years for mood disorders, anxiety disorders, and schizophrenia [59].

A validation study was completed on the effectiveness of using the Apple Watch to collect passive sensor data with "ecological momentary assessments" from a watch-based questionnaire app recording patient feedback to assess and monitor substance abuse in young adults [89]. The response from participants on the perceived burden of engaging with the app was low; however, it was noted that the relative ease of completing the surveys was easier on an iPhone than on the Apple Watch. Burdensome interactions within wearable devices could reduce uptake and willingness to use technology for mental health monitoring. However, the benefits of engaging users through

health notifications and alerts can assist in seeking medical assistance or outpatient care [29]. A longitudinal observational study using cognitive assessment delivered through the Apple Watch in patients with major depressive disorder noted excellent adherence for both mood and cognitive tests (95% and 96%, respectively) over the 6-week study period, and it was not influenced by symptom severity or cognitive function at the study onset and did not deteriorate over time, supporting the feasibility of this approach [90].

Health and Sensor Data Access

The availability of sensor and health data collected from the Apple Watch and patient input relies on the application programming interface frameworks available from Apple for iOS and watchOS. The main frameworks are HealthKit, ResearchKit, CareKit, and SensorKit [91-93]. HealthKit is the most comprehensive as it implements a central repository for all collected health data related to the user. Developers can write apps that request permission to access the HealthKit data store to record, access, and share user health data. SensorKit is used in the event that raw access to sensors is required. ResearchKit may be used to build research study apps, whereas the CareKit framework is suited to the development of ongoing care capabilities. Together, these frameworks allow for the implementation of apps that can collect raw data and store and analyze collected data (including passively collected data) and provide tracking feedback to the end user as well as the clinician.

Within the HealthKit framework, a range of rigid data classes and methods can be used to collect, store, and retrieve data. In this way, virtually all types of health-related data can be stored as numerical data (eg, HR) and categorical data objects (eg, blood type). It categorizes the data systematically, reducing duplication and allowing for straightforward statistical data analysis. HealthKit supports units of measurement within each of these categories such as length, mass, volume, and energy. Conversion between measurement systems is automatically supported within the framework but can also be explicitly defined. Developers cannot create custom data types or units but can use the metadata fields to store additional data.

Most of the identified studies investigating wearable devices collected the activity level (steps and caloric expenditure), HR, and sleep data without indicating how the data were collected from the device, the frequency of data recording, or which measures were extracted from HealthKit. We believe this to be important information to be provided by studies, especially those that develop a custom app, to ensure a comprehensive understanding of the data, allow for comparative analysis with other studies, and inform future developments.

Data Analysis and Digital Phenotyping Approaches

Digital phenotyping approaches have been an active area of development enabled by the popularity of smartphones [94]. By collecting data from sensors in a smartphone on a moment-by-moment basis, it is hoped that information about the user's behaviors can be inferred to personalize patient care [95]. Active and passive data collection techniques have been explored, including data such as location, activity, app use, phone use, Bluetooth signals, and voice samples [96]. Research has focused on correlating such data with reported and diagnosed conditions to determine the most valid signals for mental health applications; however, this is still considered to be in its infancy.

Early studies suggest that data surrounding activity and geolocation could serve as early signs of mania or depression [97]. Furthermore, the monitoring of movement and light data was able to detect and assess depression severity [98]. Research into schizophrenia shows that digital phenotyping approaches have merit in identifying relapse events [99], that collected accelerometer and GPS data have a good correlation with future patient survey scores [100], and that such an approach was tolerated by outpatients [101].

Issues surrounding noise, privacy preservation, missing data, and data quality have been acknowledged and pose challenges in data analysis as the sensors may not be able to provide a complete context [102]. However, such approaches still require considerations of clinical relevance, social equity, development of common data standards, and multidisciplinary collaboration [103,104]. This may include the need to improve digital health literacy through training programs tailored to the needs of the target population [105].

Although it may be theoretically possible to combine smartwatch data with those collected from a smartphone to improve data quality for digital phenotyping approaches, as a smartwatch is more likely to be worn on the body than to be left behind, such an approach may be incompatible with smartwatches, which are much more resource constrained in terms of computational power, storage, connectivity, and (most importantly) battery power. The continuous collection of sensor data on smartphones has been shown to have a significant impact on battery life, which is a factor against user acceptance [103]. The impact on smartwatches, which typically have smaller batteries and rely extensively on sleep power-saving techniques to achieve all-day battery life, is anticipated to be significant.

As a result, it seems most prudent to identify the relevant physiological and physiologically related signals that relate to mental health and build algorithms focusing on data from those metrics alone rather than taking a dragnet correlation approach as is traditionally used in digital phenotyping. Such an approach will also serve to address some of the concerns regarding privacy and user perceptions that such a system is fated to diagnose users with conditions simply based on overcollection of data and misunderstanding of cause and effect [106].

Personal Health Information

The issue of personal health information regulation is important for maintaining user trust and privacy. Regulations have usually lagged behind rapid technology development, with concerns about data ownership. As such, there is some suggestion that wearable technology be considered differently from consumer technology because of inherent personal health information concerns.

Consumer wellness devices are not considered medical devices and, thus, may not be as accurate or reliable for remote health monitoring. Establishing their accuracy would require independent verification or undergoing regulatory approval processes. Constraints surrounding medical device regulation

are a source of concern as the long process can stifle innovation and the development of new technologies [107]. However, some features may be able to individually receive clearance from regulators (eg, the ECG app with the Apple Watch) [108]. The ECG app and IRNF are both classified as De Novo within the FDA regulations, which is a marketing pathway for novel devices of low to moderate risk where a predicate device does not exist. In this manner, the FDA creates a classification for the device, which can be used for future premarket approvals of equivalent devices to ensure that new and emerging novel technologies are not held back during classification.

In addition, most device manufacturers provide their own independent platforms, very similar to HealthKit for the Apple Watch, for users' data storage. These platforms may be limited in terms of data access and sharing, forming a vendor lock-in that prevents users from being able to migrate their personal health information to other platforms and reducing the research value of the devices. There are concerns over the control larger companies may have over the health data of users; this can conflict with informed consent, which is integral to medical practice [69]. Passive data collection is less intrusive and time consuming for the wearer; however, it can capture a large amount of personal data that can be stored unknown to the user, even if they have authorized the data to be recorded. Typically, the average person is more relaxed with security implementation when using personal devices and may be unaware of the level of security that third-party apps provide [13]. Similar concerns surround wearable devices and their use in workplace wellness programs and health insurance provisions if there is ambiguity regarding how the data will be used and the potential for surveillance [13,14]. The ethos behind the Apple HealthKit framework's rigid type structures and fine-grained authorization process is designed to ensure that only necessary data are collected or accessed [109,110].

The use of wearable technology for health care service provision is still in its infancy, and evidence to support its implementation is still being developed. Known concerns exist regarding passive data collection, data ownership, data use, user trust, and user attitudes toward wearable technologies, leading to potentially high abandonment rates [44,103].

Current Applications

Perhaps the best model for how the Apple Watch can be applied to mental health can be found in the insurance sector, where some insurance providers have embraced wearable technologies to promote healthier lifestyles. Incentive programs involving wearable devices have been used by numerous US health insurance providers, including United Health Care, Anthem, Humana, Health Care Service Corporation, Centene, CVS Health (Aetna), WellCare, Kaiser Permanente, GuideWell, and Molina [17]. AIA Insurance Australia has a specific program using the Apple Watch called the Vitality Apple Watch Benefit, which reduces the monthly loan repayment of the device through the achievement of weekly activity targets [16]. Loss-framed incentivized policies using the Apple Watch achieve a 34% increase in tracked activity days over 1 month in comparison with a standard gain-framed policy [12]. This offers a potential solution to individuals who may not have the financial flexibility to pay the full upfront cost of the Apple Watch device but can still have access to the benefits of the device as a wellness monitor for personal health. Another study investigated the "incentivize and persuade" health-tracking approach from both insurers and employers for enhancing business chain value. It was concluded that persuaded self-tracking, whereby service firms or employers encourage consumers and employees to collect and share data via self-tracking, is heavily influenced by service firm and individual determinants. Understanding consumer perceptions and consumer reactions within a conceptual framework should reflect values in use, privacy and security, and perceived fairness or justice as the technology itself may perpetuate inequalities [15]. Both studies noted the effects of physical activity on physical wellness, as well as mental health, but did not specifically note the impact on policy holders with severe mental illnesses. Investigation into mental health monitoring for insurance purposes could potentially create contention and the consensus that balancing privacy and confidentiality is critical for engendering trust with users and policy holders through transparency [111].

Conclusions

The Apple Watch has presented itself as a capable wearable device that is able to monitor several physiological parameters and track overall health and wellness. Its use within the mental health sphere is encouraging, particularly as more research emerges correlating changes in the emotional and physiological states of the body. Measures of HRV are key indicators of changes in both physical and emotional states. In combination with other sensors to monitor general activity, sleep, and more, health data can be aggregated with user-provided information to assist in the monitoring and even diagnosis of mental health disorders. Particular benefits may be derived through the avoidance of recall bias by providing a more objective, data-driven record of events in a passive manner. The lack of methodologically robust and replicated evidence of user benefits and a supportive health economic analysis, as well as concerns about storage, access, and security of personal health information, remain key factors that must be addressed to enable broader uptake for mental health applications.

Acknowledgments

The authors acknowledge funding in combination from Virtual Psychologist and the Theme Champion Grant Assistance Funding Program of Western Sydney University. CP received partial funding from Virtual Psychologist (author DL) for this research.

Conflicts of Interest

None declared.

References

1. Balcombe L, De Leo D. An integrated blueprint for digital mental health services amidst COVID-19. JMIR Ment Health 2020 Jul 22;7(7):e21718 [FREE Full text] [doi: 10.2196/21718] [Medline: 32668402]
2. Jo A, Coronel BD, Coakes CE, Mainous 3rd AG. Is there a benefit to patients using wearable devices such as Fitbit or health apps on mobiles? A systematic review. Am J Med 2019 Dec;132(12):1394-400.e1. [doi: 10.1016/j.amjmed.2019.06.018] [Medline: 31302077]
3. Wen D, Zhang X, Liu X, Lei J. Evaluating the consistency of current mainstream wearable devices in health monitoring: a comparison under free-living conditions. J Med Internet Res 2017 Mar 07;19(3):e68 [FREE Full text] [doi: 10.2196/jmir.6874] [Medline: 28270382]
4. Mental Health: Strengthening Mental Health Promotion. Fact sheet No. 220. World Health Organization. 2007. URL: http://www.who.int/mediacentre/factsheets/fs220/en/ [accessed 2022-07-30]
5. Mehta N, Croudace T, Davies SC. Public mental health: evidenced-based priorities. Lancet 2015 Apr 11;385(9976):1472-1475. [doi: 10.1016/S0140-6736(14)61400-8] [Medline: 25217115]
6. Ohrnberger J, Fichera E, Sutton M. The relationship between physical and mental health: a mediation analysis. Soc Sci Med 2017 Dec;195:42-49 [FREE Full text] [doi: 10.1016/j.socscimed.2017.11.008] [Medline: 29132081]
7. Paluska SA, Schwenk TL. Physical activity and mental health: current concepts. Sports Med 2000 Mar;29(3):167-180. [doi: 10.2165/00007256-200029030-00003] [Medline: 10739267]
8. Empowering people to live a healthier day. Apple Inc. 2022 Jul. URL: https://www.apple.com/newsroom/pdfs/Health-Report-July-2022.pdf [accessed 2022-07-21]
9. Lim S, Shah N. Global Smartwatch Shipments Jump 35% YoY in Q1 2021. Counterpoint Research. 2021 May 26. URL: https://www.counterpointresearch.com/global-smartwatch-shipments-q1-2021/ [accessed 2021-08-07]
10. K201525 ECG 2.0 App Electrocardiograph Software for Over-The-Counter Use Class II. U.S. Food & Drug Administration. 2020 Oct 8. URL: https://www.accessdata.fda.gov/cdrh_docs/pdf20/K201525.pdf [accessed 2021-09-30]
11. Bidargaddi N, Schrader G, Klasnja P, Licinio J, Murphy S. Designing m-Health interventions for precision mental health support. Transl Psychiatry 2020 Jul 07;10(1):222 [FREE Full text] [doi: 10.1038/s41398-020-00895-2] [Medline: 32636358]
12. Hafner M, Pollard J, Van Stolk C. Incentives and physical activity: an assessment of the association between Vitality's Active Rewards with Apple Watch benefit and sustained physical activity improvements. Rand Health Q 2020 Jun;9(1):4 [FREE Full text] [Medline: 32742746]
13. Mares CM. To cover or not to cover? The relationship between the Apple Watch and the health insurance portability and accountability act. DePaul J Health Care Law 2016;18(2):159-178. [doi: 10.4135/9781452276250.n118]
14. Chung CF, Gorm N, Shklovski IA, Munson S. Finding the right fit: understanding health tracking in workplace wellness programs. In: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems. 2017 Presented at: CHI '17; May 6-11, 2017; Denver, CO, USA p. 4875-4886. [doi: 10.1145/3025453.3025510]
15. Paluch S, Tuzovic S. Persuaded self-tracking with wearable technology: carrot or stick? J Serv Market 2019 Aug 12;33(4):436-448. [doi: 10.1108/jsm-03-2018-0091]
16. AIA Vitality Apple Watch Benefit. AIA Insurance. 2021. URL: https://www.aia.com.au/en/individual/aia-vitality/partners-and-rewards/apple-watch.html [accessed 2021-07-21]
17. Krüger S, Ní Bhroin N. Vital signs: innovations in self-tracking health insurance and social change. J Media Innov 2020 Mar 06;6(1):93-108. [doi: 10.5617/jomi.7836]
18. Hunkin H, King DL, Zajac IT. Perceived acceptability of wearable devices for the treatment of mental health problems. J Clin Psychol 2020 Jun 05;76(6):987-1003. [doi: 10.1002/jclp.22934] [Medline: 32022908]
19. Comer JS, Conroy K, Timmons AC. Ensuring wearable devices don't wear out their welcome: cautions for the mental health care road ahead. Clin Psychol (New York) 2019 Sep;26(3):e12297. [doi: 10.1111/cpsp.12297]
20. Li C, Lin SH, Chib A. The state of wearable health technologies: a transdisciplinary literature review. Mobile Media Commun 2020 Oct 29;9(2):353-376. [doi: 10.1177/2050157920966023]
21. Binsch O, Wabeke T, Valk P. Comparison of three different physiological wristband sensor systems and their applicability for resilience- and work load monitoring. In: Proceedings of the IEEE 13th International Conference on Wearable and Implantable Body Sensor Networks. 2016 Presented at: BSN '16; June 14-17, 2016; San Francisco, CA, USA p. 272-276. [doi: 10.1109/bsn.2016.7516272]
22. Shcherbina A, Mattsson CM, Waggott D, Salisbury H, Christle JW, Hastie T, et al. Accuracy in wrist-worn, sensor-based measurements of heart rate and energy expenditure in a diverse cohort. J Pers Med 2017 May 24;7(2):3 [FREE Full text] [doi: 10.3390/jpm7020003] [Medline: 28538708]
23. Dooley EE, Golaszewski NM, Bartholomew JB. Estimating accuracy at exercise intensities: a comparative study of self-monitoring heart rate and physical activity wearable devices. JMIR Mhealth Uhealth 2017 Mar 16;5(3):e34 [FREE Full text] [doi: 10.2196/mhealth.7043] [Medline: 28302596]
24. Wang R, Blackburn G, Desai M, Phelan D, Gillinov L, Houghtaling P, et al. Accuracy of wrist-worn heart rate monitors. JAMA Cardiol 2017 Jan 01;2(1):104-106. [doi: 10.1001/jamacardio.2016.3340] [Medline: 27732703]

25. Hernando D, Roca S, Sancho J, Alesanco Á, Bailón R. Validation of the Apple Watch for heart rate variability measurements during relax and mental stress in healthy subjects. Sensors (Basel) 2018 Aug 10;18(8):2619 [FREE Full text] [doi: 10.3390/s18082619] [Medline: 30103376]
26. Abt G, Bray J, Benson AC. Measuring moderate-intensity exercise with the Apple Watch: validation study. JMIR Cardio 2018 Feb 28;2(1):e6 [FREE Full text] [doi: 10.2196/cardio.8574] [Medline: 31758766]
27. Abt G, Bray J, Benson AC. The validity and inter-device variability of the Apple Watch™ for measuring maximal heart rate. J Sports Sci 2018 Jul;36(13):1447-1452. [doi: 10.1080/02640414.2017.1397282] [Medline: 29090987]
28. Roomkham S, Hittle M, Cheung J, Lovell D, Mignot E, Perrin D. Sleep monitoring with the Apple Watch: comparison to a clinically validated actigraph [version 1; peer review: 2 approved with reservations, 1 not approved]. F1000Res 2019 May 29;8:754. [doi: 10.12688/f1000research.19020.1]
29. Perez MV, Mahaffey KW, Hedlin H, Rumsfeld JS, Garcia A, Ferris T, Apple Heart Study Investigators. Large-scale assessment of a smartwatch to identify atrial fibrillation. N Engl J Med 2019 Nov 14;381(20):1909-1917 [FREE Full text] [doi: 10.1056/NEJMoa1901183] [Medline: 31722151]
30. Nuss KJ, Thomson EA, Courtney JB, Comstock A, Reinwald S, Blake S, et al. Assessment of accuracy of overall energy expenditure measurements for the Fitbit Charge HR 2 and Apple Watch. Am J Health Behav 2019 May 01;43(3):498-505. [doi: 10.5993/AJHB.43.3.5] [Medline: 31046881]
31. Thomson EA, Nuss K, Comstock A, Reinwald S, Blake S, Pimentel RE, et al. Heart rate measures from the Apple Watch, Fitbit Charge HR 2, and electrocardiogram across different exercise intensities. J Sports Sci 2019 Jun;37(12):1411-1419. [doi: 10.1080/02640414.2018.1560644] [Medline: 30657025]
32. Nelson BW, Allen NB. Accuracy of consumer wearable heart rate measurement during an ecologically valid 24-hour period: intraindividual validation study. JMIR Mhealth Uhealth 2019 Mar 11;7(3):e10828 [FREE Full text] [doi: 10.2196/10828] [Medline: 30855232]
33. Falter M, Budts W, Goetschalckx K, Cornelissen V, Buys R. Accuracy of Apple Watch measurements for heart rate and energy expenditure in patients with cardiovascular disease: cross-sectional study. JMIR Mhealth Uhealth 2019 Mar 19;7(3):e11889 [FREE Full text] [doi: 10.2196/11889] [Medline: 30888332]
34. Düking P, Giessing L, Frenkel MO, Koehler K, Holmberg HC, Sperlich B. Wrist-worn wearables for monitoring heart rate and energy expenditure while sitting or performing light-to-vigorous physical activity: validation study. JMIR Mhealth Uhealth 2020 May 06;8(5):e16716 [FREE Full text] [doi: 10.2196/16716] [Medline: 32374274]
35. Espinosa HG, Thiel DV, Sorell M, Rowlands D. Can we trust inertial and heart rate sensor data from an Apple Watch device? Proceedings 2020 Jun 15;49(1):128. [doi: 10.3390/proceedings2020049128]
36. Seshadri DR, Bittel B, Browsky D, Houghtaling P, Drummond CK, Desai M, et al. Accuracy of the Apple Watch 4 to measure heart rate in patients with atrial fibrillation. IEEE J Transl Eng Health Med 2019 Dec 13;8:2700204 [FREE Full text] [doi: 10.1109/JTEHM.2019.2950397] [Medline: 32128290]
37. Seshadri DR, Bittel B, Browsky D, Houghtaling P, Drummond CK, Desai MY, et al. Accuracy of Apple Watch for detection of atrial fibrillation. Circulation 2020 Feb 25;141(8):702-703. [doi: 10.1161/CIRCULATIONAHA.119.044126] [Medline: 32091929]
38. Glasheen E, Domingo A, Kressler J. Accuracy of Apple Watch fitness tracker for wheelchair use varies according to movement frequency and task. Ann Phys Rehabil Med 2021 Jan;64(1):101382 [FREE Full text] [doi: 10.1016/j.rehab.2020.03.007] [Medline: 32335302]
39. Huynh P, Shan R, Osuji N, Ding J, Isakadze N, Marvel FA, et al. Heart rate measurements in patients with obstructive sleep apnea and atrial fibrillation: prospective pilot study assessing Apple Watch's agreement with telemetry data. JMIR Cardio 2021 Feb 08;5(1):e18050 [FREE Full text] [doi: 10.2196/18050] [Medline: 33555260]
40. Lu TC, Fu CM, Ma MH, Fang CC, Turner AM. Healthcare applications of smart watches. A systematic review. Appl Clin Inform 2016 Sep 14;7(3):850-869 [FREE Full text] [doi: 10.4338/ACI-2016-03-R-0042] [Medline: 27623763]
41. Reeder B, David A. Health at hand: a systematic review of smart watch uses for health and wellness. J Biomed Inform 2016 Oct;63:269-276 [FREE Full text] [doi: 10.1016/j.jbi.2016.09.001] [Medline: 27612974]
42. Kim HG, Cheon EJ, Bai DS, Lee YH, Koo BH. Stress and heart rate variability: a meta-analysis and review of the literature. Psychiatry Investig 2018 Mar;15(3):235-245 [FREE Full text] [doi: 10.30773/pi.2017.08.17] [Medline: 29486547]
43. Shin G, Jarrahi MH, Fei Y, Karami A, Gafinowitz N, Byun A, et al. Wearable activity trackers, accuracy, adoption, acceptance and health impact: a systematic literature review. J Biomed Inform 2019 May;93:103153 [FREE Full text] [doi: 10.1016/j.jbi.2019.103153] [Medline: 30910623]
44. Attig C, Franke T. Abandonment of personal quantification: a review and empirical study investigating reasons for wearable activity tracking attrition. Comput Human Behav 2020 Jan;102:223-237 [FREE Full text] [doi: 10.1016/j.chb.2019.08.025]
45. Guillodo E, Lemey C, Simonnet M, Walter M, Baca-García E, Masetti V, HUGOPSY Network, et al. Clinical applications of mobile health wearable-based sleep monitoring: systematic review. JMIR Mhealth Uhealth 2020 Apr 01;8(4):e10733 [FREE Full text] [doi: 10.2196/10733] [Medline: 32234707]
46. O'Driscoll R, Turicchi J, Beaulieu K, Scott S, Matu J, Deighton K, et al. How well do activity monitors estimate energy expenditure? A systematic review and meta-analysis of the validity of current technologies. Br J Sports Med 2020 Mar;54(6):332-340. [doi: 10.1136/bjsports-2018-099643] [Medline: 30194221]

47. Hickey BA, Chalmers T, Newton P, Lin CT, Sibbritt D, McLachlan CS, et al. Smart devices and wearable technologies to detect and monitor mental health conditions and stress: a systematic review. Sensors (Basel) 2021 May 16;21(10):3461 [FREE Full text] [doi: 10.3390/s21103461] [Medline: 34065620]
48. Monitor your heart rate with Apple Watch. Apple Inc. 2021. URL: https://support.apple.com/en-us/HT204666 [accessed 2021-06-02]
49. Apple Watch. Helping your patients identify early warning signs. Apple Inc. 2021. URL: https://www.apple.com/au/healthcare/apple-watch/ [accessed 2021-09-10]
50. Using Apple Watch for Arrhythmia Detection. Apple Inc. 2020 Dec. URL: https://www.apple.com/ca/healthcare/docs/site/Apple_Watch_Arrhythmia_Detection.pdf [accessed 2021-06-21]
51. Turakhia MP, Desai M, Hedlin H, Rajmane A, Talati N, Ferris T, et al. Rationale and design of a large-scale, app-based study to identify cardiac arrhythmias using a smartwatch: the Apple Heart Study. Am Heart J 2019 Jan;207:66-75 [FREE Full text] [doi: 10.1016/j.ahj.2018.09.002] [Medline: 30392584]
52. Weng D, Ding J, Sharma A, Yanek L, Xun H, Spaulding EM, et al. Heart rate trajectories in patients recovering from acute myocardial infarction: a longitudinal analysis of Apple Watch heart rate recordings. Cardiovasc Digit Health J 2021 Oct;2(5):270-281 [FREE Full text] [doi: 10.1016/j.cvdhj.2021.05.003] [Medline: 35265918]
53. Turki A, Behbehani K, Ding K, Zhang R, Li M, Bell K. Estimation of heart rate variability measures using Apple Watch and evaluating their accuracy: estimation of heart rate variability measures using Apple Watch. In: Proceedings of the 14th PErvasive Technologies Related to Assistive Environments Conference. 2021 Presented at: PETRA '21; June 29-July 2, 2021; Corfu, Greece p. 565-574. [doi: 10.1145/3453892.3462647]
54. Lam E, Aratia S, Wang J, Tung J. Measuring heart rate variability in free-living conditions using consumer-grade photoplethysmography: validation study. JMIR Biomed Eng 2020 Nov 3;5(1):e17355. [doi: 10.2196/17355]
55. Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology. Heart rate variability: standards of measurement, physiological interpretation and clinical use. Circulation 1996 Mar 01;93(5):1043-1065. [Medline: 8598068]
56. Stone JD, Ulman HK, Tran K, Thompson AG, Halter MD, Ramadan JH, et al. Assessing the accuracy of popular commercial technologies that measure resting heart rate and heart rate variability. Front Sports Act Living 2021 Mar 1;3:585870 [FREE Full text] [doi: 10.3389/fspor.2021.585870] [Medline: 33733234]
57. HealthKit Documentation - Type Property - heartRateVariabilitySDNN. Apple Inc. 2021. URL: https://developer.apple.com/documentation/healthkit/hkquantitytypeidentifier/2881127-heartratevariabilitysdnn [accessed 2021-06-02]
58. Dalmeida KM, Masala GL. HRV features as viable physiological markers for stress detection using wearable devices. Sensors (Basel) 2021 Apr 19;21(8):2873 [FREE Full text] [doi: 10.3390/s21082873] [Medline: 33921884]
59. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders (DSM-5). 5th edition. Arlington, VA, USA: American Psychiatric Association; 2013.
60. Waxenbaum JA, Reddy V, Varacallo M. Anatomy, Autonomic Nervous System. Treasure Island, FL, USA: StatPearls Publishing; Jul 29, 2021.
61. Quintana DS, Heathers JA. Considerations in the assessment of heart rate variability in biobehavioral research. Front Psychol 2014 Jul 22;5:805 [FREE Full text] [doi: 10.3389/fpsyg.2014.00805] [Medline: 25101047]
62. Düking P, Tafler M, Wallmann-Sperlich B, Sperlich B, Kleih S. Behavior change techniques in wrist-worn wearables to promote physical activity: content analysis. JMIR Mhealth Uhealth 2020 Nov 19;8(11):e20820 [FREE Full text] [doi: 10.2196/20820] [Medline: 33211023]
63. Use the Activity app on your Apple Watch. Apple Inc. 2020. URL: https://support.apple.com/en-au/HT204523 [accessed 2021-08-09]
64. Wallen MP, Gomersall SR, Keating SE, Wisløff U, Coombes JS. Accuracy of heart rate watches: implications for weight management. PLoS One 2016 May 27;11(5):e0154420 [FREE Full text] [doi: 10.1371/journal.pone.0154420] [Medline: 27232714]
65. Boudreaux BD, Hebert EP, Hollander DB, Williams BM, Cormier CL, Naquin MR, et al. Validity of wearable activity monitors during cycling and resistance exercise. Med Sci Sports Exerc 2018 Mar;50(3):624-633. [doi: 10.1249/MSS.0000000000001471] [Medline: 29189666]
66. Workout Types on Apple Watch. Apple Inc. 2020. URL: https://support.apple.com/en-au/HT207934 [accessed 2021-07-12]
67. watchOS 8 brings new access, connectivity and mindfulness features to Apple Watch. Apple Inc. 2021 Jun 8. URL: https://www.apple.com/au/newsroom/2021/06/watchos-8-brings-new-access-connectivity-and-mindfulness-features-to-apple-watch/ [accessed 2021-08-24]
68. Ryan J, Edney S, Maher C. Anxious or empowered? A cross-sectional study exploring how wearable activity trackers make their owners feel. BMC Psychol 2019 Jul 03;7(1):42 [FREE Full text] [doi: 10.1186/s40359-019-0315-y] [Medline: 31269972]
69. Patel S, Saunders KE. Apps and wearables in the monitoring of mental health disorders. Br J Hosp Med (Lond) 2018 Dec 02;79(12):672-675. [doi: 10.12968/hmed.2018.79.12.672] [Medline: 30526097]
70. Tushar AK, Kabir MA, Ahmed SI. Mental health and sensing. In: Ahad MA, Ahmed MU, editors. Signal Processing Techniques for Computational Health Informatics. Cham, Switzerland: Springer; 2021:247-260.

71. Knight A, Bidargaddi N. Commonly available activity tracker apps and wearables as a mental health outcome indicator: a prospective observational cohort study among young adults with psychological distress. J Affect Disord 2018 Aug 15;236:31-36. [doi: 10.1016/j.jad.2018.04.099] [Medline: 29709718]
72. Track your sleep with Apple Watch. Apple Inc. 2021. URL: https://support.apple.com/en-au/guide/watch/apd830528336/watchos [accessed 2021-09-30]
73. Dove J. How to track your sleep with an Apple Watch. Digital Trends. 2022. URL: https://www.digitaltrends.com/mobile/how-to-track-sleep-with-apple-watch/ [accessed 2021-09-30]
74. Ameen MS, Cheung LM, Hauser T, Hahn MA, Schabus M. About the accuracy and problems of consumer devices in the assessment of sleep. Sensors (Basel) 2019 Sep 25;19(19):4160 [FREE Full text] [doi: 10.3390/s19194160] [Medline: 31557952]
75. Liang Z, Ploderer B. Sleep tracking in the real world: a qualitative study into barriers for improving sleep. In: Proceedings of the 28th Australian Conference on Computer-Human Interaction. 2016 Presented at: OzCHI '16; November 29-December 2 2016; Launceston, Tasmania, Australia p. 527-541. [doi: 10.1145/3010915.3010988]
76. Apple Watch Battery. Apple Inc. 2021. URL: https://www.apple.com/au/watch/battery/ [accessed 2021-09-08]
77. Can YS, Arnrich B, Ersoy C. Stress detection in daily life scenarios using smart phones and wearable sensors: a survey. J Biomed Inform 2019 Apr;92:103139 [FREE Full text] [doi: 10.1016/j.jbi.2019.103139] [Medline: 30825538]
78. Apple Watch Series 7. Apple Inc. 2021. URL: https://www.apple.com/au/apple-watch-series-7/ [accessed 2021-06-02]
79. Edwardes S. Apple Watch Photoplethysmography (PPG). Helix Apps. 2021. URL: https://www.helixapps.co.uk/blog/apple-watch-photoplethysmography-ppg [accessed 2021-09-08]
80. De Novo Classification Request for Irregular Rhythm Notificaiton Feature. Food and Drug Administration and Center for Drugs Evaluation Research. 2018 Aug 8. URL: https://www.accessdata.fda.gov/cdrh_docs/reviews/DEN180042.pdf [accessed 2022-02-16]
81. Heart Health Notifications on your Apple Watch. Apple Inc. 2021. URL: https://support.apple.com/en-us/HT208931 [accessed 2021-06-02]
82. De Novo Classification Request for ECG App. Food and Drug Administration and Center for Drugs Evaluation Research. 2018 Aug 14. URL: https://www.accessdata.fda.gov/cdrh_docs/reviews/DEN180044.pdf [accessed 2022-02-16]
83. Isakadze N, Martin SS. How useful is the smartwatch ECG? Trends Cardiovasc Med 2020 Oct;30(7):442-448 [FREE Full text] [doi: 10.1016/j.tcm.2019.10.010] [Medline: 31706789]
84. Self-care monitoring Web based application software. Theraputic Goods Administration. 2021. URL: https://www.tga.gov.au/resources/artg/354056 [accessed 2021-06-21]
85. Cardiac Electrophysiology Application Software. Theraputic Goods Administration. 2021. URL: https://www.tga.gov.au/resources/artg/355992 [accessed 2021-06-21]
86. Ringwald M, Crich A, Beysard N. Smart watch recording of ventricular tachycardia: case study. Am J Emerg Med 2020 Apr;38(4):849.e3-849.e5. [doi: 10.1016/j.ajem.2019.10.040] [Medline: 31785973]
87. Garrido S, Millington C, Cheers D, Boydell K, Schubert E, Meade T, et al. What works and what doesn't work? A systematic review of digital mental health interventions for depression and anxiety in young people. Front Psychiatry 2019 Nov 13;10:759 [FREE Full text] [doi: 10.3389/fpsyt.2019.00759] [Medline: 31798468]
88. Raja JM, Elsakr C, Roman S, Cave B, Pour-Ghaz I, Nanda A, et al. Apple Watch, wearables, and heart rhythm: where do we stand? Ann Transl Med 2019 Sep;7(17):417 [FREE Full text] [doi: 10.21037/atm.2019.06.79] [Medline: 31660316]
89. Kunchay S, Abdullah S. WatchOver: using Apple Watches to assess and predict substance co-use in young adults. In: Adjunct Proceedings of the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2020 ACM International Symposium on Wearable Computers. 2020 Sep Presented at: UbiComp-ISWC '20; September 12-17, 2020; Virtual p. 488-493. [doi: 10.1145/3410530.3414373]
90. Cormack F, McCue M, Taptiklis N, Skirrow C, Glazer E, Panagopoulos E, et al. Wearable technology for high-frequency cognitive and mood assessment in major depressive disorder: longitudinal observational study. JMIR Ment Health 2019 Nov 18;6(11):e12814 [FREE Full text] [doi: 10.2196/12814] [Medline: 31738172]
91. SensorKit Framework Documentation. Apple Inc. 2021. URL: https://developer.apple.com/documentation/sensorkit [accessed 2021-08-08]
92. HealthKit Health and Fitness. Apple Inc. 2021. URL: https://developer.apple.com/health-fitness/ [accessed 2021-08-08]
93. Apple Inc. 2021. URL: https://www.researchandcare.org/ [accessed 2021-08-08]
94. Liang Y, Zheng X, Zeng DD. A survey on big data-driven digital phenotyping of mental health. Inf Fusion 2019 Dec;52:290-307 [FREE Full text] [doi: 10.1016/j.inffus.2019.04.001]
95. Insel TR. Digital phenotyping: technology for a new science of behavior. JAMA 2017 Oct 03;318(13):1215-1216. [doi: 10.1001/jama.2017.11295] [Medline: 28973224]
96. Onnela JP, Rauch SL. Harnessing smartphone-based digital phenotyping to enhance behavioral and mental health. Neuropsychopharmacology 2016 Jun;41(7):1691-1696 [FREE Full text] [doi: 10.1038/npp.2016.7] [Medline: 26818126]
97. Insel TR. Digital phenotyping: a global tool for psychiatry. World Psychiatry 2018 Oct;17(3):276-277 [FREE Full text] [doi: 10.1002/wps.20550] [Medline: 30192103]

98. Jacobson NC, Weingarden H, Wilhelm S. Using digital phenotyping to accurately detect depression severity. J Nerv Ment Dis 2019 Oct;207(10):893-896. [doi: 10.1097/NMD.0000000000001042] [Medline: 31596769]
99. Marsch LA. Opportunities and needs in digital phenotyping. Neuropsychopharmacology 2018 Jul;43(8):1637-1638 [FREE Full text] [doi: 10.1038/s41386-018-0051-7] [Medline: 29703995]
100. Torous J, Staples P, Barnett I, Sandoval LR, Keshavan M, Onnela JP. Characterizing the clinical relevance of digital phenotyping data quality with applications to a cohort with schizophrenia. NPJ Digit Med 2018 Apr 6;1:15 [FREE Full text] [doi: 10.1038/s41746-018-0022-8] [Medline: 31304300]
101. Raugh IM, James SH, Gonzalez CM, Chapman HC, Cohen AS, Kirkpatrick B, et al. Digital phenotyping adherence, feasibility, and tolerability in outpatients with schizophrenia. J Psychiatr Res 2021 Jun;138:436-443 [FREE Full text] [doi: 10.1016/j.jpsychires.2021.04.022] [Medline: 33964681]
102. Onnela JP. Opportunities and challenges in the collection and analysis of digital phenotyping data. Neuropsychopharmacology 2021 Jan;46(1):45-54 [FREE Full text] [doi: 10.1038/s41386-020-0771-3] [Medline: 32679583]
103. Huckvale K, Venkatesh S, Christensen H. Toward clinical digital phenotyping: a timely opportunity to consider purpose, quality, and safety. NPJ Digit Med 2019 Sep 6;2:88 [FREE Full text] [doi: 10.1038/s41746-019-0166-1] [Medline: 31508498]
104. Birk RH, Samuel G. Can digital data diagnose mental health problems? A sociological exploration of 'digital phenotyping'. Sociol Health Illn 2020 Nov;42(8):1873-1887. [doi: 10.1111/1467-9566.13175] [Medline: 32914445]
105. Hoffman L, Wisniewski H, Hays R, Henson P, Vaidyam A, Hendel V, et al. Digital opportunities for outcomes in recovery services (DOORS): a pragmatic hands-on group approach toward increasing digital health and smartphone competencies, autonomy, relatedness, and alliance for those with serious mental illness. J Psychiatr Pract 2020 Mar;26(2):80-88 [FREE Full text] [doi: 10.1097/PRA.0000000000000450] [Medline: 32134881]
106. Stanghellini G, Leoni F. Digital phenotyping: ethical issues, opportunities, and threats. Front Psychiatry 2020 May 27;11:473 [FREE Full text] [doi: 10.3389/fpsyt.2020.00473] [Medline: 32536882]
107. Arnow G. Apple Watch-ing you: why wearable technology should be federally regulated. Loy LA Law Rev 2016;49(3):607.
108. Expanded use of Apple ECG for supporting remote heart rhythm evaluation during the COVID-19 pandemic. Apple Inc. 2020. URL: https://www.apple.com/healthcare/docs/site/Apple_ECG_app_during_COVID-19.pdf [accessed 2022-01-17]
109. Authorizing Access to Health Data. Apple Inc. 2020. URL: https://developer.apple.com/documentation/healthkit/authorizing_access_to_health_data [accessed 2021-07-21]
110. HealthKit Human Interface Guidelines. Apple Inc. 2021. URL: https://developer.apple.com/design/human-interface-guidelines/healthkit/overview/ [accessed 2021-08-08]
111. Abdullah S, Choudhury T. Sensing technologies for monitoring serious mental illnesses. IEEE MultiMedia 2018 Jan;25(1):61-75. [doi: 10.1109/mmul.2018.011921236]

Abbreviations

AF: atrial fibrillation
bpm: beats per minute
ECG: electrocardiogram
EE: energy expenditure
FDA: Food and Drug Administration
HR: heart rate
HRV: heart rate variability
IRNF: Irregular Rhythm Notification Feature
LED: light-emitting diode
PPG: photoplethysmography
SDNN: SD of NN intervals

Edited by J Torous; submitted 17.02.22; peer-reviewed by L Balcombe, G Abt; comments to author 20.05.22; revised version received 29.07.22; accepted 03.08.22; published 07.09.22.

Please cite as:
Lui GY, Loughnane D, Polley C, Jayarathna T, Breen PP
The Apple Watch for Monitoring Mental Health–Related Physiological Symptoms: Literature Review
JMIR Ment Health 2022;9(9):e37354
URL: https://mental.jmir.org/2022/9/e37354
doi:10.2196/37354
PMID:36069848

Review

Efficacy and Conflicts of Interest in Randomized Controlled Trials Evaluating Headspace and Calm Apps: Systematic Review

Alison O'Daffer[1], BA; Susannah F Colt[2], MA; Akash R Wasil[3], MA; Nancy Lau[1,4], PhD

[1]Palliative Care and Resilience Research Program, Center for Clinical and Translational Research, Seattle Children's Research Institute, Seattle, WA, United States

[2]Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, United States

[3]Department of Psychology, University of Pennsylvania, Philadelphia, PA, United States

[4]Department of Psychiatry and Behavioral Sciences, University of Washington, Seattle, WA, United States

Corresponding Author:
Alison O'Daffer, BA
Palliative Care and Resilience Research Program
Center for Clinical and Translational Research
Seattle Children's Research Institute
1920 Terry Ave
Seattle, WA, 98101
United States
Phone: 1 206 884 4211
Email: alison.odaffer@seattlechildrens.org

Abstract

Background: Although there are thousands of mental health apps, 2 apps, Headspace and Calm, claim a large percentage of the marketplace. These two mindfulness and meditation apps have reached tens of millions of active users. To guide consumers, clinicians, and researchers, we performed a systematic review of randomized controlled trials (RCTs) of Headspace and Calm.

Objective: Our study aimed to evaluate intervention efficacy, risk of bias, and conflicts of interest (COIs) in the evidence base for Headspace and Calm, the two most popular mental health apps at the time of our search.

Methods: To identify studies, we searched academic databases (Google Scholar, MEDLINE, and PsycINFO) and the websites of Headspace and Calm in May 2021 for RCTs of Headspace and Calm testing efficacy via original data collection, published in English in peer-reviewed journals. For each study, we coded (1) study characteristics (eg, participants, sample size, and outcome measures), (2) intervention characteristics (eg, free vs paid version of the app and intended frequency of app usage), (3) all study outcomes, (4) Cochrane risk of bias variables, and (5) COI variables (eg, presence or absence of a preregistration and the presence or absence of a COI statement involving the company).

Results: We identified 14 RCTs of Headspace and 1 RCT of Calm. Overall, 93% (13/14) of RCTs of Headspace and 100% (1/1) of RCTs of Calm recruited participants from a nonclinical population. Studies commonly measured mindfulness, well-being, stress, depressive symptoms, and anxiety symptoms. Headspace use improved depression in 75% of studies that evaluated it as an outcome. Findings were mixed for mindfulness, well-being, stress, and anxiety, but at least 40% of studies showed improvement for each of these outcomes. Studies were generally underpowered to detect “small” or “medium” effect sizes. Furthermore, 50% (7/14) of RCTs of Headspace and 0% (0/1) of RCTs of Calm reported a COI that involved Headspace or Calm (the companies). The most common COI was the app company providing premium app access for free for participants, and notably, 14% (2/14) of RCTs of Headspace reported Headspace employee involvement in study design, execution, and data analysis. Only 36% (5/14) of RCTs of Headspace were preregistered, and the 1 RCT of Calm was not preregistered.

Conclusions: The empirical research on Headspace appears promising, whereas there is an absence of randomized trials on Calm. Limitations of this study include an inability to compare Headspace and Calm owing to the dearth of RCTs studying Calm and the reliance on author reports to evaluate COIs. When determining whether or not mental health apps are of high quality, identification of high-quality apps and evaluation of their effectiveness and investigators’ COIs should be ensured.

(JMIR Ment Health 2022;9(9):e40924) doi:10.2196/40924

KEYWORDS

mHealth; psychological interventions; mobile health; mental health; health applications; health apps; mindfulness; meditation app; digital health application; digital health intervention

Introduction

Background

Mental health problems are leading contributors to the global burden of disease [1]. As a result, efforts to improve population-level mental health and wellness are a public health priority. Although empirically supported treatments exist for mental health problems, most people in need of support do not access traditional mental health treatments [2]. Common barriers to treatment access include high costs, low supply and availability of clinicians, stigma toward professional treatments, and preferences for self-help [3,4].

Mental health help-seekers have gravitated toward low-barrier, cost-effective prevention and intervention programs, mainly mental health apps. Although there are thousands of mental health apps, data through 2021 have shown that 2 apps, Calm and Headspace, are the most popular and consistently rank the highest in the number of downloads and user activity [5-10]. Both apps include mindfulness meditation and deep breathing content and allow users the ability to select the topic (eg, sleep or stress relief), length, and modality of a guided sessions each time they use the app (with the option to follow specific modules in order). The app landscape is dynamic, but 2019 estimates suggest that each app reaches approximately 5-9 million monthly active users, and the apps are responsible for approximately 90% of total monthly active users [5,11]. Given the widespread dissemination of these apps, evaluation of the quality of the evidence for these apps is a public health priority. Such a review could help identify if, for whom, and for which conditions these mental health apps have been shown to be effective. Although previous systematic reviews and meta-analyses have shown that mental health apps can be effective for depression and anxiety [7,12-14], there is little overlap between the apps that are evaluated in academic research [15] and those that are widely disseminated on public-facing app stores [5,16]. Thus, reviewing Headspace and Calm is an important priority, and the findings from existing reviews of mental health apps may not generalize to these commercially popular apps.

Headspace Inc and Calm are both for-profit companies, and both companies use research findings to promote their products. Increasing interest in the clinical robustness of these apps [17] presents a potential conflict of interest (COI): companies may have incentives to publish "positive" findings and suppress negative or inconclusive results. Even among academic researchers, incentives to publish "positive" results has contributed to biased literature, leading to concerns about the reproducibility of psychological science [15]. There has also been a growing conversation about "researcher degrees of freedom"— decisions in data collection and analysis that may contribute to the elevated rate of false positives in psychological science [18]. While these concerns are always worth considering when reviewing academic literature, they may be especially salient when for-profit companies are performing or funding research on their own products (eg, elevated estimates of the effectiveness of antidepressant medications [19]). It is plausible that similar concerns could be present in digital mental health interventions [20], especially in cases where companies are explicitly funding, sponsoring, or participating in clinical trials.

Objectives

In this study, we systematically reviewed randomized controlled trials (RCTs) of Headspace and Calm, the two most popular mental health apps. These two apps dominate the mental health app market, both in absolute terms (reaching millions of users each month) and relative terms (reaching up to 90% of mental health app users). We aimed to (1) evaluate the efficacy of these apps and (2) evaluate the risk of bias and COIs in the studies contributing to this evidence base. Owing to a wide range of outcomes of interest across studies, we did not conduct a meta-analysis. The purpose of this review is to provide researchers, clinicians, and consumers with up-to-date information regarding the evidence base, risk of bias, and COIs of the two most popular mental health apps.

Methods

Search Strategy

Our approach is outlined in detail in the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) diagram (Multimedia Appendix 1). Two authors (AO and SC) conducted a literature review via Google Scholar, MEDLINE, and PsycINFO databases using the search terms "*[app name]*" AND "*smartphone*" in May 2021 to identify peer reviewed RCTs of Headspace and Calm. To supplement this procedure, we also identified articles via the websites for Headspace and Calm, which list peer-reviewed publications on their respective apps. The date range for this search had no start date and ended in May 2021. Inclusion criteria were as follows: RCTs of Headspace and Calm testing efficacy, published in peer reviewed journals, in English only, and solely including original data collection. Exclusion criteria included non-Headspace or -Calm papers non-RCTs, nonoriginal data collection, conference abstracts, or student theses, non–English-language papers, and papers not published in a peer reviewed journal.

Three authors (SC, AO, and NL) retrieved and independently reviewed the full text of all eligible studies. Two authors (AO and NL) coded half of the included articles, and one author (SC) coded all the articles, such that each article was coded by at least 2 coders. To resolve discrepancies, coders conducted consensus conversations and referred to the articles for resolution.

Data Extraction

Trial Outcomes

We extracted the following information from each included article: participants, sample size, intervention adherence,

treatment condition, and type of control condition. We extracted all reported outcomes, regardless of whether the outcome in question was examined as a primary, secondary, or exploratory outcome, and we documented whether each outcome measured was positive or negative (or null). A positive outcome was defined as the intervention condition outperforming the control with statistically significant findings. A negative outcome was defined as the control condition outperforming the intervention with statistically significant findings. A null outcome was defined as nonsignificant differences between the intervention and control conditions. To better characterize the studies and to understand how participants engaged with the apps, we also examined additional variables: (1) whether or not users had access to the premium (paid) versions of the app, (2) whether users were instructed to use specific parts of the app, or if they were told to use the app freely and choose which content to access, (3) the intended frequency of use, (4) the actual of frequency of use (if measured), (5) the length of the intervention (eg, 4 weeks), and (6) incentives that were provided to participants for their participation in the study.

Power calculations were performed using G*Power (version 3.1) assuming an α of .05 and a desired power of at least 0.823) [21]. We considered a study's power as "high" if the study included enough participants to detect a between-group standardized mean difference of 0.3, "medium" if it included enough participants to detect a standardized mean difference of 0.5, and "low" if it did not include enough participants to detect a standardized mean difference of 0.5. Thus, studies with over 278 participants were coded as "high," studies with between 102 and 278 participants were coded as "medium," and those with fewer than 102 participants were coded as "low" in power.

Risk of Bias

We assessed risk of bias using the Cochrane collaboration's risk of bias tool. Three authors (NL, SC, and AO) independently assessed risk of bias by applying 7 criteria from the Cochrane collaboration's tool for assessing risk of bias: (1) evaluating random sequence generation (selection bias), (2) allocation concealment (selection bias), (3) blinding of participants and personnel (performance bias), (4) blinding of outcome assessment (detection bias), (5) incomplete outcome data (attrition bias), and (6) selective reporting [22].

Assessment of COIs

There have been concerns about the reproducibility in psychological science [15], which may be especially salient when research is conducted or supported by profit-driven companies [19]. Thus, it is important to apply additional codes to assess study rigor and bias, going beyond those included in the Cochrane framework. To develop these additional codes, we reviewed relevant work on risk of bias disclosure recommendations [15,22] and open science practices [23,24]. Specifically, we examined if (1) the companies had any role in the study, (2) the companies initiated the study, (3) the companies were involved in analyzing data, (4) the companies provided funding for the study, (5) the companies were mentioned in the acknowledgments section, (6) members of the companies were included as coauthors, and (7) trial preregistration. Preregistration—the act of specifying research questions, relevant variables, and planned analyses before data collection—is a highly valued open science practice that is thought to reduce the use of questionable research practices [25].

Results

Study Details

Overview

Our final sample consisted of 15 studies. We identified 14 RCTs of Headspace [26-39] and 1 RCT of Calm [40]. Table 1 summarizes the characteristics and findings of each study on all outcomes measured. We categorized specific study outcomes into 5 overarching constructs (mindfulness, psychological well-being, stress, anxiety, and depression) representative of the psychosocial outcomes that mental health apps including Headspace and Calm purport to target. This categorization scheme enabled us to descriptively synthesize the various outcome measures into overarching domains.

Table 1. Included studies (N=15).

Authors (year; country)	Participants (sample size)	Intervention	Control group	Adherence	Positive results	Null or inconclusive results	Gender (% female) and age (mean)
Calm							
Huberty et al (2019; United States) [40]	College students (n=88)	10 minutes of daily use for 8 weeks	Waitlist	On average, intervention participants completed 37.9/70 (54%) minutes of meditation per week over the course of the study	Improved stress, mindfulness, and self-compassion	N/A[a]	• 88 • Intervention: 20.41 years; Control: 21.85 years
Headspace							
Bennike et al (2017; Denmark) [26]	University staff novice meditators (n=95)	10 minutes daily for week 1, 15 minutes daily for week 2, and 20 minutes daily for week 3	Cognitive-training app use for 30 days	On average, intervention participants completed 302.7/450 (67%) minutes of the required meditation minutes over the study period	Improved dispositional mindfulness and mind wandering	N/A	• 69% in the intervention group; 71% in the control group • Intervention: 41.4 years; control: 43.4 years
Bjorkstrand et al (2019; Sweden) [27]	Adults without extensive meditation experience (n=26)	Daily 10-20–minute guided mindfulness meditation sessions over 4 weeks	Waitlist	On average, intervention participants completed 13.2 minutes of meditation per day	Improved retention of extinction learning on day 2	No effect on fear acquisition or extinction of conditioned response on day 1	• 79% (86% in the intervention group; 73% int he control group) • 35.1 years (intervention: 35.6 years; control: 34.5 years)
Bostock et al (2019; United Kingdom) [28]	Adult employees of 2 firms in the United Kingdom reporting work stress (n=238)	45 sessions of guided mindfulness meditation over 8 weeks	Waitlist	On average, participants completed 16.6/45 sessions (37%)	Improved global well-being, daily positive affect, anxiety and depressive symptoms, job strain, and workplace social support	Marginally significant improvement in systolic blood pressure. No effect on diastolic blood pressure	• 59% • 35.5 years
Champion et al (2018; United Kingdom) [29]	Adult novice meditators (n=74)	Daily use for 30 days	Waitlist	On average, intervention participants completed 6.21/10 (62%) sessions in the first 10 days and 11.66/20 (58%) sessions in the second 20 days[b]	Improved satisfaction with life, stress, and resilience	N/A	• 55% • 39.4 years

Authors (year; country)	Participants (sample size)	Intervention	Control group	Adherence	Positive results	Null or inconclusive results	Gender (% female) and age (mean)
DeSteno et al (2018; United States) [30]	College student novice meditators (n=46)	Daily meditation training (approximately 15 min) for 3 weeks	Daily logic problem	53/77 (68%) intervention participants completed all required sessions	Improved aggression	No effect on anger or executive control	• Gender % not reported • Age range: 18-24 years (average not reported)
Economides et al (2018; United States) [31]	Adult novice meditators (n=88)	10 sessions in 1 month	Mindfulness or meditation psychoeducational audiobook	69/88 (78%) participants completed all sessions	Improved irritability, affect, and stress from external issues	No effect on stress from internal pressure	• 57% • 28% aged 18-24 years; 26% aged 25-29 years, 27% aged 30-39 years, and 19% aged 40-49 years
Flett et al (2018; New Zealand) [32]	College students (n=208)	Daily use for 10 days	Intervention arm 2: Smiling Mind app use; control: Evernote app use	On average, intervention participants completed 8.24/10 sessions (82%)[b]	Improved depressive symptoms and college adjustment (for both Headspace and Smiling Mind users). Improved mindfulness for Headspace users. (Improved resilience for Smiling Mind users). Improvements were maintained for participants who continued to use intervention apps	No differences in flourishing, stress, or anxiety. No effect on resilience for Headspace users. (No effect on mindfulness for Smiling Mind users)	• Gender % not reported • 20.08 years
Howells et al (2016; United Kingdom) [33]	Adult app users (n=121)	10 minutes daily for 10 days	List-making app use (Catch Notes)	Not reported	Improved positive affect and depressive symptoms	No effect on satisfaction with life, flourishing, or negative affect	• 87% • 40.7 years
Kubo et al (2019; United States) [34]	Arm 1: patients with a diagnosis of cancer (n=72). Arm 2: their caregivers (26)	8 weeks of daily mindfulness sessions delivered via Headspace app	Waitlist	Not reported	Patients: improved overall well-being. (Caregivers: improved FFMQ[c] observing mindfulness domain score)	Patients: no statistically significant differences in change in anxiety, depression, sleep, or fatigue	• Arm 1: 69% • Arm 2: 58% • Mean age not reported
Lim et al (2015; United States) [35]	College student novice meditators (n=56)	14 sessions plus daily quiz over 3 weeks	14 sessions of cognitive-training app plus daily questionnaire	Not reported	Improved compassionate responding	No effect on empathic accuracy	• 54% • 19.4 years
Noone and Hogan (2018; Ireland) [36]	College students (n=91)	30 mindfulness meditation sessions over 6 weeks	30 sham meditations delivered through Headspace interface	On average, intervention participants completed 15/30 (50%) sessions	N/A	No difference between groups in mindful disposition, critical thinking, or executive functioning	• 76% • 20.92 years
Quinones and Griffiths (2019; United Kingdom) [37]	Adult novice meditators with signs of compulsive internet use (n=994)	Daily 10-minute mindfulness podcast	Active control: muscle relaxation podcast. Passive control: waitlist	Not reported	Improved mindfulness and compulsive internet use in the intervention group compared to active control and waitlist groups		

Authors (year; country)	Participants (sample size)	Intervention	Control group	Adherence	Positive results	Null or inconclusive results	Gender (% female) and age (mean)
						No differences between mindfulness and active control groups in anxiety or depression, but both outperformed waitlist group	• Intervention group: 38%; active control: 42%; waitlist control: 37% • Intervention group: 39 years; active control group: 40 years; waitlist control: 41 years
Rosen et al (2018; United States) [38]	Women diagnosed with breast cancer (n=112)	Self-guided app-delivered mindfulness training for 8 weeks	Waitlist	On average, intervention patients used the app 18/72 (25%) days.	Improved quality of life and mindfulness.	N/A	• 100% • Intervention group: 51.4 years; control group: 53.22 years
Yang et al (2018; United States) [39]	Medical students (n=88)	App-delivered mindfulness training over 30 days	Waitlist	On average, intervention participants completed 11.97/30 (40%) sessions[b]	Improved well-being and stress	No differences between groups for mindfulness	• 64% • 25.11 years

[a]N/A: not applicable.

[b]Self-report data.

[c]FFMQ: Five Facet Mindfulness Questionnaire

RCTs of the Headspace App

Among the RCTs of the Headspace app, 43% (6/14) of studies recruited novice meditators (individuals with some experience with meditation practices prior to the study), 29% (4/14) of them included college students, 14% (2/14) of them included patients with cancer, and 7% (1/14) of them included individuals with compulsive internet use. In other words, most studies focused on samples from the general population, rather than individuals with elevated levels of depression, anxiety, or another mental disorders. Overall, 50% (7/14) of studies included a measure of mindfulness, 57% (8/14) of them measured well-being, 36% (5/14) of them measured stress, 29% (4/14) of them measured depressive symptoms, and 29% (4/14) of them measured anxiety symptoms. Furthermore, 43% (6/14) of the studies used waitlist control conditions, 43% (6/14) of them had active control conditions, and 14% (2/14) of them had both an active and a waitlist control condition. Of the 5 RCTs with only active control conditions, 33% (2/6) of studies used cognitive training apps, 17% (1/6) of studies had participants do daily logic problems, 17% (1/6) of studies used a mindfulness or meditation psychoeducation audiobook, and 17% (1/6) of studies had participants complete sham meditation sessions through the Headspace app.

In 93% (13/14) of studies, participants were allowed to access content from the premium (paid) version of the app. In 93% (13/14) of studies, participants were instructed to use specific parts of the app (as opposed to navigating the app freely). The intervention period ranged from 14 days to 70 days (mean 33.14, SD 14.53 days). In 79% (11/14) of studies, participants were instructed to use the app for at least 10 minutes each day. Overall, 29% (4/14) of studies did not report app adherence data, 21% (3/14) of studies asked participants for self-reported usage data, and 50% (7/14) of studies used backend app usage data from Headspace to evaluate app usage. App adherence metrics varied greatly (including days of meditation completed, minutes of meditation completed, number of participants who completed entire intervention, and number of completed sessions), and these data are provided in Multimedia Appendix 2. No paper reported lower than 25% adherence or higher than 90.16% adherence on the measure they utilized. Furthermore, 71% (10/14) of studies offered some sort of incentive to participants (eg, gift card, course credit, premium app access, and lottery entry) and 29% (4/14) of studies offered no incentive for participation.

RCT of the Calm App

The 1 RCT of the Calm app recruited college students, who were not required to have a mental health diagnosis or clinically significant distress. Of our 5 outcome domains of interest, the Calm RCT measured stress and mindfulness. The active control condition was a waitlist control. Participants were allowed to access content from the premium (paid) version of the app and were instructed to use specific modules within the app. Participants were instructed to use the app daily for at least 10

minutes a day for 8 weeks. On average, intervention participants completed 37.9 out of 70 minutes (54%) of meditation per week over the course of the study. Participants were given gift cards as an incentive for completing questionnaires.

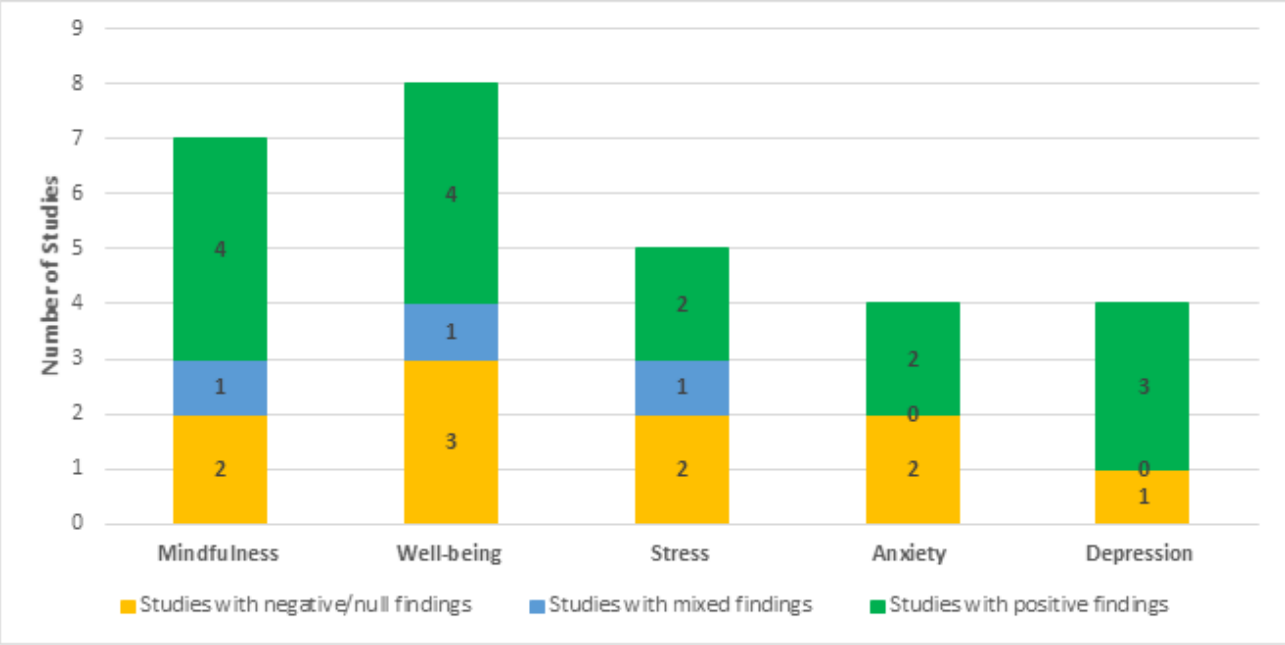
Trial Outcomes

RCTs of the Headspace App

Figure 1 presents a summary of main findings from the 14 RCTs of Headspace. We categorized findings in the 5 domains of interest as "positive" (ie, the intervention group showed significant improvement compared to the control group), "mixed" (ie, 2 or more measures of the same outcome domain that yielded conflicting results), or "null" (ie, the intervention group did not outperform the control group) for each outcome domain. Of the RCTs of Headspace that evaluated mindfulness, 57% (4/7) had positive findings, 14% (1/7) had mixed findings, and 29% (2/7) had null findings. Of the RCTs of Headspace that evaluated well-being, 50% (4/8) had positive findings, 13% (1/8) had mixed findings, and 38% (3/8) had null findings. For the RCTs of Headspace that evaluated stress, 40% (2/5) had positive findings, 20% (1/5) had mixed findings, and 40% (2/5) had null findings. For anxiety, 50% (2/4) of studies had positive findings, and 50% (2/4) had null findings. Finally, for RCTs of Headspace evaluating depression, 75% (3/4) had positive findings and 25% (1/4) had null findings. We were unable to calculate effect size pooled estimates owing to the small number of studies, the variability in outcome measures, and the wide range of timing for administration of postassessments.

Sample sizes (number of participants included in analyses) ranged from 46 to 994 (mean 174, median 102, SD 234). Applying our coding system, 64% (9/14) of the studies had low power (<102 participants), 29% (4/14) studies had medium power (between 102 and 278 participants), and 7% (1/14) of studies had high power (>278 participants). More specifics on outcomes (including additional outcomes from each RCT) are provided in Multimedia Appendix 2.

Figure 1. Summary of findings from randomized controlled trials (RCTs) of Headspace. "Mixed findings" refers to when 2 or more measures were used to evaluate the same outcome domain in an RCT and these measures yielded conflicting results.



RCT of the Calm App

In the single RCT of the Calm app, participants in the intervention arm showed significantly improved stress and mindfulness scores than those in the control arm. In total, 88 participants were included in the analyses, and the study had low power (<102 participants).

Risk of Bias

RCTs of the Headspace App

Overall, 100% (14/14) of studies were judged as having a low risk of bias on two of the Cochrane criteria: *random sequence generation* and *allocation concealment*. On the *blinding ofparticipants and personnel* domain and the *blinding of outcome assessment* domain, 50% (7/14) studies received a rating of low risk and 50% (7/14) received a rating of high risk. On the *incomplete outcomedata* domain, 57% (8/14) of studies received a rating of low risk and 43% (6/14) received a rating of high risk. Finally, on the *selective reporting* domain, 29% (4/14) of studies received a rating of low risk and 71% (10/14) of studies received a rating of unclear. These 10 studies were not preregistered, so we could not determine if the authors engaged in selective reporting of outcomes. For the *other bias* category, 79% (11/14) of studies were rated as having a low risk, 14% (2/14) of them were rated as having a high risk, and 7% (1/14) of them were rated as unclear. Itemized Cochrane risk of bias results can be found in Multimedia Appendix 2.

RCT of the Calm App

The singular RCT of the Calm app was judged as having a low risk of bias on the *random sequence generation* and *allocation concealment* domains. On the *blinding of participants and personnel* and *blinding of outcome assessment* domains, it received a high risk rating. *Incomplete outcome data* was rated

as having a low risk of bias and *selective reporting* was rated as having a high risk of bias. The RCT of Calm had a low risk of bias for the *other bias* category. Itemized Cochrane risk of bias results can be found in Multimedia Appendix 2.

Assessment of COIs

In addition to the standard Cochrane risk of bias categories, we evaluated preregistration and COIs related to the involvement of app companies in the 15 RCTs identified.

RCTs of the Headspace App

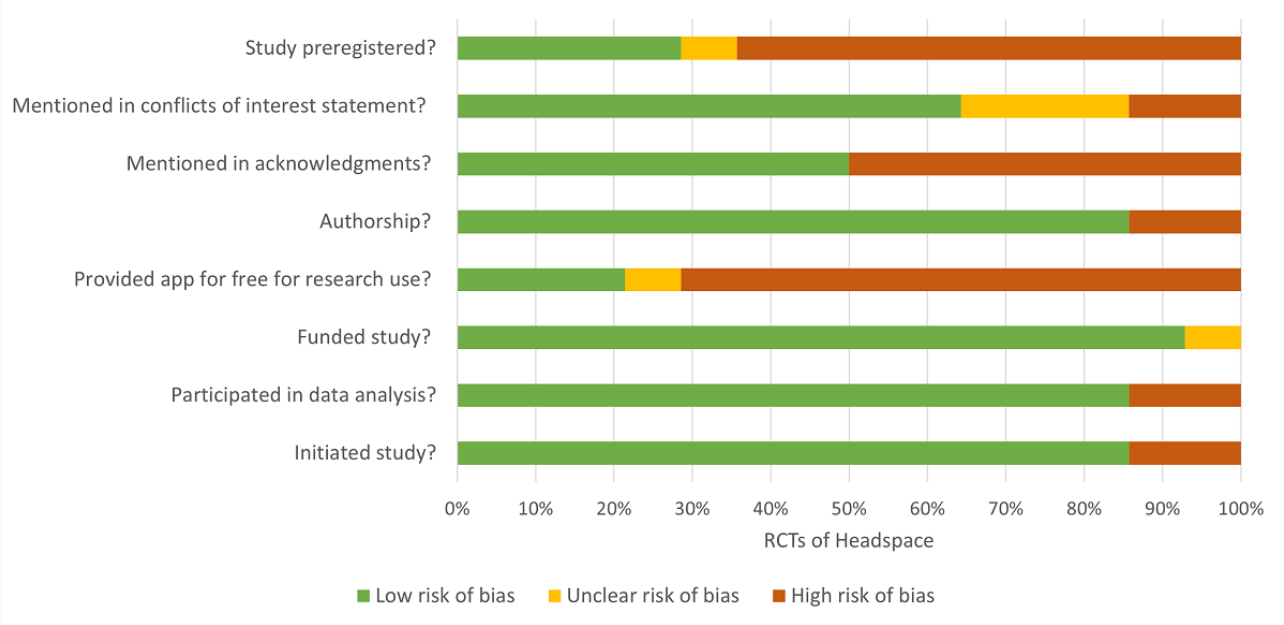
Preregistration

Of the 14 RCTs of Headspace, only 36% (5/14) were preregistered.

Assessment of COIs

Half of the studies (7/14) mentioned a COI in their COI statement that involved Headspace or Calm, 21% (3/14) of them did not include a COI statement, and 29% (4/14) of them explicitly stated that there was no COI. In 14% (2/14) of studies, individuals from Headspace or Calm were involved in the study's conception or execution. Overall, 93% (13/14) of studies were not funded by Headspace, and for 7% (1/14) of studies, it was unclear whether the app company had funded the study. In 14% (2/14) of studies, individuals from Headspace or Calm were involved in data analysis and were included as coauthors. In 71% (10/14) of studies, Headspace Inc provided premium app access at no cost to researchers for participants to use, 21% (3/14) of studies did not use complimentary access from Headspace, and for 7% (1/14) of studies, this usage was unclear. Figure 2 depicts the COI data of these studies.

Figure 2. Evaluation of conflicts of interest in randomized controlled trials (RCTs) of Headspace.



RCT of the Calm App

Preregistration

The 1 RCT of the Calm app was not preregistered.

Assessment of COIs

For the 1 RCT of the Calm app, Calm (the company) was not involved in initiating the study, data analysis, study funding, or authorship. Researchers did not specify in the paper whether Calm provided app use free of charge for this study. The company was not mentioned in the acknowledgments or COI statements.

Discussion

Principal Findings

We performed a systematic review of RCTs evaluating Headspace and Calm, the two most popular mental health apps. First, we evaluated the efficacy of Calm and Headspace. For Calm, additional RCTs will be needed before the question of efficacy can be addressed empirically. For Headspace, our review of RCTs demonstrated that the efficacy findings are inconclusive. RCTs of Headspace showed that Headspace use reliably improved depression. Findings were mixed for mindfulness, well-being, stress, and anxiety, but at least 40% of studies for each of these 4 outcomes showed improvement from the intervention. The studies mostly focused on members of the general population; we found that relatively few studies have examined the efficacy of these apps with clinical samples. Most studies were not powered to detect "small" or "medium" effects. App adherence data were measured inconsistently. Second, our review characterized the risk of bias and COIs in the available evidence. For all studies, lack of preregistration was a main concern for risk of bias. Direct app company involvement in authorship and study procedures was low for both apps. For Headspace papers, the provision of free use of the premium version of the app was another key finding. The single RCT evaluating Calm did not find COIs with regard to the company's involvement of study conduct, analysis, or authorship.

Comparison to Prior Work

Our general discussion of studies in aggregate mainly refer to Headspace owing to the limited number of RCTs identified for Calm. Despite the mixed findings and underpowered studies, we believe that the evidence supporting an intervention should be considered in light of its costs. Even if only a small proportion of individuals who use a mental health app experience symptom reduction attributable to the app, this "small proportion" could include millions of individuals who would not have accessed other forms of evidence-based support [41,42]. Furthermore, users who do not benefit from the apps can discontinue using them with low opportunity costs. Headspace and Calm are unguided self-help apps with relatively lower costs than other kinds of mental health promotion interventions (eg, psychotherapy, medications, and professional coaching). Both Headspace and Calm offer a free version of the app, and the premium versions cost US $13 per month and US $15 per month, respectively (or both offer an annual plan for US $70 per year), which are considerably more affordable than traditional mental health interventions [42] (eg, US $60-$250 per session for private-pay psychotherapy [43]).

Notably, there are several ways in which app usage in RCTs may differ from app usage in naturalistic settings. The trials included in this review focused on college students, healthy volunteers from the general population, and novice meditators. In most of the trials, users were instructed to access specific content within the apps. In contrast, when apps are used in naturalistic settings, users are free to choose the content that they want to access. Additionally, engagement with apps tends to be higher in trials, as investigators can promote engagement through financial incentives, and participants in research trials may feel committed to participating fully in the study [44]. Thus, findings from randomized trials may not fully generalize to app usage in real-world settings. We were unable to draw conclusions on app adherence data owing to variability in measurement. App adherence is a crucial component of understanding the real-world validity of mobile health (mHealth) interventions; hence, adoption of standardized reporting tools is necessary for appropriate evaluation of adherence in future systematic reviews on mHealth interventions [45].

Given the low cost of Headspace, the fact that multiple randomized trials have supported its effectiveness in some samples and individuals who do not benefit from Headspace can discontinue its use with a low opportunity cost, Headspace may be a promising intervention. More evidence on Calm is needed. The funding that Headspace provides promotes the acquisition of empirical evidence on mental health apps, which is positive, but the provision of app use free of charge for research presents several relevant risks for bias. First, there is evidence of a higher potential for bias when people who work at a for-profit company are involved in study design, conduct, and analysis [46]. Second, researchers who are interested in studying psychological interventions or constructs including mindfulness will be more likely to study a mindfulness app being offered free of charge [7]. Other apps that may be equally or more effective may not be able to financially support research in this way. This difference accounts for the imbalance between Headspace and Calm with respect to the number of published RCTs we found in our review. The resulting plethora of research on one app in comparison to other mental health apps may lead consumers to believe that that app is the "best" intervention or the most evidence based, when the lack of studies on other mental health apps is potentially attributable to financial inaccessibility.

Future Directions

Our review demonstrates several gaps to be addressed by future research on popular mental health apps. First, future research could examine for whom these apps are effective, and how much of the intervention someone must complete to achieve desired positive effects. Precision mental health techniques could be used to identify individuals who are most likely to respond to apps, minimum intervention time, and what content might be most helpful for a given individual.

Future randomized trials of mental health apps could also evaluate the effectiveness of apps when users are instructed to use the app freely rather than when they are instructed to access specific preselected modules within the app, particularly in naturalistic settings. For example, Headspace gifted free app access to educators during the COVID-19 pandemic [47], and future similar circumstances could provide an opportunity to study app efficacy and engagement.

The variability in outcomes across RCTs prevented us from calculating effect sizes or other statistics in these data, limiting our ability to draw conclusions. Future work could attempt to standardize patient-reported outcomes in clinical trials on mental health apps to enable future comparisons, especially via meta-analysis.

The involvement of app companies in the research process introduces a risk of bias in studies evaluating mental health apps. We recommend that when evaluating an existing intervention that is provided free of charge, researchers should use an active control to demonstrate how Headspace and Calm perform in comparison with alternative apps. With the goal of improving mental health outcomes for users, strategies should be explored to increase the number of open access apps available for research.

Limitations

There are a few important limitations to our study. First, the app market is dynamic, and new apps may increase in popularity rapidly or over time. This study is time bound, since our search was conducted in May 2021 when Headspace and Calm were the most downloaded and widely used mental health apps. This may change by the time of or after publication of this review. Second, this review was not preregistered, and a protocol was not published ahead of time, thus potentially increasing the risk of bias in our review. Third, the disparity in the number of RCTs for Headspace compared to those of Calm limited our ability to investigate the Calm app thoroughly. We were not able to directly compare efficacy and COI variables between Headspace and Calm owing to only finding 1 RCT of Calm. Since the number of RCTs for Headspace and Calm was beyond our control, we discussed the results without comparing the two apps and encouraged additional RCTs on Calm. Fourth, we only captured risk of bias and COI information based on what authors

reported; hence, we may not be aware of all potential COIs. Fifth, we reported results on the basis of significance, but significant findings do not necessarily mean that improvements in psychosocial outcomes are clinically significant, and we did not evaluate the data with respect to clinically meaningful differences.

Conclusions

The wide adoption of apps including Headspace and Calm provides an opportunity to address population-level mental health. We hope that this review inspires further work on mental health apps, both adding to the current evidence base on Headspace and Calm and evaluating other mental health apps. We advise clinicians, researchers, and consumers of clinical research to ask similar questions about COIs when consuming research, particularly research evaluating products from for-profit companies using science-based marketing to promote their product. Once a product such as Headspace or Calm is widely used, it can be easily accepted on face value as effective, but we want to inspire other researchers to evaluate the nuances of the evidence base, especially since popular mental health apps are already reaching millions of people each month. If effective apps disseminate widely, they may play an extremely important role in improving mental health and wellness worldwide.

Acknowledgments

NL is funded as an Implementation Science Scholar through the National Heart, Lung, and Blood Institute of the National Institutes of Health (grant 5K12 HL137940-02). The opinions herein represent those of the authors and not necessarily the funders.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) Diagram.
[PDF File (Adobe PDF File), 55 KB - mental_v9i9e40924_app1.pdf]

Multimedia Appendix 2
Cochrane Risk of Bias Summary.
[PNG File , 13 KB - mental_v9i9e40924_app2.png]

References

1. GBD 2019 Mental Disorders Collaborators. Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet Psychiatry 2022 Feb;9(2):137-150 [FREE Full text] [doi: 10.1016/S2215-0366(21)00395-3] [Medline: 35026139]
2. Kazdin AE, Blase SL. Rebooting psychotherapy research and practice to reduce the burden of mental illness. Perspect Psychol Sci 2011 Jan 03;6(1):21-37 [FREE Full text] [doi: 10.1177/1745691610393527] [Medline: 26162113]
3. Andrade LH, Alonso J, Mneimneh Z, Wells JE, Al-Hamzawi A, Borges G, et al. Barriers to mental health treatment: results from the WHO World Mental Health surveys. Psychol Med 2013 Aug 09;44(6):1303-1317. [doi: 10.1017/s0033291713001943]
4. Gulliver A, Griffiths KM, Christensen H. Perceived barriers and facilitators to mental health help-seeking in young people: a systematic review. BMC Psychiatry 2010 Dec 30;10:113 [FREE Full text] [doi: 10.1186/1471-244X-10-113] [Medline: 21192795]
5. Wasil AR, Gillespie S, Shingleton R, Wilks CR, Weisz JR. Examining the Reach of Smartphone Apps for Depression and Anxiety. Am J Psychiatry 2020 May 01;177(5):464-465. [doi: 10.1176/appi.ajp.2019.19090905] [Medline: 32354266]
6. Josephine K, Josefine L, Philipp D, David E, Harald B. Internet- and mobile-based depression interventions for people with diagnosed depression: a systematic review and meta-analysis. J Affect Disord 2017 Dec 01;223:28-40. [doi: 10.1016/j.jad.2017.07.021] [Medline: 28715726]
7. Lau N, O'Daffer A, Colt S, Yi-Frazier JP, Palermo TM, McCauley E, et al. Android and iPhone Mobile Apps for Psychosocial Wellness and Stress Management: Systematic Search in App Stores and Literature Review. JMIR Mhealth Uhealth 2020 May 22;8(5):e17798 [FREE Full text] [doi: 10.2196/17798] [Medline: 32357125]
8. Wang X, Markert C, Sasangohar F. Investigating popular mental health mobile application downloads and activity during the COVID-19 pandemic. Hum Factors 2021 Mar 07:1-12. [doi: 10.1177/00187208211998110] [Medline: 33682467]
9. Chapple C. Downloads of top English-language mental wellness apps surged by 2 million in April amid COVID-19 pandemic. Sensor Tower. 2020. URL: https://sensortower.com/blog/top-mental-wellness-apps-april-2020-downloads [accessed 2022-08-28]
10. Chan S. Sensor Tower. 2021. URL: https://sensortower.com/blog/state-of-health-and-fitness-apps-report-2021 [accessed 2022-08-28]

11. Firth J, Torous J, Nicholas J, Carney R, Rosenbaum S, Sarris J. Can smartphone mental health interventions reduce symptoms of anxiety? A meta-analysis of randomized controlled trials. J Affect Disord 2017 Aug 15;218:15-22 [FREE Full text] [doi: 10.1016/j.jad.2017.04.046] [Medline: 28456072]
12. Linardon J, Wade T, de la Piedad Garcia X, Brennan L. The efficacy of cognitive-behavioral therapy for eating disorders: A systematic review and meta-analysis. J Consult Clin Psychol 2017 Nov;85(11):1080-1094 [FREE Full text] [doi: 10.1037/ccp0000245] [Medline: 29083223]
13. Linardon J, Cuijpers P, Carlbring P, Messer M, Fuller-Tyszkiewicz M. The efficacy of app-supported smartphone interventions for mental health problems: a meta-analysis of randomized controlled trials. World Psychiatry 2019 Oct;18(3):325-336 [FREE Full text] [doi: 10.1002/wps.20673] [Medline: 31496095]
14. Open Science Collaboration. PSYCHOLOGY. Estimating the reproducibility of psychological science. Science 2015 Aug 28;349(6251):aac4716. [doi: 10.1126/science.aac4716] [Medline: 26315443]
15. Hansen WB, Scheier LM. Specialized smartphone intervention apps: review of 2014 to 2018 NIH funded grants. JMIR Mhealth Uhealth 2019 Jul 29;7(7):e14655 [FREE Full text] [doi: 10.2196/14655] [Medline: 31359866]
16. Simmons J, Nelson L, Simonsohn U. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychol Sci 2011 Nov;22(11):1359-1366. [doi: 10.1177/0956797611417632] [Medline: 22006061]
17. Day S, Shah V, Kaganoff S, Powelson S, Mathews SC. Assessing the Clinical Robustness of Digital Health Startups: Cross-sectional Observational Analysis. J Med Internet Res 2022 Jun 20;24(6):e37677 [FREE Full text] [doi: 10.2196/37677] [Medline: 35723914]
18. John L, Loewenstein G, Prelec D. Measuring the prevalence of questionable research practices with incentives for truth telling. Psychol Sci 2012 May 01;23(5):524-532 [FREE Full text] [doi: 10.1177/0956797611430953] [Medline: 22508865]
19. Turner E, Matthews A, Linardatos E, Tell R, Rosenthal R. Selective publication of antidepressant trials and its influence on apparent efficacy. N Engl J Med 2008 Jan 17;358(3):252-260 [FREE Full text] [doi: 10.1056/nejmsa065779]
20. Driessen E, Hollon SD, Bockting CLH, Cuijpers P, Turner EH. Does publication bias inflate the apparent efficacy of psychological treatment for major depressive disorder? A systematic review and meta-analysis of US National Institutes of Health-funded trials. PLoS One 2015 Sep 30;10(9):e0137864 [FREE Full text] [doi: 10.1371/journal.pone.0137864] [Medline: 26422604]
21. Faul F, Erdfelder E, Buchner A, Lang A. Statistical power analyses using G*Power 3.1: tests for correlation and regression analyses. Behav Res Methods 2009 Nov;41(4):1149-1160 [FREE Full text] [doi: 10.3758/brm.41.4.1149]
22. Higgins JPT, Altman DG, Gøtzsche PC, Jüni P, Moher D, Oxman AD, Cochrane Bias Methods Group, Cochrane Statistical Methods Group. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. BMJ 2011 Oct 18;343:d5928-d5928 [FREE Full text] [doi: 10.1136/bmj.d5928] [Medline: 22008217]
23. Allen C, Mehler D. Open science challenges, benefits and tips in early career and beyond. PLoS Biol 2019 May 1;17(5):e3000246 [FREE Full text] [doi: 10.1371/journal.pbio.3000246]
24. Turner E, Matthews A, Linardatos E, Tell R, Rosenthal R. Selective Publication of Antidepressant Trials and Its Influence on Apparent Efficacy. N Engl J Med 2008 Jan 17;358(3):252-260 [FREE Full text] [doi: 10.1056/nejmsa065779]
25. Simmons J, Nelson L, Simonsohn U. Pre‐registration: why and how. J Consum Psychol 2021 Feb 15;31(1):151-162 [FREE Full text] [doi: 10.1002/jcpy.1208]
26. Bennike I, Wieghorst A, Kirk U. Online-based mindfulness training reduces behavioral markers of mind wandering. J Cogn Enhanc 2017 Apr 25;1(2):172-181. [doi: 10.1007/s41465-017-0020-9]
27. Björkstrand J, Schiller D, Li J, Davidson P, Rosén J, Mårtensson J, et al. The effect of mindfulness training on extinction retention. Sci Rep 2019 Dec 27;9(1):19896 [FREE Full text] [doi: 10.1038/s41598-019-56167-7] [Medline: 31882606]
28. Bostock S, Crosswell AD, Prather AA, Steptoe A. Mindfulness on-the-go: Effects of a mindfulness meditation app on work stress and well-being. J Occup Health Psychol 2019 Feb;24(1):127-138 [FREE Full text] [doi: 10.1037/ocp0000118] [Medline: 29723001]
29. Champion L, Economides M, Chandler C. The efficacy of a brief app-based mindfulness intervention on psychosocial outcomes in healthy adults: a pilot randomised controlled trial. PLoS One 2018;13(12):e0209482 [FREE Full text] [doi: 10.1371/journal.pone.0209482] [Medline: 30596696]
30. DeSteno D, Lim D, Duong F, Condon P. Meditation inhibits aggressive responses to provocations. Mindfulness 2017 Nov 2;9(4):1117-1122. [doi: 10.1007/s12671-017-0847-2]
31. Economides M, Martman J, Bell MJ, Sanderson B. Improvements in stress, affect, and irritability following brief use of a mindfulness-based smartphone app: a randomized controlled trial. Mindfulness (N Y) 2018;9(5):1584-1593 [FREE Full text] [doi: 10.1007/s12671-018-0905-4] [Medline: 30294390]
32. Flett J, Hayne H, Riordan B, Thompson L, Conner T. Mobile mindfulness meditation: a randomised controlled trial of the effect of two popular apps on mental health. Mindfulness 2018 Oct 31;10(5):863-876. [doi: 10.1007/s12671-018-1050-9]
33. Howells A, Ivtzan I, Eiroa-Orosa F. Putting the 'app' in happiness: a randomised controlled trial of a smartphone-based mindfulness intervention to enhance wellbeing. J Happiness Stud 2014 Oct 29;17(1):163-185. [doi: 10.1007/s10902-014-9589-1]

34. Kubo A, Kurtovich E, McGinnis M, Aghaee S, Altschuler A, Quesenberry C, et al. A randomized controlled trial of mHealth mindfulness intervention for cancer patients and informal cancer caregivers: a feasibility study within an integrated health care delivery system. Integr Cancer Ther 2019 May 16;18:153473541985063. [doi: 10.1177/1534735419850634]
35. Lim D, Condon P, DeSteno D. Mindfulness and compassion: an examination of mechanism and scalability. PLoS One 2015;10(2):e0118221 [FREE Full text] [doi: 10.1371/journal.pone.0118221] [Medline: 25689827]
36. Noone C, Hogan MJ. A randomised active-controlled trial to examine the effects of an online mindfulness intervention on executive control, critical thinking and key thinking dispositions in a university student sample. BMC Psychol 2018 Apr 05;6(1):13 [FREE Full text] [doi: 10.1186/s40359-018-0226-3] [Medline: 29622047]
37. Quinones C, Griffiths MD. Reducing compulsive Internet use and anxiety symptoms via two brief interventions: a comparison between mindfulness and gradual muscle relaxation. J Behav Addict 2019 Sep 01;8(3):530-536 [FREE Full text] [doi: 10.1556/2006.8.2019.45] [Medline: 31505967]
38. Rosen K, Paniagua S, Kazanis W, Jones S, Potter J. Quality of life among women diagnosed with breast cancer: a randomized waitlist controlled trial of commercially available mobile app-delivered mindfulness training. Psycho-Oncology 2018 Jun 01;27(8):2023-2030. [doi: 10.1002/pon.4764]
39. Yang E, Schamber E, Meyer RML, Gold JI. Happier healers: randomized controlled trial of mobile mindfulness for stress management. J Altern Complement Med 2018 May;24(5):505-513. [doi: 10.1089/acm.2015.0301] [Medline: 29420050]
40. Huberty J, Green J, Glissmann C, Larkey L, Puzia M, Lee C. Efficacy of the mindfulness meditation mobile app "Calm" to reduce stress among college students: randomized controlled trial. JMIR Mhealth Uhealth 2019 Jun 25;7(6):e14273 [FREE Full text] [doi: 10.2196/14273] [Medline: 31237569]
41. Wasil A, Venturo-Conerly K, Shingleton R, Weisz JR. A review of popular smartphone apps for depression and anxiety: Assessing the inclusion of evidence-based content. Behav Res Ther 2019 Dec;123:103498. [doi: 10.1016/j.brat.2019.103498] [Medline: 31707224]
42. Baumel A, Muench F, Edan S, Kane JM. Objective user engagement with mental health apps: systematic search and panel-based usage analysis. J Med Internet Res 2019 Sep 25;21(9):e14567 [FREE Full text] [doi: 10.2196/14567] [Medline: 31573916]
43. Fleming T, Bavin L, Lucassen M, Stasiak K, Hopkins S, Merry S. Beyond the trial: systematic review of real-world uptake and engagement with digital self-help interventions for depression, low mood, or anxiety. J Med Internet Res 2018 Jun 06;20(6):e199 [FREE Full text] [doi: 10.2196/jmir.9275] [Medline: 29875089]
44. Mehta A, Niles A, Vargas J, Marafon T, Couto DD, Gross JJ. Acceptability and effectiveness of artificial intelligence therapy for anxiety and depression (Youper): longitudinal observational study. J Med Internet Res 2021 Jun 22;23(6):e26771 [FREE Full text] [doi: 10.2196/26771] [Medline: 34155984]
45. Liptáková S, Světlák M, Matis J, Slezáčková A, Rastislav R. Signing up is not yet mindfulness practice: A systematic review of adherence to eHealth and mHealth mindfulness-based programs in the pre-pandemic period. CESK PSYCHOL 2022 Jun 30;66(3):233-254. [doi: 10.51561/cspsych.66.3.233]
46. Tierney WM, Meslin EM, Kroenke K. Industry support of medical research: important opportunity or treacherous pitfall? J Gen Intern Med 2016 Feb 26;31(2):228-233 [FREE Full text] [doi: 10.1007/s11606-015-3495-z] [Medline: 26307387]
47. How is Headspace helping those impacted by COVID-19? Headspace. URL: https://help.headspace.com/hc/en-us/articles/360045857254-How-is-Headspace-helping-those-impacted-by-COVID-19- [accessed 2022-08-10]

Abbreviations

COI: conflict of interest
mHealth: mobile health
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RCT: randomized controlled trial

Edited by J Torous; submitted 09.07.22; peer-reviewed by M Svetlak, H van Marwijk; comments to author 01.08.22; revised version received 15.08.22; accepted 28.08.22; published 20.09.22.

Please cite as:
O'Daffer A, Colt SF, Wasil AR, Lau N
Efficacy and Conflicts of Interest in Randomized Controlled Trials Evaluating Headspace and Calm Apps: Systematic Review
JMIR Ment Health 2022;9(9):e40924
URL: https://mental.jmir.org/2022/9/e40924
doi:10.2196/40924
PMID:36125880

Review

Content and Effectiveness of Web-Based Treatments for Online Behavioral Addictions: Systematic Review

Jennifer J Park[1], BHSc (Hons); Daniel L King[2], PhD; Laura Wilkinson-Meyers[1], PhD; Simone N Rodda[3], PhD

[1]School of Population Health, The University of Auckland, Auckland, New Zealand

[2]College of Education, Psychology & Social Work, Flinders University, Adelaide, Australia

[3]Department of Psychology and Neuroscience, Auckland University of Technology, Auckland, New Zealand

Corresponding Author:
Jennifer J Park, BHSc (Hons)
School of Population Health
The University of Auckland
Building 507, 22-30 Park Ave, Grafton
Auckland, 1023
New Zealand
Phone: 64 210 822 6685
Email: jpar956@aucklanduni.ac.nz

Abstract

Background: Very few people seek in-person treatment for online behavioral addictions including gaming and gambling or problems associated with shopping, pornography use, or social media use. Web-based treatments have the potential to address low rates of help seeking due to their convenience, accessibility, and capacity to address barriers to health care access (eg, shame, stigma, cost, and access to expert care). However, web-based treatments for online behavioral addictions have not been systematically evaluated.

Objective: This review aimed to systematically describe the content of web-based treatments for online behavioral addictions and describe their therapeutic effectiveness on symptom severity and consumption behavior.

Methods: A database search of MEDLINE, Embase, PsycInfo, Web of Science, Cochrane Central Register of Controlled Trials, and Google Scholar was conducted in June 2022. Studies were eligible if the study design was a randomized controlled trial or a pre-post study with at least 1 web-based intervention arm for an online behavioral addiction and if the study included the use of a validated measure of problem severity, frequency, or duration of online behavior. Data on change techniques were collected to analyze intervention content, using the Gambling Intervention System of CharacTerization. Quality assessment was conducted using the Effective Public Health Practice Project Quality Assessment Tool.

Results: The review included 12 studies with 15 intervention arms, comprising 7 randomized controlled trials and 5 pre-post studies. The primary focus of interventions was gaming (n=4), followed by internet use inclusive of screen time and smartphone use (n=3), gambling (n=3), and pornography (n=2). A range of different technologies were used to deliver content, including websites (n=6), email (n=2), computer software (n=2), social media messaging (n=1), smartphone app (n=1), virtual reality (n=1), and videoconferencing (n=1). Interventions contained 15 different change techniques with an average of 4 per study. The techniques most frequently administered (>30% of intervention arms) were cognitive restructuring, relapse prevention, motivational enhancement, goal setting, and social support. Assessment of study quality indicated that 7 studies met the criteria for moderate or strong global ratings, but only 8 out of 12 studies evaluated change immediately following the treatment. Across included studies, two-thirds of participants completed after-treatment evaluation, and one-quarter completed follow-up evaluation. After-intervention evaluation indicated reduced severity (5/9, 56%), frequency (2/3, 67%), and duration (3/7, 43%). Follow-up evaluation indicated that 3 pre-post studies for gaming, gambling, and internet use demonstrated reduced severity, frequency, and duration of consumption. At 3-month evaluation, just 1 pre-post study indicated significant change to mental health symptoms.

Conclusions: Web-based treatments for online behavioral addictions use an array of mechanisms to deliver cognitive and behavioral change techniques. Web-based treatments demonstrate promise for short-term reduction in symptoms, duration, or frequency of online addictive behaviors. However, there is limited evidence on the effectiveness of web-based treatments over the longer term due to the absence of controlled trials.

(JMIR Ment Health 2022;9(9):e36662) doi:10.2196/36662

KEYWORDS

systematic review; gambling; gaming; internet intervention; pornography; treatment; social media

Introduction

There is growing recognition that some individuals engage in problematic and potentially addictive behaviors across a wide range of online activities, including gaming, gambling, shopping, social media use, and pornography use [1-3]. The *International Classification of Diseases 11th Revision* (ICD-11) includes 2 behavioral addictions associated with gaming and gambling [4,5]. Gambling disorder was the first recognized behavioral addiction and is characterized by gambling to escape negative mood, tolerance, repeated unsuccessful attempts to change, and gambling despite negative consequences. Gambling disorder encompasses both land-based activities as well as online casino gambling and web-based betting on sports and racing, which have increased for adults and adolescents over recent years [6,7]. Gaming disorder has characteristics that are consistent with gambling disorder, but there is less focus on money, chasing losses, and financial impacts of gambling on other people. The ICD-11 describes gaming disorder as a condition involving impaired control (eg, over the onset, duration, frequency, and context of play), increasing prioritization of gaming over other activities and life interests, and continued involvement despite negative consequences (eg, impairment in social, educational, and occupational functioning). Some online behavioral addictions are not yet identified under any diagnostic classification of the ICD-11 (eg, pornography and social media use), and some excessive behaviors may be encapsulated by existing categories (eg, online shopping within compulsive buying disorder). Although the literature on different classes of behavioral addictions is still developing, it is often argued that there is a need for evidence-based interventions and other countermeasures to prevent and reduce problematic use.

The literature on interventions for online behavioral addictions has generally been focused on in-person treatment which is intensive and typically involves 6 or more weekly sessions [8,9]. A recent review of treatment for gaming disorder reported it was predominantly psychotherapeutic, face-to-face, and targeted to those with more severe problems [10]. At the same time, reviews have tended to focus on in-person treatment studies and excluded web-based options as evidenced by a recent Cochrane review on psychological therapies for gambling [9]. The lack of scholarly attention on web-based interventions may be overlooking an important modality that is accessed by many affected by behavioral addictions. Online behavioral addictions reportedly affects between 1% and 3% of the population [11,12], but help-seeking rates are quite low [13,14]. These findings suggest that either few people want or require help to resolve their problem or that available clinical options are not meeting the needs of the population. Help seeking may be impeded by structural issues such as the homogeneity of available treatments, prohibitive cost and accessibility, or individual barriers like depression, introversion, or a preference for self-management [15-20].

Web-based treatment appears to be a viable alternative to in-person treatment and has demonstrated effectiveness in reducing symptom severity and consumption patterns of addictive behaviors [21]. Web-based treatment has the potential to reach a wider group of help seekers, such as those seeking anonymity, to reduce perceived shame and stigma [20]. Web-based options may also be attractive for their relatively lower cost compared to individual sessions or retreats [10] and for their convenience and flexibility [20,22,23]. Furthermore, these options may be optimally positioned in the online environment (ie, at the site where users are experiencing psychological difficulties) despite concerns around the appropriateness of web-based delivery for online problems [24]. Online delivery may occur via email, websites, social media, apps, online calls, instant messaging, and virtual reality and may involve smartphones, laptops, and computers, among other online devices. Currently, it is unknown how each of these diverse options might be leveraged effectively to deliver mental health services or other public health measures to address the problematic use of online activities and applications.

Reviews on treatment for online addictive behaviors have not yet explicitly focused on the mode of intervention delivery. Past treatment reviews have also tended to be narrow in focus and overlooked the wide variability in the scope of online activities. For example, reviews of online behaviors have examined interventions for problems related to gaming [8,10,25-29], cybersex [30], both internet use and gaming [31-33], internet use and smartphone use [34], and general internet addiction or problematic internet use [2,35-38]. Reviews focused on gambling problems have examined the effectiveness of web-based treatment for prevention [39] and treatment [23,40], but these were not restricted to samples of online gamblers. Only 1 previous review has examined web-based treatments specifically for problematic internet use, reporting on 3 studies and without examining the effectiveness of treatment [36]. This review included the search terms “online intervention,” “eIntervention,” “eTherapy,” and “eHealth,” which meant other forms of web-based treatments such as online psychotherapy, psychoeducation, and self-help were overlooked. Given these limitations and that considerable time (ie, 5 years) had passed since the previous review, it was timely to evaluate the content and effectiveness of web-based treatments for online behavioral addictions.

This systematic review aimed to summarize and critique the available literature on web-based treatments for online behavioral addictions. Specifically, this review aimed to do the following: describe the content of web-based treatments inclusive of any intervention type for online gaming, gambling, shopping, pornography use, social media use, smartphone use, or nonspecified online use; and describe the effectiveness of web-based treatments on severity, duration, or frequency of consumption. Although only gaming and gambling are currently recognized as addictive disorders in the ICD-11, the scope of this review was expanded to include other online activities (social media, pornography, and shopping) that have been proposed to share similarities to these disorders and which have been studied using addiction-based approaches [1]. It is

acknowledged that, over time, there may be important changes to the classification of these behaviors as disorders, including their status of inclusion in *The Diagnostic and Statistical Manual of Mental Disorders* and ICD nomenclatural systems.

Methods

This systematic review was registered and published on PROSPERO (International Prospective Register of Systematic Reviews; registration code CRD42021224595) and followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [41].

Eligibility Criteria

Studies were selected on the basis of the following six inclusion criteria: (1) at least 1 intervention arm was web-based; (2) the behavioral addiction was predominantly a web-based activity and involved gaming, gambling, shopping, pornography use, social media use, smartphone use, or nonspecified internet use; (3) the intervention was intended to reduce the severity, frequency, or duration of the behavioral addiction inclusive of mild and moderate problems; (4) the behavioral addiction was assessed with a validated screen, self-report, or participant registration in a treatment program; (5) the study had a comparison group including a passive or active control or comparative intervention, or was a pre-post study; and (6) there was at least 1 evaluation conducted after the intervention. Unpublished reports, conference papers, presentations, theses, posters, opinion pieces, letters, or protocols were excluded. Studies were also excluded based on the following four criteria: (1) interventions not targeted at web-based behaviors, such as land-based electronic gaming-machine gambling; (2) web-based behavior considered to not be addictive (eg, cyberbullying); (3) prevention programs designed to reduce the risk of future harm or where there were no reported problems; and (4) where the majority of the intervention content was not web-based.

Identification and Selection of Studies

A database search of MEDLINE, Embase, PsycInfo, Web of Science, Cochrane Central Register of Controlled Trials, and Google Scholar was conducted in June 2022. The search strategy is provided in Multimedia Appendix 1. The search was limited to studies in English language, published in the last 22 years (ie, 2000-2022), and available in full text. To identify potential studies that met the inclusion criteria, recent systematic reviews, reference lists within these reviews, and reference lists of included studies were also searched. Titles and abstracts of the studies returned from the search strategy were screened independently by 2 researchers (JJP and another researcher) against the inclusion and exclusion criteria. The full text of the studies returned from this process was also screened independently by the 2 aforementioned researchers with a third researcher (SNR) involved to resolve any disagreements.

Data Extraction and Analysis

A structured data extraction form was developed for the study in Microsoft Excel. The data extraction included information on the behavioral addiction type; recruitment and study methods; participant demographics; outcome measures; intervention characteristics; mode of intervention delivery; comparison conditions; and outcomes for frequency, duration, severity, and mental health. To systematically identify the content of interventions, each paper was assessed against the 18 categories of change techniques identified in the Gambling Intervention System of CharacTerization (GIST-1) [42]. The GIST-1 provides an efficient way to classify change techniques sourced from published articles as opposed to assessing the smaller behavior change techniques reported in treatment manuals [43]. Two independent coders (JJP and SNR) assessed each article for the presence of the 18 GIST-1 categories and extracted qualitative data describing each technique.

Quality Assessment

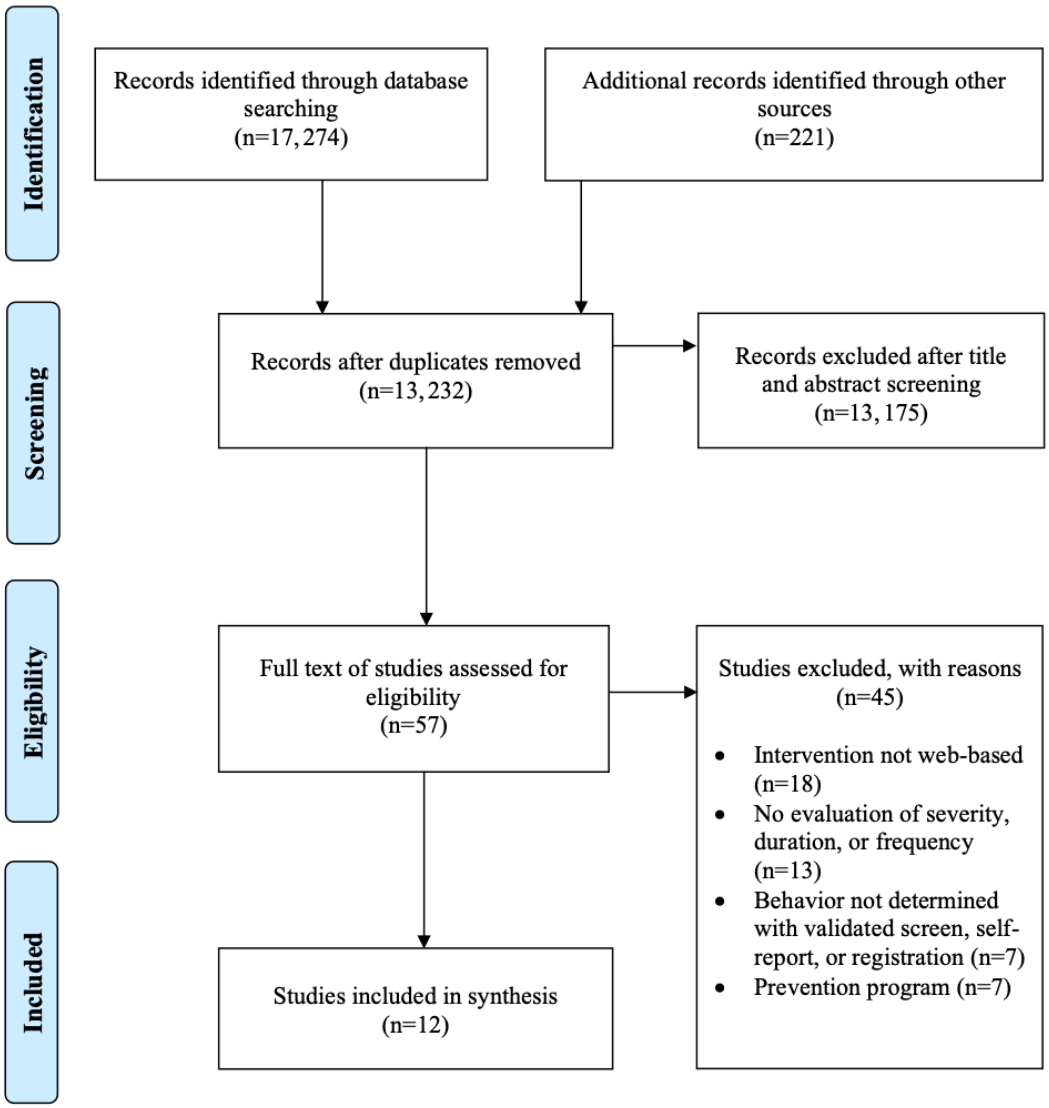
Each study was assessed for quality using the Effective Public Health Practice Project (EPHPP) Quality Assessment Tool for Quantitative Studies [44]. The EPHPP assesses each study for selection bias, study design, confounders, blinding, data collection method, and study attrition. Each component was rated as strong, moderate, or weak by 2 independent reviewers (JJP and SNR). Each included study was then given a global rating of strong (no weak ratings), moderate (1 weak rating), or weak (2 or more weak ratings).

Results

Search Results and Flow Diagram

The search yielded a total of 17,274 studies which included the results of the following 6 databases: MEDLINE (n=2448), Embase (n=3410), PsycInfo (n=2872), Web of Science (n=5750), Cochrane Central Register of Controlled Trials (n=2630), and Google Scholar (n=164). After accounting for duplicates, there were 13,232 studies remaining, of which 13,175 were removed following the review of the title and abstracts of studies against the inclusion criteria (see Figure 1). There was a high number of records requiring screening because of terms such as “internet” and “social media.” The remaining 57 studies were reviewed in full to examine their eligibility for inclusion, which excluded 45 studies. A total of 12 studies with 15 intervention arms, published between 2010 and 2021, were identified for inclusion in the review. The included studies reported on 2218 participants, with individual study sample sizes ranging from 10 to 1122 (mean 184.8, SD 294.3).

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram of study selection.



Study Characteristics

Multimedia Appendix 2 presents a summary of included studies. Of the 12 included studies, 7 were randomized controlled trials (RCTs) and 5 were pre-post studies without randomization. Studies recruited participants from Europe (n=8), Asia (n=4), North America (n=3), and Oceania (n=1). Studies predominantly recruited from the community (n=10) via social media, treatment or industry websites, online panels, or online message boards. The primary focus of interventions was gaming (n=4), followed by internet use inclusive of screentime and smartphone use (n=3), gambling (n=3), and pornography (n=2). A range of different technologies were used to deliver content, including websites (n=6), email (n=2), computer software (n=2), social media messaging (n=1), smartphone app (n=1), virtual reality (n=1), and videoconferencing (n=1).

The average age of participants was 33.9 (SD 10.9) years old, and the percentage of males ranged from 10% to 100% (mean 71.8%, SD 31.7%). Almost all participants met the cutoff for problematic behavior, with 8 studies including only people with current problems and 3 studies reporting that the majority had a problem (70%-92%). Participant engagement with the addictive behavior was reported for duration (sessions, days, and weeks) as well as frequency. The average session duration was 57 minutes [45,46], and when measured over 1 week, the average was 27 hours [47-51]. There were 2 studies for internet reduction that reported an average of 5.5 hours of screen time per day [52,53]. The average frequency of engagement was 6 times per week [20,45,46], with 2 studies involving gamblers reporting a frequency of 13 times a month for internet gambling [54] and another study reporting 62 sessions a month for online poker [55].

Intervention Content

Intervention content was examined in 15 web-based intervention arms from the 12 included studies. To determine the exact content of interventions, the components were assessed and coded into the GIST-1 framework of change techniques [42]. A total of 17 different change techniques were identified (Multimedia Appendix 3). The average number of change techniques per study was 4, with a range of 1 to 10 different techniques. The change techniques most frequently administered (>30% of arms) were cognitive restructuring, relapse prevention, motivational enhancement, goal setting, and social support. No

studies included imaginal desensitization or financial management which were previously identified in the GIST-1.

Eight studies (ten arms) reported the delivery of cognitive restructuring [45,51,54,55], cognitive bias modification (CBM) [47,50], exposure therapy [49], or mindfulness [53]. Cognitive restructuring prompted participants to identify, challenge, and replace automatic negative thoughts associated with gaming, gambling, pornography, and nonspecified internet use. Studies also identified triggers and beliefs mediating the relationship between situations and subsequent addictive behavior. Two studies delivered CBM with the aim of altering automatic responses to gaming stimuli. These studies delivered CBM using a device similar to games, where participants used a joystick to push away gaming cues and pull forward neutral or positive associations. Only 1 study delivered exposure therapy that aimed to reduce gaming via virtual reality technology. Exposure therapy involved repeated exposure to high-risk situations, such as scenes from popular games paired with aversion-inducing noise (eg, siren). Just 1 study included mindfulness activities, which were delivered via messaging across 7 days. Participants were prompted to focus their attention on the present moment through engagement with pleasurable activities, including physical activity, breathing, eating, and letting go of disruptive thoughts.

Six studies (eight arms) delivered problem solving [45,48,52], relapse prevention [45,48,51,55], or social skills training [54]. Problem solving prompted participants to identify high-risk situations or triggers that were barriers to sustained behavior change. Participants were also prompted to develop action plans and if-then plans for addressing barriers to change. Relapse prevention prompted participants to review previous goals and plans on what worked well or needed improvement with the view to make plans and prevent future relapse. Just 1 study delivered social skills training for pornography reduction, which focused on improving communication skills and strengthening relationships.

Five studies (five arms) reported the delivery of stimulus control [46,53,56], behavioral substitution [45], or self-monitoring [45,48], and five studies (five arms) delivered social support [45,48,52,54,55]. Stimulus control involved periods of exclusion from online gambling venues or reducing prompts, inclusive of removing notifications, placing the phone out of sight, or turning it off and establishing phone-free periods during the day (eg, before sleep). Conversely, behavioral substitution involved adding pleasurable activities into everyday life. Self-monitoring involved tracking consumption or mood against a self-constructed plan. Social support was provided by clinical or nonclinical professionals who prompted engagement with the intervention and, in 2 cases, delivered the content via videoconferencing or email. Just 2 studies provided peer support, with 1 offering online forums and another integrating fictional characters discussing lived experience within cognitive behavioral therapy (CBT) modules.

Five studies (six arms) reported the delivery of motivational enhancement [45,48,51,52], decisional balance [45,48,51], or goal setting [45,48,51,53]. Motivational enhancement aimed to reduce consumption or increase help seeking by increasing readiness to change. Studies administered motivational interviewing techniques through person-to-person exchanges via videoconferences or nonclinical project support. Motivational enhancement was included as the first module of CBT in 1 study, and another study assessed readiness as a method of tailoring CBT. Three of the studies administering motivational enhancement also offered decisional balance where participants considered the advantages and disadvantages of consumption and reasons for change. Studies that included a goal setting activity prompted participants to establish goal intentions, inclusive of frequency and duration of gaming, pornography use, internet use, and smartphone use.

Seven studies (eight arms) reported the delivery of information gathering [51,52,54], information provision [48,52], feedback on assessment [45,51,53,55], or social comparison [51,55]. Information gathering explored the development of the problem, family history, motives for the addictive behavior, past change attempts, and an assessment of comorbid psychiatric disorders. Information provision included guidelines for reduction and tailored information on support options. Feedback on assessment included a single written and visual report on consumption patterns and severity of addiction, and repeated feedback delivered across 7 days. Two studies provided extended assessment feedback to detail how each individual's results compared with people of similar age and gender.

Intervention Effectiveness

Intervention effectiveness was determined by change in problem severity, duration of use, or frequency (see Multimedia Appendix 2). As presented in the following sections, the review also examined change in mental health or psychosocial functioning.

Problem Severity

Ten studies examined problem severity, including six RCTs and four pre-post studies. Nine studies conducted after-treatment evaluation, where two RCTs and three pre-post studies indicated reduced problem severity for internet use [51,52], gaming [48], pornography use [45], and smartphone use [53]. One study compared web-based exposure therapy against in-person CBT and reported a reduction in symptom severity and no difference between treatments [49]. Three studies reported no change in internet gambling [46,55] or gaming [47] after treatment. Five studies conducted follow-up evaluation, where three pre-post studies reported reduced severity of gaming [48], gambling [56], and internet use [52]. Two studies reported no change in severity of internet gambling at follow-up evaluation [46,55].

Duration

Eight studies assessed duration, including four RCTs and four pre-post studies. Seven studies conducted after-treatment evaluation, where two RCTs reported reduced duration compared with a control group for gaming [50] and internet use [51], and one pre-post study indicated reduced internet use [52]; the remaining four studies indicated no change in duration after treatment [45,46,53] or did not measure change immediately after treatment [48]. Four studies conducted follow-up evaluation, where three pre-post studies reported reduced duration of internet use [52], gaming [48], and gambling [56].

One internet gambling reduction study reported no change in duration at follow-up evaluation [46].

Frequency

Five studies assessed frequency, including two RCTs and three pre-post studies. Three studies conducted after-treatment evaluation, where one RCT and one pre-post study indicated reduced frequency of pornography use [45,54]. One study reported no change [55] or did not measure frequency immediately after treatment [48,56]. Three studies conducted follow-up evaluation, where two pre-post studies indicated reduced frequency of gaming [48] and gambling [56]. One RCT indicated no change in frequency of internet gambling at follow-up evaluation [55].

Mental Health

Three studies assessed mental health or psychosocial functioning, including one RCT and two pre-post studies. One RCT for gaming demonstrated a reduction in anxiety symptoms after treatment, but not for depression [47]. Two pre-post studies demonstrated an increase in well-being for smartphone use and gaming [48,53] and a reduction in psychological distress for gaming [48]. One pre-post study for gaming conducted follow-up evaluation which indicated improved psychological distress and well-being [48].

Assessment of Study Quality

On the EPHPP Quality Assessment Tool, 7 out of 12 studies scored a “moderate” or “strong” global rating (see Multimedia Appendix 2). There were 4 studies that had a “weak” global rating due to selection bias, confounds, and low study retention. Just 2 of 12 studies had an associated protocol or registered their study with a trials board [45,55]. Participant retention after treatment was 64.8% (SD 37.5%) with a range of 11% to 100%. The lowest retention was found in 2 gambling studies with 1 on web-based self-exclusion (11%) and 1 delivering CBT to online poker players who were not actively seeking help (15%). One study administering motivational interviewing by videoconferencing reported that almost half of study participants did not complete after treatment evaluation. Only 5 of 12 studies completed follow-up evaluation that was most frequently 3 months [46,48,52,55], with 1 study conducting a 12-month follow-up evaluation [56]. The average follow-up retention rate was 24.0% (SD 30.7%) with a range of 8% to 70%.

Discussion

This systematic review aimed to summarize and critique the available literature on web-based treatments for online behavioral addictions. The review described and evaluated 12 studies that administered web-based treatments for problems related to online gaming, gambling, pornography, and internet or smartphone use. Treatment was delivered via a range of different technologies inclusive of websites, email, computer software, social media messaging, smartphone apps, virtual reality, and videoconferencing. Treatment delivered an average of 4 different change techniques and, like previous studies involving in-person treatment [10,26,35,42], the most-employed change techniques were cognitive restructuring, relapse prevention, motivational enhancement, goal setting, and social support. The least-used techniques have demonstrated effectiveness for other addictive behaviors or in-person delivery, including exposure therapy, social comparison, feedback on assessment, self-monitoring, and mindfulness. These findings suggest an opportunity to enhance or develop new intervention types that incorporate these techniques.

This review described the effectiveness of web-based treatments on severity, duration, or frequency of consumption. Immediately following treatment, participant evaluation indicated that 5 out of 9 studies that evaluated problem severity reported significant improvements for gaming, pornography, and internet or smartphone use. Out of 7 studies that conducted after-treatment evaluation of duration, 3 reported reduced gaming and internet use. Out of 3 studies that conducted after-treatment evaluation of frequency, 2 reported reduced pornography use. Just 5 of 12 studies conducted follow-up evaluation, and this was most frequently 3 months with one 12-month evaluation. Follow-up indicated that treatment was effective at improving symptoms, duration, or frequency. However, 4 out of 5 studies that conducted follow-up evaluation included completers only, with just 1 RCT for gambling [55] using intent-to-treat analysis which indicated no effect of the intervention. Taken together, this evidence suggests findings should be treated with caution given studies retained just 1 in 4 participants. Easy access is related to high attrition rates because people can easily step away from treatment without interacting with another person [57,58]. One pre-post study on gaming [48] reported a retention rate of 70%, and this study had addressed the risk of attrition by including a coach for advice and support during engagement with the intervention. Future studies should investigate methods such as support or other mechanisms like incentives for increasing retention in web-based treatment for online addictions.

Participants in the included studies were predominantly male and aged around 25 years old. Research indicates that being male and more frequently engaged in addictive online activities is associated with an increased risk of online addictions (gaming disorder, gambling disorder, compulsive buying disorder, and issues related to pornography use and social media) [59], which suggests that most online interventions had appropriate target groups. Participants reported various online intervention components (not content) that were important or helpful, such as privacy and convenience when accessing the intervention, staying engaged with the intervention instead of being overwhelmed or bored, and staying connected to professional support systems through online messaging [48,60,61]. This aligns with research reporting that help seekers have preferences for web-based treatments due to their convenience, accessibility, time efficiency, and ability to connect with professional support in a nontraditional manner (ie, not face-to-face) [20].

The majority of included studies recruited participants from Europe, North America, and Oceania. Only 4 studies recruited from Asian countries despite a significant amount of in-person intervention research for online addictions being conducted in East Asia [62,63]. It bears noting, however, that East Asian studies may often be published in non–English language journals. In South Korea and China, there have been parallel developments in structural and technological restrictions, including content filters, shutdown features, and time limits

[64-67]. However, research shows that people experiencing problems with their online addictive behaviors (specifically gaming, in this case) report disapproval with modifications to the structure of activities. Instead, they report stronger support for education, free online screening, self-monitoring tools, and warning labels [11] that are online in nature or can be adapted to be delivered online.

Several limitations of the review should be considered. First, a meta-analysis could not be conducted due to the limited quality of studies, limited data available, and varying study designs. Second, just 7 RCTs were included, but only 3 conducted short- or medium-term follow-up evaluation. In addition, the findings from the included studies were limited due to high rates of attrition at follow-up evaluation, with 4 out of 5 studies reporting on completer analysis only. To determine the effectiveness of web-based treatment, there needs to be a greater focus on RCT study design as well as participant retention and long-term follow-up evaluation. Third, we did not include non–English language literature, which might have excluded a large body of research conducted in East Asia. A strength of our study was the inclusion of studies that were focused on treatment rather than on prevention or early intervention; however, due to the heterogenous study focus and design, we were unable to determine who would likely benefit from web-based treatment. Future studies might consider examining the effectiveness of web-based treatment, level of problem severity, and type of addictive behavior. Just 1 included study compared in-person and web-based treatment and reported significant improvements in symptom severity after 8 sessions with no difference between groups. If future research finds web-based outcomes are similar to in-person treatment, then there is a strong case for expansion of web-based options.

There were also several limitations in relation to describing the content of interventions. We used the GIST-1 [42] to categorize change techniques instead of the 93-item behavior change technique (BCT) taxonomy [43] because of the absence of associated protocols or study registration. Just 2 studies referenced a published or manualized protocol or trial registration, which may reflect the exploratory nature of the research at this time. The lack of detailed reporting is common in behavioral addictions and was a reason for the development of the GIST-1 classification system which enables researchers to reliably code brief treatment descriptions [42]. Future studies may consider obtaining treatment manuals or working with a study developer to map the content of effective treatments onto the 93-item BCT taxonomy [43]. The current study was also limited to describing the content of interventions because of the limited sample. Future studies should consider examining the relationship between change techniques and participant outcomes. Studies should also consider examining the theoretical underpinnings or mechanisms of interventions and whether these also have an impact on severity, duration, or frequency of use.

This systematic review identified 12 studies assessing web-based treatments for online behavioral addictions. These findings highlight the potential of emerging web-based treatments, but the current evidence base varied greatly in study quality. This review also highlights the importance of having treatment protocols registered or published alongside an article and reporting components as aligned with BCTs or change techniques to be able to replicate studies with the exact components. Enhanced research designs are needed to develop a stronger evidence base to inform health care guidelines. Future research should also consider the relative appropriateness and cost-effectiveness of web-based treatments to guide the provision and allocation of funding across health systems. The review should be updated as more evidence on intervention effectiveness across online behavior types becomes available.

Acknowledgments

The authors wish to thank Irene Lu for assistance with screening studies.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategy.
[DOCX File , 13 KB - mental_v9i9e36662_app1.docx]

Multimedia Appendix 2
Characteristics of included studies (N=12).
[DOCX File , 28 KB - mental_v9i9e36662_app2.docx]

Multimedia Appendix 3
Content of web-based treatment in included studies (N=12).
[DOCX File , 23 KB - mental_v9i9e36662_app3.docx]

References

1. Brand M, Rumpf H, Demetrovics Z, MÜller A, Stark R, King DL, et al. Which conditions should be considered as disorders in the International Classification of Diseases (ICD-11) designation of "other specified disorders due to addictive behaviors"? J Behav Addict 2020 Jun 30:150-159. [doi: 10.1556/2006.2020.00035] [Medline: 32634114]
2. King DL, Delfabbro PH, Griffiths MD, Gradisar M. Assessing clinical trials of Internet addiction treatment: a systematic review and CONSORT evaluation. Clin Psychol Rev 2011 Nov;31(7):1110-1116. [doi: 10.1016/j.cpr.2011.06.009] [Medline: 21820990]
3. Gainsbury SM, Russell A, Hing N, Wood R, Lubman D, Blaszczynski A. How the Internet is changing gambling: findings from an Australian Prevalence Survey. J Gambl Stud 2015 Mar;31(1):1-15 [FREE Full text] [doi: 10.1007/s10899-013-9404-7] [Medline: 23934369]
4. International Classification of Diseases, Eleventh Revision (ICD-11). World Health Organization. URL: https://icd.who.int/browse11 [accessed 2021-01-01]
5. Reed GM, First MB, Billieux J, Cloitre M, Briken P, Achab S, et al. Emerging experience with selected new categories in the ICD-11: complex PTSD, prolonged grief disorder, gaming disorder, and compulsive sexual behaviour disorder. World Psychiatry 2022 Jun;21(2):189-213 [FREE Full text] [doi: 10.1002/wps.20960] [Medline: 35524599]
6. King DL, Russell A, Hing N. Adolescent land-based and internet gambling: Australian and international prevalence rates and measurement issues. Curr Addict Rep 2020 Jun 01;7(2):137-148. [doi: 10.1007/s40429-020-00311-1]
7. Gainsbury SM, Angus DJ, Blaszczynski A. Isolating the impact of specific gambling activities and modes on problem gambling and psychological distress in internet gamblers. BMC Public Health 2019 Oct 25;19(1):1372 [FREE Full text] [doi: 10.1186/s12889-019-7738-5] [Medline: 31653242]
8. Zajac K, Ginley MK, Chang R. Treatments of internet gaming disorder: a systematic review of the evidence. Expert Rev Neurother 2020 Jan;20(1):85-93 [FREE Full text] [doi: 10.1080/14737175.2020.1671824] [Medline: 31544539]
9. Cowlishaw S, Merkouris S, Dowling N, Anderson C, Jackson A, Thomas S. Psychological therapies for pathological and problem gambling. Cochrane Database Syst Rev 2012 Nov 14;11:CD008937. [doi: 10.1002/14651858.CD008937.pub2] [Medline: 23152266]
10. King DL, Delfabbro PH, Wu AMS, Doh YY, Kuss DJ, Pallesen S, et al. Treatment of Internet gaming disorder: An international systematic review and CONSORT evaluation. Clin Psychol Rev 2017 Jun;54:123-133. [doi: 10.1016/j.cpr.2017.04.002] [Medline: 28458097]
11. Stevens MWR, Delfabbro PH, King DL. Prevention strategies to address problematic gaming: An evaluation of strategy support among habitual and problem gamers. J Prim Prev 2021 Apr;42(2):183-201 [FREE Full text] [doi: 10.1007/s10935-021-00629-0] [Medline: 33710442]
12. Calado F, Griffiths MD. Problem gambling worldwide: An update and systematic review of empirical research (2000-2015). J Behav Addict 2016 Dec;5(4):592-613. [doi: 10.1556/2006.5.2016.073] [Medline: 27784180]
13. Konkolÿ Thege B, Woodin EM, Hodgins DC, Williams RJ. Natural course of behavioral addictions: a 5-year longitudinal study. BMC Psychiatry 2015 Jan 22;15:4 [FREE Full text] [doi: 10.1186/s12888-015-0383-3] [Medline: 25608605]
14. Bijker R, Booth N, Merkouris SS, Dowling NA, Rodda SN. Global prevalence of help-seeking for problem gambling: A systematic review and meta-analysis. Addiction 2022 Jul 13:1-14. [doi: 10.1111/add.15952] [Medline: 35830876]
15. Yeh YC, Wang PW, Huang MF, Lin PC, Chen CS, Ko CH. The procrastination of internet gaming disorder in young adults: The clinical severity. Psychiatry Research 2017;254:258-262. [doi: 10.1080/09638237.2016.1276530] [Medline: 28132570]
16. Müller KW, Beutel ME, Egloff B, Wölfling K. Investigating risk factors for Internet gaming disorder: a comparison of patients with addictive gaming, pathological gamblers and healthy controls regarding the big five personality traits. Eur Addict Res 2014;20(3):129-136. [doi: 10.1159/000355832] [Medline: 24247280]
17. Lee YS, Son JH, Park JH, Kim SM, Kee BS, Han DH. The comparison of temperament and character between patients with internet gaming disorder and those with alcohol dependence. J Ment Health 2017 Jun;26(3):242-247. [doi: 10.1080/09638237.2016.1276530] [Medline: 28132570]
18. Burleigh TL, Griffiths MD, Sumich A, Stavropoulos V, Kuss DJ. A systematic review of the co-occurrence of gaming disorder and other potentially addictive behaviors. Curr Addict Rep 2019 Sep 07;6(4):383-401. [doi: 10.1007/s40429-019-00279-7]
19. Suurvali H, Cordingley J, Hodgins DC, Cunningham J. Barriers to seeking help for gambling problems: a review of the empirical literature. J Gambl Stud 2009 Sep;25(3):407-424. [doi: 10.1007/s10899-009-9129-9] [Medline: 19551495]
20. Park JJ, Wilkinson-Meyers L, King DL, Rodda SN. Person-centred interventions for problem gaming: a stepped care approach. BMC Public Health 2021 May 06;21(1):872 [FREE Full text] [doi: 10.1186/s12889-021-10749-1] [Medline: 33957877]
21. Danielsson A, Eriksson A, Allebeck P. Technology-based support via telephone or web: a systematic review of the effects on smoking, alcohol use and gambling. Addict Behav 2014 Dec;39(12):1846-1868. [doi: 10.1016/j.addbeh.2014.06.007] [Medline: 25128637]
22. Rodda SN, Lubman DI, Dowling NA, McCann TV. Reasons for using web-based counselling among family and friends impacted by problem gambling. Asian J of Gambling Issues and Public Health 2013 Jul 29;3(1):1-11. [doi: 10.1186/2195-3007-3-12]

23. van der Maas M, Shi J, Elton-Marshall T, Hodgins DC, Sanchez S, Lobo DS, et al. Internet-based interventions for problem gambling: Scoping review. JMIR Ment Health 2019 Jan 07;6(1):e65 [FREE Full text] [doi: 10.2196/mental.9419] [Medline: 30617046]
24. Hing N, Russell AMT, Gainsbury SM, Blaszczynski A. Characteristics and help-seeking behaviors of Internet gamblers based on most problematic mode of gambling. J Med Internet Res 2015 Jan 07;17(1):e13 [FREE Full text] [doi: 10.2196/jmir.3781] [Medline: 25567672]
25. Costa S, Kuss DJ. Current diagnostic procedures and interventions for Gaming Disorders: A systematic review. Front Psychol 2019;10:578 [FREE Full text] [doi: 10.3389/fpsyg.2019.00578] [Medline: 30971971]
26. King DL, Delfabbro PH. Internet gaming disorder treatment: a review of definitions of diagnosis and treatment outcome. Journal of clinical psychology 2014;70(10):942-955. [doi: 10.1037/adb0000315] [Medline: 28921996]
27. Nazlıgül M, Baş S, Akyüz Z, Yorulmaz O. Internet gaming disorder and treatment approaches: A systematic review. Addicta: The Turkish Journal on Addictions 2018;5:A. [doi: 10.15805/addicta.2018.5.1.0018]
28. Raveendran RB, Kumar SV. Cognitive behaviour therapy for 'Internet Gaming Disorder' A systematic review. Indian Journal of Public Health 2020 Jul:657-661. [doi: 10.37506/ijphrd.v11i7.10161]
29. Stevens MWR, King DL, Dorstyn D, Delfabbro PH. Cognitive-behavioral therapy for Internet gaming disorder: A systematic review and meta-analysis. Clin Psychol Psychother 2019 Mar;26(2):191-203. [doi: 10.1002/cpp.2341] [Medline: 30341981]
30. Wéry A, Billieux J. Problematic cybersex: Conceptualization, assessment, and treatment. Addict Behav 2017 Jan;64:238-246. [doi: 10.1016/j.addbeh.2015.11.007] [Medline: 26646983]
31. King DL, Delfabbro PH, Griffiths MD. Clinical interventions for technology-based problems: excessive internet and video game use. J Cogn Psychother 2012 Jan 01;26(1):43-56. [doi: 10.1891/0889-8391.26.1.43]
32. Lemos IL, Abreu CND, Sougey EB. Internet and video game addictions: a cognitive behavioral approach. Rev. psiquiatr. clín 2014 Jul;41(3):82-88. [doi: 10.1590/0101-60830000000016]
33. Zajac K, Ginley MK, Chang R, Petry NM. Treatments for Internet gaming disorder and Internet addiction: A systematic review. Psychol Addict Behav 2017 Dec;31(8):979-994 [FREE Full text] [doi: 10.1037/adb0000315] [Medline: 28921996]
34. Malinauskas R, Malinauskiene V. A meta-analysis of psychological interventions for Internet/smartphone addiction among adolescents. J Behav Addict 2019 Dec 01;8(4):613-624 [FREE Full text] [doi: 10.1556/2006.8.2019.72] [Medline: 31891316]
35. Kuss DJ, Lopez-Fernandez O. Internet addiction and problematic Internet use: A systematic review of clinical research. World J Psychiatry 2016 Mar 22;6(1):143-176 [FREE Full text] [doi: 10.5498/wjp.v6.i1.143] [Medline: 27014605]
36. Lam LT, Lam MK. eHealth intervention for problematic internet use (PIU). Curr Psychiatry Rep 2016 Dec;18(12):107. [doi: 10.1007/s11920-016-0747-5] [Medline: 27766532]
37. Liu J, Nie J, Wang Y. Effects of group counseling programs, cognitive behavioral therapy, and sports intervention on internet addiction in East Asia: a systematic review and meta-analysis. Int J Environ Res Public Health 2017 Nov 28;14(12):1-17 [FREE Full text] [doi: 10.3390/ijerph14121470] [Medline: 29182549]
38. Weinstein A, Lejoyeux M. Internet addiction or excessive internet use. Am J Drug Alcohol Abuse 2010 Sep;36(5):277-283. [doi: 10.3109/00952990.2010.491880] [Medline: 20545603]
39. Rodda SN. A systematic review of internet delivered interventions for gambling: prevention, harm reduction and early intervention. J Gambl Stud 2021 Sep 13:967-991. [doi: 10.1007/s10899-021-10070-x] [Medline: 34515903]
40. Sagoe D, Griffiths MD, Erevik EK, Høyland T, Leino T, Lande IA, et al. Internet-based treatment of gambling problems: A systematic review and meta-analysis of randomized controlled trials. J Behav Addict 2021 Sep 17;10(3):546-565 [FREE Full text] [doi: 10.1556/2006.2021.00062] [Medline: 34546971]
41. Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Ann Intern Med 2009 Aug 18;151(4):264-9, W64. [Medline: 19622511]
42. Rodda S, Merkouris SS, Abraham C, Hodgins DC, Cowlishaw S, Dowling NA. Therapist-delivered and self-help interventions for gambling problems: A review of contents. J Behav Addict 2018 Jun 01;7(2):211-226 [FREE Full text] [doi: 10.1556/2006.7.2018.44] [Medline: 29895185]
43. Michie S, Richardson M, Johnston M, Abraham C, Francis J, Hardeman W, et al. The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. Ann Behav Med 2013 Aug;46(1):81-95. [doi: 10.1007/s12160-013-9486-6] [Medline: 23512568]
44. Thomas BH, Ciliska D, Dobbins M, Micucci S. A process for systematically reviewing the literature: providing the research evidence for public health nursing interventions. Worldviews Evid Based Nurs 2004;1(3):176-184. [doi: 10.1111/j.1524-475X.2004.04006.x] [Medline: 17163895]
45. Bőthe B, Baumgartner C, Schaub MP, Demetrovics Z, Orosz G. Hands-off: Feasibility and preliminary results of a two-armed randomized controlled trial of a web-based self-help tool to reduce problematic pornography use. J Behav Addict 2021 Oct 29:1015-1035 [FREE Full text] [doi: 10.1556/2006.2021.00070] [Medline: 34727088]
46. Caillon J, Grall-Bronnec M, Perrot B, Leboucher J, Donnio Y, Romo L, et al. Effectiveness of at-risk gamblers' temporary self-exclusion from internet gambling sites. J Gambl Stud 2019 Jun;35(2):601-615. [doi: 10.1007/s10899-018-9782-y] [Medline: 29974308]
47. He J, Pan T, Nie Y, Zheng Y, Chen S. Behavioral modification decreases approach bias in young adults with internet gaming disorder. Addict Behav 2021 Feb;113:106686. [doi: 10.1016/j.addbeh.2020.106686] [Medline: 33069109]

48. Park JJ, Booth N, Bagot KL, Rodda SN. A brief internet-delivered intervention for the reduction of gaming-related harm: A feasibility study. Computers in Human Behavior Reports 2020 Aug;2:100027. [doi: 10.1016/j.chbr.2020.100027]
49. Park SY, Kim SM, Roh S, Soh M, Lee SH, Kim H, et al. The effects of a virtual reality treatment program for online gaming addiction. Comput Methods Programs Biomed 2016 Jun;129:99-108. [doi: 10.1016/j.cmpb.2016.01.015] [Medline: 26860055]
50. Rabinovitz S, Nagar M. Possible end to an endless quest? Cognitive bias modification for excessive multiplayer online gamers. Cyberpsychol Behav Soc Netw 2015 Oct;18(10):581-587. [doi: 10.1089/cyber.2015.0173] [Medline: 26383549]
51. Su W, Fang X, Miller JK, Wang Y. Internet-based intervention for the treatment of online addiction for college students in China: a pilot study of the Healthy Online Self-helping Center. Cyberpsychology, Behavior, and Social Networking 2011;14(9):497-503. [doi: 10.1093/ndt/5.suppl_1.63] [Medline: 2129463]
52. Bottel L, Brand M, Dieris-Hirche J, Herpertz S, Timmesfeld N, Te Wildt BT. Efficacy of short-term telemedicine motivation-based intervention for individuals with Internet Use Disorder - A pilot-study. J Behav Addict 2021 Nov 17:1005-1014 [FREE Full text] [doi: 10.1556/2006.2021.00071] [Medline: 34797218]
53. Kent S, Masterson C, Ali R, Parsons CE, Bewick BM. Digital Intervention for Problematic Smartphone Use. Int J Environ Res Public Health 2021 Dec 14;18(24):1-16 [FREE Full text] [doi: 10.3390/ijerph182413165] [Medline: 34948774]
54. Hardy SA, Ruchty J, Hull TD, Hyde R. A preliminary study of an online psychoeducational program for hypersexuality. Sexual Addiction & Compulsivity 2010 Nov 30;17(4):247-269. [doi: 10.1080/10720162.2010.533999]
55. Luquiens A, Tanguy M, Lagadec M, Benyamina A, Aubin H, Reynaud M. The efficacy of three modalities of internet-based psychotherapy for non-treatment-seeking online problem gamblers: A randomized controlled trial. J Med Internet Res 2016 Feb 15;18(2):e36 [FREE Full text] [doi: 10.2196/jmir.4752] [Medline: 26878894]
56. Hayer T, Meyer G. Internet self-exclusion: Characteristics of self-excluded gamblers and preliminary evidence for its effectiveness. Int J Ment Health Addiction 2010 Aug 4;9(3):296-307. [doi: 10.1007/s11469-010-9288-z]
57. Eysenbach G. Design and evaluation of consumer health information web sites. In: Lewis D, Eysenbach G, Kukafka R, Stavri PZ, Jimison HB, editors. Consumer Health Informatics. New York, NY: Springer,; 2005:34-60.
58. Mohr DC, Cuijpers P, Lehman K. Supportive accountability: a model for providing human support to enhance adherence to eHealth interventions. J Med Internet Res 2011;13(1):e30 [FREE Full text] [doi: 10.2196/jmir.1602] [Medline: 21393123]
59. Soule LC, Shell LW, Kleen BA. Exploring internet addiction: Demographic characteristics and stereotypes of heavy internet users. Journal of Computer Information Systems 2016 Feb 01;44(1):64-73. [doi: 10.1080/08874417.2003.11647553]
60. Babic MJ, Smith JJ, Morgan PJ, Lonsdale C, Plotnikoff RC, Eather N, et al. Intervention to reduce recreational screen-time in adolescents: Outcomes and mediators from the 'Switch-Off 4 Healthy Minds' (S4HM) cluster randomized controlled trial. Prev Med 2016 Oct;91:50-57. [doi: 10.1016/j.ypmed.2016.07.014] [Medline: 27471018]
61. Stieger S, Lewetz D. A week without using social media: Results from an ecological momentary intervention study using smartphones. Cyberpsychol Behav Soc Netw 2018 Oct;21(10):618-624. [doi: 10.1089/cyber.2018.0070] [Medline: 30334650]
62. Shek DTL, Yu L. Longitudinal impact of the project PATHS on adolescent risk behavior: what happened after five years? ScientificWorldJournal 2012;2012:316029 [FREE Full text] [doi: 10.1100/2012/316029] [Medline: 22649287]
63. Yeun YR, Han SJ. Effects of psychosocial interventions for school-aged children's internet addiction, self-control and self-esteem: Meta-Analysis. Healthc Inform Res 2016 Jul;22(3):217-230 [FREE Full text] [doi: 10.4258/hir.2016.22.3.217] [Medline: 27525163]
64. Davies B, Blake E. Evaluating existing strategies to limit video game playing time. IEEE Comput Graph Appl 2016;36(2):47-57. [doi: 10.1109/MCG.2016.25] [Medline: 26960027]
65. King DL, Delfabbro PH, Doh YY, Wu AMS, Kuss DJ, Pallesen S, et al. Policy and prevention approaches for disordered and hazardous gaming and internet use: an international perspective. Prev Sci 2018 Feb;19(2):233-249. [doi: 10.1007/s11121-017-0813-1] [Medline: 28677089]
66. Király O, Griffiths MD, King DL, Lee H, Lee S, Bányai F, Zsila, et al. Policy responses to problematic video game use: A systematic review of current measures and future possibilities. J Behav Addict 2017 Sep 01:1-15. [doi: 10.1556/2006.6.2017.050] [Medline: 28859487]
67. Wang Q, Ren H, Long J, Liu Y, Liu T. Research progress and debates on gaming disorder. Gen Psychiatr 2019;32(3):e100071 [FREE Full text] [doi: 10.1136/gpsych-2019-100071] [Medline: 31423477]

Abbreviations

BCT: behavior change technique
CBM: cognitive bias modification
CBT: cognitive behavioral therapy
EPHPP: Effective Public Health Practice Project
GIST-1: Gambling Intervention System of CharacTerization
ICD-11: International Classification of Diseases 11th Revision
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PROSPERO: International Prospective Register of Systematic Reviews
RCT: randomized controlled trial

Edited by J Torous; submitted 19.01.22; peer-reviewed by G Humphreys; comments to author 19.05.22; revised version received 25.07.22; accepted 06.08.22; published 09.09.22.

Please cite as:
Park JJ, King DL, Wilkinson-Meyers L, Rodda SN
Content and Effectiveness of Web-Based Treatments for Online Behavioral Addictions: Systematic Review
JMIR Ment Health 2022;9(9):e36662
URL: https://mental.jmir.org/2022/9/e36662
doi:10.2196/36662
PMID:36083612

Original Paper

Assessing the Impact of Conversational Artificial Intelligence in the Treatment of Stress and Anxiety in Aging Adults: Randomized Controlled Trial

Morena Danieli[1], PsyD; Tommaso Ciulli[2], PhD; Seyed Mahed Mousavi[1], MSc; Giorgia Silvestri[2], MSc; Simone Barbato[2], MSc; Lorenzo Di Natale[2], MSc; Giuseppe Riccardi[1], PhD

[1]Signal & Interactive Systems Lab, Dipartimento di Ingegneria e Scienze dell'Informazione, Università degli Studi di Trento, Povo di Trento - Trento, Italy

[2]IDEGO - Digital Psychology srl, Rome, Italy

Corresponding Author:
Morena Danieli, PsyD
Signal & Interactive Systems Lab
Dipartimento di Ingegneria e Scienze dell'Informazione
Università degli Studi di Trento
via Sommarive 9
Povo di Trento - Trento, 38123
Italy
Phone: 39 5381237 ext 686
Email: morena.danieli@unitn.it

Abstract

Background: While mental health applications are increasingly becoming available for large populations of users, there is a lack of controlled trials on the impacts of such applications. Artificial intelligence (AI)-empowered agents have been evaluated when assisting adults with cognitive impairments; however, few applications are available for aging adults who are still actively working. These adults often have high stress levels related to changes in their work places, and related symptoms eventually affect their quality of life.

Objective: We aimed to evaluate the contribution of TEO (Therapy Empowerment Opportunity), a mobile personal health care agent with conversational AI. TEO promotes mental health and well-being by engaging patients in conversations to recollect the details of events that increased their anxiety and by providing therapeutic exercises and suggestions.

Methods: The study was based on a protocolized intervention for stress and anxiety management. Participants with stress symptoms and mild-to-moderate anxiety received an 8-week cognitive behavioral therapy (CBT) intervention delivered remotely. A group of participants also interacted with the agent TEO. The participants were active workers aged over 55 years. The experimental groups were as follows: group 1, traditional therapy; group 2, traditional therapy and mobile health (mHealth) agent; group 3, mHealth agent; and group 4, no treatment (assigned to a waiting list). Symptoms related to stress (anxiety, physical disease, and depression) were assessed prior to treatment (T1), at the end (T2), and 3 months after treatment (T3), using standardized psychological questionnaires. Moreover, the Patient Health Questionnaire-8 and General Anxiety Disorders-7 scales were administered before the intervention (T1), at mid-term (T2), at the end of the intervention (T3), and after 3 months (T4). At the end of the intervention, participants in groups 1, 2, and 3 filled in a satisfaction questionnaire.

Results: Despite randomization, statistically significant differences between groups were present at T1. Group 4 showed lower levels of anxiety and depression compared with group 1, and lower levels of stress compared with group 2. Comparisons between groups at T2 and T3 did not show significant differences in outcomes. Analyses conducted within groups showed significant differences between times in group 2, with greater improvements in the levels of stress and scores related to overall well-being. A general worsening trend between T2 and T3 was detected in all groups, with a significant increase in stress levels in group 2. Group 2 reported higher levels of perceived usefulness and satisfaction.

Conclusions: No statistically significant differences could be observed between participants who used the mHealth app alone or within the traditional CBT setting. However, the results indicated significant differences within the groups that received treatment and a stable tendency toward improvement, which was limited to individual perceptions of stress-related symptoms.

Trial Registration: ClinicalTrials.gov NCT04809090; https://clinicaltrials.gov/ct2/show/NCT04809090

(JMIR Ment Health 2022;9(9):e38067) doi:10.2196/38067

KEYWORDS

mental health care; conversational artificial intelligence; mobile health; mHealth; personal health care agent

Introduction

Background

The multiplicity of issues related with active aging has been on the agenda of national institutions and health agencies for many years worldwide. The European Union framework directive on health and safety at work (89/391/ EEC) [1] indicates that practicable adjustments to physical and social working environments are necessary to prevent or reduce excessive physical and mental demands on aging workers. Many studies have identified high levels of stress in the workplace as a major factor for developing age-related health risks, including cardiovascular diseases, sickness absence, anxiety, depression, and burnout syndrome [2-5]. As a consequence, several interventions have been implemented and evaluated for the prevention of physical diseases and mental disorders, and the strengthening of older employees, as reported in a recent systematic review [6]. Although this review did not focus only on the older population of workers, it reported some interesting relevant findings for our research. Based on moderate evidence that emerged from the review, cognitive behavioral therapy (CBT) and stress management programs are expected to reduce perceived stress. Nevertheless, the persistence and sustainability of these interventions were insufficient or limited.

Another systematic review analyzed the results of studies providing evidence for digital psychological interventions in the workplace [7]. The authors reviewed digital interventions aimed to address the well-known problem of accessibility of mental health care for the working population in general, due to limited resources in the occupational health sector and to stigma. The adjective “digital” in the reviewed studies stands for interventions whose primary modality of delivery was a website, where participants could access different types of assignments and receive feedback after completing the assignments from a coach or therapist by email, text, or phone call. All the summarized studies were randomized controlled trials (RCTs), but only one study reported data about a mobile app, and no study mentioned artificial intelligence (AI)-empowered treatments.

The demand for accessible and large-scale mental health care support has been previously pointed out [8] and aggravated by the COVID-19 pandemic and its consequences [9,10]. A growing number of studies have indicated that the development of conversational AI systems (also known as chatbots) as applications in the mental health domain can improve access to mental health care support in an easy and inexpensive manner [8,11,12]. Even though traditional in-person therapy sessions remain the most frequent framework for support provision, conversational AI agents have been shown to be an effective alternative regarding various mental disorders, such as stress, anxiety, and depression [8]. In particular, during the COVID-19 pandemic, the problem of accessibility to mental health treatments increased users’ appreciation of remote therapy, thus providing video therapy an opportunity to develop its potential in a world where these kinds of communications represent the new normal [13].

TEO (Therapy Empowerment Opportunity) is a mobile personal health care agent (m-PHA) designed to provide CBT support for the prevention and treatment of stress and anxiety [12]. It has been designed and developed in collaboration with CBT therapists [12]. In the course of the intervention, TEO converses with users through text-based dialogues. From these conversations, TEO recognizes users’ emotional states, beliefs, and personal events, and implements strategies designed by professionals.

Objective

The observational study discussed in this paper was designed for evaluating the impact of introducing AI technology in the psychological treatment of aging workers presenting a variety of stress symptoms hypothetically related with moderate to high levels of perceived stress in the workplace. The experimental protocol was designed to answer the following questions: (1) whether the use of AI-empowered conversational technologies could contribute to people’s psychological well-being; (2) whether there are differences in terms of symptom reduction between receiving support from an AI-empowered conversational technology and traditional psychotherapy; (3) whether the observed changes are different when comparing a group of people receiving treatment or not receiving it; and (4) whether there are differences compared with a group of people receiving a standard course with a psychologist in a remote setting.

Methods

Design

The experimental design included comparison of the presence of several different symptoms, like anxiety and depression, and psychological attitudes, measured by standardized self-assessed psychological questionnaires. We applied these metrics before treatment (T1) and at the end of treatment (T2). An additional measurement was performed 3 months after the end of treatment to longitudinally assess the effects (T3). The self-assessment scales we applied were Symptom Checklist-90-Revised (SCL-90-R), Occupational Stress Indicator (OSI), and Perceived Stress Scale (PSS). SCL-90-R is a self-administered questionnaire that assesses a broad spectrum of psychopathological symptoms like depression, anxiety, psychoticism, and others. OSI is a questionnaire for the evaluation of psychosocial stress in organizations. PSS is a brief questionnaire for the detection of generalized psychological stress. In addition, 2 brief versions of Patient Health Questionnaire-8 (PHQ-8) and General Anxiety Disorders-7 (GAD-7) were administered at the beginning of treatment (T1), after 4 weeks (T2), at the end of treatment (T3), and after 3 months (T4). PHQ-8 is an 8-item questionnaire for assessing

and monitoring depression severity [14], while GAD-7 is a short questionnaire for assessing and monitoring generalized anxiety disorders [15].

The treatment involved administering 8 weeks of cognitive behavioral psychotherapy, specifically oriented toward the acquisition of stress management skills. In addition, the experimental design included the possibility of supporting stress management training-CBT with the continuous assistance of an AI-based conversational agent for mental health care (TEO) [12]. The experimental design included 4 groups of subjects as follows: group 1 received traditional psychotherapy from CBT therapists in a remote setting; group 2 received both traditional therapy and the support of the conversational AI agent; group 3 received only the support of the conversational AI agent; and group 4 was the control group not receiving any treatment. Participants assigned to group 4 were also assigned to a waiting list and received treatment at the end of the 8 weeks of the experiment.

IDEGO (Digital Psychology srl, Rome, Italy) carried out the psychometric tests and their data analysis. The experimental design of the RCT, training, and evaluation of the AI algorithms and systems were performed by the University of Trento.

Ethics Approval

This methodology was approved by the Ethics Committee of the University of Trento within the context of the research activities of the HORIZON2020 CO-ADAPT project, and the experimental protocol has been registered on ClinicalTrials.gov (NCT04809090).

Recruitment

We collected the data of this study between Spring and Fall 2021, when the third wave of the COVID-19 pandemic was hitting Italy, starting from the Northern regions of the country. The traditional recruitment strategies were inadequate or limited owing to social distancing measures. To overcome these difficulties, we designed new strategies on social media with recruiting campaigns involving engaging posts and graphics. Comparing the usage statistics of the 2 social networks Facebook and Instagram, Facebook provided the highest percentage of users in our target group (21.3% and 11.7% for Facebook and Instagram, respectively) [16]. The campaigns were widespread throughout Italy, with the goal to motivate people to reach our website [17] and enroll in our research. The site included all the information about the research and a form where the users could request to participate. Moreover, the users could ask for more information, resulting in one-to-one interviews to answer all the questions. In order to select eligible participants, several questionnaires and a clinical interview with each subject were conducted. Exclusion criteria were the presence of severe depression (PHQ-8 score ≥20), suicidal thoughts, substance abuse, and mild cognitive impairment (Montreal Cognitive Assessment score <26) [18].

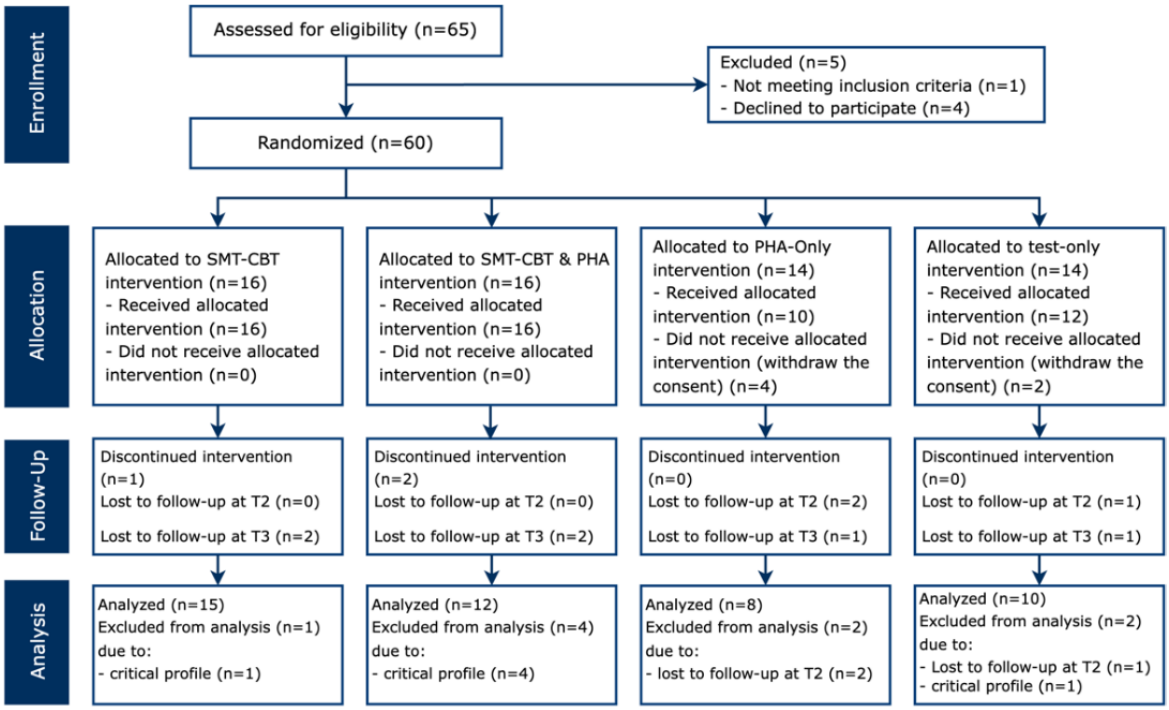
Participants

The characteristics of the samples are described in Table 1. A total of 65 potential participants were examined, and of these, 60 were recruited. A code was assigned to each participant, and through a random generator of numbers, the selected subjects were distributed into 4 groups. After the assignment, 2 subjects (1 in group 3 and 1 in group 4) showed mental health issues that made it necessary to reassign them to groups 1 and 2 to provide more accurate monitoring, where they could receive psychological support throughout the experiment. Other subjects showed a critical profile during the experiment, and they were directed to a standard psychological support service. Subsequently, these subjects were excluded from the analyses (Figure 1). Only 45 subjects were considered for the analysis. Group 1 included 27% (4/15) men and 73% (11/15) women, with a mean age of 54.08 (SD 4.11; median 54) years. Group 2 included 17% (2/12) men and 83% (10/12) women, with a mean age of 55.17 (SD 3.69; median 55) years. Group 3 included 25% (2/8) men and 75% (6/8) women, with a mean age of 55.63 (SD 4.50; median 55.5) years. Group 4 included 20% (2/10) men and 80% (8/10) women, with a mean age of 57.20 (SD 7.96; median 60) years.

Table 1. Sample characteristics (N=45).

Characteristic	Value
Age (years), mean (SD)	55.58 (5.08)
Gender, n (%)	
Male	10 (22)
Female	35 (78)
Group, n (%)	
Group 1	15 (33)
Group 2	12 (27)
Group 3	8 (18)
Group 4	10 (22)
Formal education, n (%)	
Secondary school	4 (9)
High school	14 (31)
Degree	16 (36)
Master's degree or PhD	2 (4)
Other	9 (20)
Marital status, n (%)	
Single	6 (13)
Cohabiting	2 (5)
Married	21 (47)
Separated	15 (33)
Widower	1 (2)

Figure 1. The CONSORT (Consolidated Standards of Reporting Trials) diagram shows the flow of the intervention, the enrollment of participants, their allocation to treatment, their follow-up, and data analysis. PHA: personal health care agent; SMT-CBT: stress management training-cognitive behavioral therapy; T2: end of the treatment; T3: 3 months after the treatment.



TEO

TEO is an m-PHA [19], a type of AI conversational agent, in the form of a mobile app that supports input/output interactions with users via natural language. Many PHAs currently developed for the mental health domain demonstrate limited flexibility of interactions with users, with system-directed interactions and a predefined dialogue flow [11]. As a result, the user has no control over the flow of the conversation and can only follow the system directives throughout the conversation. These limitations lead to shallow conversations and weak user engagement [8].

TEO allows users to share their thoughts and emotions using free-form natural language and engages users in personalized interactions about the events that are specific to each user. TEO can engage users in 2 types of dialogues. For the first type, TEO is designed to facilitate ABC (Activation, Belief, and Consequence) note writing for users. ABC notes are worksheets used by CBT therapists to help their patients in the identification of activating events (A), their beliefs related to the events (B), and the consequences of the events (C). Upon initiatives from a user to share a moment he/she is experiencing, TEO engages the user in dialogues where it asks a controlled set of questions designed by CBT therapists and collects an ABC note from the user in the form of a personal narrative about the event and his/her emotions. For the second type of dialogue (follow-up), TEO notifies the user about the ABC note written the day before and asks the user how he/she feels about the events, whether the issue is resolved, or whether the user is experiencing a different emotion [20]. TEO then tends to engage the user in a short personalized dialogue where it detects the recurrence of emotions and life events the user is experiencing [21], and provides helpful suggestions to ensure a healthier mental state.

Furthermore, TEO benefits from a knowledge base of therapeutic suggestions, recommendations, and exercises, which have been collected from therapists and domain experts. Users receive personalized tips and exercises weekly based on their progress of the therapy intervention. All the interactions with TEO are provided to the therapist weekly prior to the therapy session, so that the therapist can provide necessary support regarding the events and emotions expressed in the recollections and notes.

Measures

According to the findings by Sullivan and Artino [22] about the power of parametric versus nonparametric tests to detect differences between small-size samples, parametric analysis with repeated measures ANOVA (with a mixed within and between-subjects design) was performed to assess the differences between times (T1, T2, and T3) and groups (group 1, group 2, group 3, and group 4), and their interaction effect related to the results obtained in the PSS, SCL-90-R, and OSI tests. Multiple comparisons were corrected by using Bonferroni adjustment. The same analysis was conducted on PHQ-8 and GAD-7, which were administered before the intervention (T1), at mid-term (T2), at the end of the intervention (T3), and after 3 months (T4) to assess the differences between times (T1, T2, T3, and T4) and groups (group 1, group 2, group 3, and group 4). Regarding the OSI test, only a few scales were considered for the analysis, that is, the ones regarding coping strategies (social support, home-work relationship, task oriented, logic, time, and involvement), mental health, and physical health.

Results

PSS and SCL-90-R Results

The results obtained by administering the PSS and SCL-90-R tests are reported in Table 2. For the PSS, Global Severity Index (GSI), Positive Symptom Total (PST), Positive Symptom Distress Index (PSDI), obsessiveness-compulsiveness, interpersonal hypersensitivity, and depression scales/subscales, the assumption of sphericity had not been violated; otherwise, for the hostility and psychoticism subscales, the assumption had been violated (Multimedia Appendix 1 presents the results of the Mauchly test).

For the PSS, lower scores indicate lower stress levels and better well-being. Significant differences within groups between times were found for group 2 between T1 (mean 22.4, standard error [SE] 1.97) and T2 (mean 11.6, SE 2.36) (SE 2.52; P<.001), between T2 (mean 11.6, SE 2.36) and T3 (mean 16.6, SE 1.90) (SE 1.80; P=.03), and between T1 (mean 22.4, SE 1.97) and T3 (mean 16.6, SE 1.90) (SE 2.01; P=.02; Table 2). Further comparisons conducted within times between groups revealed a significant difference between groups at T1 ($F_{3,32}$=3.34; P=.03; η^2p=0.24), specifically between group 2 (mean 22.4, SE 1.97) and group 4 (mean 13.88, SE 2.20) (SE 2.95; P=.04) at T1.

For the GSI subscale of the SCL-90-R, lower values indicate less psychological distress. Significant differences within groups between times were found for group 2 between T1 (mean 59.4, SE 2.64) and T2 (mean 48.9, SE 3.99) (SE 2.83; P=.002; Table 2). Further comparisons conducted within times between groups did not highlight any significant difference.

For the PST subscale, lower scores indicate fewer reported symptoms. Significant differences within groups between times were found for group 2 between T1 (mean 59.7, SE 2.50) and T2 (mean 51.9, SE 3.17) (SE 2.24; P=.004; Table 2). Further comparisons conducted within times between groups did not highlight any significant difference.

For the PSDI subscale, lower scores indicate lower intensity of distress. Significant differences within groups between times were found for group 2 between T1 (mean 57, SE 2.51) and T2 (mean 45.1, SE 3.94) (SE 3.37; P=.004; Table 2). Further comparisons conducted within times between groups did not highlight any significant difference.

For the obsessiveness-compulsiveness subscale, lower scores indicate less symptomatology. Significant differences within groups between times were found for group 2 between T1 (mean 57.6, SE 2.29) and T2 (mean 47.9, SE 3.53) (SE 3.02; P=.009; Table 2). Further comparisons conducted within times between groups did not highlight any significant difference.

For the interpersonal hypersensitivity subscale, lower scores indicate less presence of feelings of inadequacy and inferiority. Significant differences within groups between times were found for group 2 between T1 (mean 54.9, SE 2.05) and T2 (mean 48, SE 2.36) (SE 2.21; P=.01; Table 2). Further comparisons

conducted within times between groups did not highlight any significant difference.

For the depression subscale, lower scores indicate less depression symptoms. Significant differences within groups between times were found for group 2 between T1 (mean 63.1, SE 3.29) and T2 (mean 51.8, SE 4.79) (SE 3.58; *P*=.01; Table 2). Further comparisons conducted within times between groups did not highlight any significant difference.

For the hostility subscale, lower scores indicate the presence of fewer anger-related personal characteristics. Significant differences within groups between times were found for group 2 between T1 (mean 57.6, SE 4.22) and T2 (mean 45.4, SE 2.21) (SE 4.49; *P*=.03; Table 2). Further comparisons conducted within times between groups did not highlight any significant difference.

For the psychoticism subscale, lower scores indicate less tendency of isolation and less presence of symptoms. Significant differences within groups between times were found for group 2 between T1 (mean 56.7, SE 3.29) and T2 (mean 50.2, SE 3.58) (SE 2.18; *P*=.02; Table 2). Further comparisons conducted within times between groups did not highlight any significant difference.

The results of the somatization, anxiety, phobic anxiety, and paranoid ideation (PAR) subscales are shown in Multimedia Appendix 1.

Table 2. Parametric analysis of repeated measures ANOVA for differences between times and groups with regard to the Perceived Stress Scale and Symptom Checklist-90-Revised test.

Scale/subscale and group[a]	Time[b]			*F* (*df*)	*P* value	η^2p
	T1, mean (SD)	T2, mean (SD)	T3, mean (SD)			
PSS[c] score						
Group 1	21.17 (6.24)	15.58 (7.81)	16.92 (5.45)	3.22 (2,31)	.053	0.17
Group 2	22.40 (5.66)	11.60 (5.85)	16.60 (6.29)	8.95 (2,31)	<.001	0.37
Group 3	21.50 (8.17)	14.00 (9.38)	18.67 (7.53)	2.86 (2,31)	.07	0.16
Group 4	13.87 (5.19)	14.63 (7.15)	15.13 (5.25)	0.17 (2,31)	.85	0.01
GSI[d]						
Group 1	58.42 (11.47)	56.33 (18.73)	53.17 (12.58)	2.16 (2,31)	.13	0.12
Group 2	59.40 (5.72)	48.90 (7.36)	54.70 (11.31)	6.77 (2,31)	.004	0.30
Group 3	54.67 (6.89)	48.83 (6.59)	50.50 (9.89)	1.44 (2,31)	.25	0.09
Group 4	53.25 (6.07)	49.88 (8.74)	51.50 (9.47)	0.57 (2,31)	.57	0.04
PST[e] score						
Group 1	57.00 (9.41)	54.92 (11.99)	53.50 (11.97)	1.10 (2,31)	.35	0.07
Group 2	59.70 (6.38)	51.90 (9.21)	55.00 (12.24)	5.97 (2,31)	.006	0.28
Group 3	56.00 (9.10)	48.67 (8.04)	50.50 (9.94)	3.14 (2,31)	.057	0.17
Group 4	56.88 (5.99)	51.63 (8.91)	53.13 (9.75)	2.14 (2,31)	.14	0.12
PSDI[f]						
Group 1	57.08 (10.25)	53.00 (16.83)	51.75 (9.96)	1.90 (2,31)	.17	0.11
Group 2	57.00 (8.49)	45.10 (6.59)	51.60 (9.57)	6.49 (2,31)	.004	0.30
Group 3	53.00 (4.73)	53.00 (14.99)	49.33 (9.48)	0.33 (2,31)	.72	0.02
Group 4	48.75 (3.88)	46.88 (7.00)	48.50 (8.19)	0.12 (2,31)	.89	0.01
Somatization score						
Group 1	57.25 (15.02)	53.83 (21.91)	52.83 (15.12)	1.15 (2,31)	.33	0.07
Group 2	55.90 (10.96)	47.30 (7.43)	49.60 (8.75)	2.99 (2,31)	.06	0.16
Group 3	49.67 (6.65)	45.83 (6.37)	42.83 (3.87)	1.40 (2,31)	.26	0.08
Group 4	51.38 (8.78)	50.50 (8.33)	50.88 (8.63)	0.02 (2,31)	.98	0.00
Obsessiveness-compulsiveness score						
Group 1	56.83 (8.62)	56.00 (15.58)	54.75 (11.34)	0.34 (2,31)	.72	0.02
Group 2	57.60 (7.04)	47.90 (7.91)	53.50 (11.08)	4.99 (2,31)	.01	0.24
Group 3	55.83 (7.63)	51.67 (7.03)	50.67 (11.24)	1.13 (2,31)	.34	0.07
Group 4	53.38 (4.21)	51.00 (8.47)	51.88 (8.06)	0.26 (2,31)	.78	0.02
Interpersonal sensitivity score						
Group 1	52.25 (6.11)	48.67 (7.23)	50.17 (10.47)	1.53 (2,31)	.23	0.09
Group 2	54.90 (6.33)	48.00 (6.60)	51.60 (11.46)	4.71 (2,31)	.02	0.23
Group 3	53.67 (8.43)	45.83 (5.31)	52.33 (9.07)	3.87 (2,31)	.03	0.20
Group 4	53.88 (5.57)	49.50 (9.84)	52.63 (15.90)	1.55 (2,31)	.23	0.09
Depression score						
Group 1	59.25 (12.13)	57.67 (22.76)	55.67 (11.26)	0.72 (2,31)	.50	0.04
Group 2	63.10 (9.79)	51.80 (9.46)	56.60 (13.13)	5.34 (2,31)	.01	0.26
Group 3	55.33 (9.27)	48.17 (7.63)	54.00 (13.23)	1.19 (2,31)	.32	0.07

Scale/subscale and group[a]	Time[b]			*F* (*df*)	*P* value	η^2p
	T1, mean (SD)	T2, mean (SD)	T3, mean (SD)			
Group 4	54.63 (8.86)	52.13 (8.86)	51.88 (9.03)	0.36 (2,31)	.70	0.02
Anxiety score						
Group 1	56.92 (13.59)	57.50 (23.62)	53.50 (10.37)	0.89 (2,31)	.42	0.05
Group 2	56.50 (11.57)	47.80 (5.45)	54.10 (9.61)	2.27 (2,31)	.12	0.13
Group 3	55.83 (6.94)	51.33 (8.57)	52.67 (11.29)	0.48 (2,31)	.62	0.03
Group 4	50.75 (5.31)	46.00 (6.16)	48.38 (6.78)	0.59 (2,31)	.56	0.04
Hostility score						
Group 1	47.50 (17.58)	48.33 (7.05)	47.33 (8.69)	0.12 (2,31)	.89	0.01
Group 2	57.60 (14.52)	45.40 (3.69)	48.80 (7.90)	4.12 (2,31)	.03	0.21
Group 3	53.50 (7.99)	49.83 (9.33)	52.00 (8.90)	0.39 (2,31)	.68	0.03
Group 4	51.75 (3.28)	46.88 (8.08)	48.63 (6.50)	0.60 (2,31)	.56	0.04
Phobic anxiety score						
Group 1	51.58 (11.02)	57.67 (21.64)	50.50 (9.56)	2.25 (2,31)	.12	0.13
Group 2	50.00 (16.67)	50.60 (6.02)	56.30 (11.38)	2.38 (2,31)	.11	0.13
Group 3	48.83 (5.14)	45.67 (3.14)	44.83 (1.60)	0.34 (2,31)	.72	0.02
Group 4	48.50 (5.43)	48.38 (5.34)	50.00 (10.92)	0.13 (2,31)	.88	0.01
Paranoid ideation score						
Group 1	59.92 (10.98)	52.33 (11.76)	53.92 (13.07)	6.58 (2,31)	.004	0.30
Group 2	54.20 (9.66)	49.50 (7.82)	54.90 (15.42)	3.49 (2,31)	.04	0.18
Group 3	56.50 (10.77)	46.83 (4.96)	48.83 (7.14)	5.35 (2,31)	.01	0.26
Group 4	52.25 (8.80)	45.25 (7.44)	50.63 (11.41)	4.57 (2,31)	.02	0.23
Psychoticism score						
Group 1	56.50 (12.75)	54.67 (15.20)	49.67 (8.79)	3.17 (2,31)	.06	0.17
Group 2	56.70 (9.56)	50.20 (8.87)	57.20 (12.64)	4.41 (2,31)	.02	0.22
Group 3	48.33 (6.15)	48.17 (6.01)	48.33 (7.47)	0.00 (2,31)	>.99	0.00
Group 4	52.63 (9.74)	52.88 (9.75)	52.38 (12.42)	0.01 (2,31)	.99	0.00

[a]Group 1 received only traditional therapy; group 2 received both traditional therapy and the support of a conversational artificial intelligence agent; group 3 received only the support of a conversational artificial intelligence agent; and group 4 did not receive any treatment (control group).

[b]T1 indicates before treatment, T2 indicates at the end of treatment, and T3 indicates 3 months after the end of treatment.

[c]PSS: Perceived Stress Scale.

[d]GSI: Global Severity Index.

[e]PST: Positive Symptom Total.

[f]PSDI: Positive Symptom Distress Index.

OSI Results

The main results of the OSI are reported in Table 3. For the task-oriented, logic, mental health, and physical health subscales, the assumption of sphericity had not been violated (Multimedia Appendix 1 presents the results of the Mauchly test).

For the task-oriented subscale, lower scores indicate criticality. Significant differences within groups between times were found for group 2 between T1 (mean 5.2, SE 0.56) and T2 (mean 6.9, SE 0.55) (SE 0.62; *P*=.04; Table 3). Further comparisons conducted within times between groups did not highlight any significant difference.

Table 3. Parametric analysis of repeated measures ANOVA for differences between times and groups with regard to the Occupational Stress Inventory.

Subscale and group[a]	Time[b]			*F* (*df*)	*P* value	η^2p
	T1, mean (SD)	T2, mean (SD)	T3, mean (SD)			
Social support score						
Group 1	7.13 (1.73)	6.50 (2.62)	5.88 (2.85)	1.57 (2,18)	.24	0.15
Group 2	5.20 (2.15)	6.10 (1.85)	6.10 (2.77)	1.68 (2,18)	.21	0.16
Group 3	5.00 (3.00)	5.00 (1.73)	5.33 (3.79)	0.07 (2,18)	.93	0.01
Group 4	7.00 (2.83)	7.50 (2.12)	7.50 (0.71)	0.10 (2,18)	.90	0.01
Task-oriented score						
Group 1	5.25 (1.98)	5.88 (2.30)	5.50 (3.12)	0.40 (2,18)	.67	0.04
Group 2	5.20 (1.55)	6.90 (1.45)	6.70 (2.50)	3.97 (2,18)	.04	0.31
Group 3	6.33 (1.16)	5.33 (0.58)	4.33 (3.06)	1.16 (2,18)	.34	0.11
Group 4	7.00 (2.83)	6.00 (0.00)	7.00 (1.41)	0.33 (2,18)	.72	0.04
Home-work relationship score						
Group 1	6.63 (1.69)	7.00 (1.07)	5.75 (1.58)	2.12 (2,18)	.15	0.19
Group 2	5.80 (1.32)	6.30 (1.64)	7.00 (1.33)	1.93 (2,18)	.17	0.18
Group 3	6.33 (2.08)	5.00 (1.73)	6.67 (2.52)	1.56 (2,18)	.24	0.15
Group 4	8.00 (1.41)	7.50 (0.71)	8.50 (0.71)	0.33 (2,18)	.72	0.04
Logic score						
Group 1	3.88 (1.64)	5.63 (2.50)	4.63 (2.33)	5.48 (2,18)	.01	0.38
Group 2	5.20 (1.81)	5.20 (1.75)	5.10 (2.08)	0.02 (2,18)	.99	0.00
Group 3	6.00 (1.00)	5.00 (0.00)	3.33 (2.52)	2.50 (2,18)	.11	0.22
Group 4	6.00 (1.41)	6.00 (2.83)	6.00 (0.00)	0.00 (2,18)	>.99	0.00
Time score						
Group 1	4.63 (2.13)	4.63 (2.26)	4.25 (1.75)	0.15 (2,18)	.86	0.02
Group 2	5.00 (2.00)	6.00 (1.83)	5.70 (2.45)	1.62 (2,18)	.23	0.15
Group 3	4.33 (1.16)	5.33 (2.08)	4.67 (1.53)	0.57 (2,18)	.57	0.06
Group 4	6.00 (4.24)	7.00 (2.83)	7.00 (1.41)	0.33 (2,18)	.72	0.04
Involvement score						
Group 1	5.25 (1.58)	6.50 (1.31)	5.13 (1.81)	1.79 (2,18)	.20	0.17
Group 2	5.70 (2.31)	6.70 (1.57)	6.00 (2.21)	1.16 (2,18)	.34	0.11
Group 3	6.33 (1.16)	6.00 (1.00)	6.00 (2.65)	0.06 (2,18)	.95	0.01
Group 4	7.50 (2.12)	6.50 (4.95)	7.50 (0.71)	0.26 (2,18)	.77	0.03
Mental health score						
Group 1	6.00 (2.98)	5.13 (2.59)	5.75 (1.67)	0.70 (2,18)	.51	0.07
Group 2	4.50 (2.01)	3.00 (2.00)	5.10 (3.32)	6.70 (2,18)	.007	0.43
Group 3	4.33 (3.06)	3.33 (2.52)	5.00 (3.61)	1.27 (2,18)	.31	0.12
Group 4	2.00 (1.41)	2.00 (1.41)	1.00 (0.00)	0.42 (2,18)	.67	0.04
Physical health score						
Group 1	6.75 (2.87)	5.75 (2.05)	5.75 (1.28)	0.90 (2,18)	.43	0.09
Group 2	7.60 (1.90)	5.70 (2.00)	5.50 (2.99)	4.17 (2,18)	.03	0.32
Group 3	6.00 (3.00)	4.67 (2.52)	5.33 (4.16)	0.61 (2,18)	.55	0.06
Group 4	4.50 (2.12)	3.00 (1.41)	7.50 (3.54)	2.70 (2,18)	.09	0.23

[a]Group 1 received only traditional therapy; group 2 received both traditional therapy and the support of a conversational artificial intelligence agent; group 3 received only the support of a conversational artificial intelligence agent; and group 4 did not receive any treatment (control group).

[b]T1 indicates before treatment, T2 indicates at the end of treatment, and T3 indicates 3 months after the end of treatment.

For the logic subscale, lower scores indicate criticality. Significant differences within groups between times were found for group 1 between T1 (mean 3.88, SE 0.59) and T2 (mean 5.63, SE 0.72) (SE 0.52; *P*=.01; Table 3). Further comparisons conducted within times between groups did not highlight any significant difference.

For the mental health subscale, lower scores indicate a higher level of mental well-being. Significant differences within groups between times were found for group 2 between T2 (mean 3.0, SE 0.72) and T3 (mean 5.1, SE 0.87) (SE 0.56; *P*=.004; Table 3). Further comparisons conducted within times between groups did not highlight any significant difference.

For the physical health subscale, lower scores indicate a higher level of physical well-being. Significant differences within groups between times were found for group 2 between T1 (mean 7.6, SE 0.77) and T2 (mean 5.7, SE 0.65) (SE 0.66; *P*=.03; Table 3). Further comparisons conducted within times between groups did not highlight any significant difference.

The results of the social support, home-work relationship, time, and involvement subscales are shown in Multimedia Appendix 1.

PHQ-8 and GAD-7 Results

The main results of PHQ-8 and GAD-7 are reported in Table 4. For the PHQ-8 test, lower scores indicate lower levels of depression. The only significant difference found was the one between groups at T1 ($F_{3,31}$=3.85; *P*=.02; η^2p=0.27), specifically between group 1 (mean 9.42, SE 1.16) and group 4 (mean 3.43, SE 1.51) (SE 1.91; *P*=.02).

For the GAD-7 test, lower scores indicate lower levels of generalized anxiety. Significant differences within groups between times were found for group 1 between T1 (mean 9.5, SE 1.21) and T4 (mean 4.83, SE 1.1) (SE 1.24; *P*=.004; Table 4). Furthermore, comparisons conducted within times between groups revealed significant differences between groups at T1 ($F_{3,31}$=3.53; *P*=.03; η^2p=0.25), specifically between group 1 (mean 9.5, SE 1.21) and group 4 (mean 3.14, SE 1.58) (SE 1.99, *P*=.02).

The PSS, SCL-90-R, OSI, PHQ-8, and GAD-7 results of the interaction effects between time and group can be found in Multimedia Appendix 1.

Table 4. Parametric analysis of repeated measures ANOVA for differences between times and groups with regard to Patient Health Questionnaire-8 and General Anxiety Disorders-7.

Scale/group[a]	Time[b]				*F* (*df*)	*P* value	η^2p
	T1, mean (SD)	T2, mean (SD)	T3, mean (SD)	T4, mean (SD)			
PHQ-8[c] score							
Group 1	9.42 (4.89)	7.50 (4.76)	7.58 (5.58)	5.83 (4.95)	3.17 (3,29)	.04	0.25
Group 2	6.30 (4.72)	6.70 (5.14)	5.60 (5.10)	6.70 (5.74)	0.52 (3,29)	.67	0.05
Group 3	4.83 (2.14)	5.50 (5.36)	5.17 (4.02)	4.50 (3.56)	0.10 (3,29)	.96	0.01
Group 4	3.43 (1.40)	5.00 (2.71)	5.43 (2.64)	3.14 (1.46)	0.90 (3,29)	.45	0.09
GAD-7[d] score							
Group 1	9.50 (5.23)	8.00 (6.67)	7.08 (5.14)	4.83 (4.02)	4.45 (3,29)	.01	0.32
Group 2	7.10 (3.64)	5.70 (2.45)	4.50 (2.76)	5.80 (3.80)	1.19 (3,29)	.33	0.11
Group 3	6.00 (4.43)	5.67 (6.25)	5.00 (4.86)	4.50 (4.76)	0.25 (3,29)	.86	0.03
Group 4	3.14 (2.04)	4.43 (1.72)	5.14 (1.77)	2.57 (2.23)	1.16 (3,29)	.34	0.11

[a]Group 1 received only traditional therapy; group 2 received both traditional therapy and the support of a conversational artificial intelligence agent; group 3 received only the support of a conversational artificial intelligence agent; and group 4 did not receive any treatment (control group).

[b]T1 indicates the beginning of treatment, T2 indicates after 4 weeks, T3 indicates at the end of treatment, and T4 indicates after 3 months.

[c]PHQ-8: Patient Health Questionnaire-8.

[d]GAD-7: General Anxiety Disorders-7.

Participant Feedback

At the end of treatment, feedback was collected from all the participants through the administration of a satisfaction questionnaire conceived for this study. For each item of the questionnaire, users were asked to indicate their degree of agreement on a 5-point Likert scale, from 1 (strongly disagree) to 5 (strongly agree). To assess satisfaction across all groups, 1 item of the questionnaire asked the users if they were satisfied overall with the received treatment. In the same way, to assess usefulness, they were asked if they felt that the treatment was useful. General results of satisfaction and perceived usefulness are shown in Table 5.

In addition to the general questions available for all groups, some specific questions were asked to assess the experience of the participants who could interact with TEO (ie, groups 2 and 3), focusing on the participants' experiences with the conversational agent. The results are shown in Table 6. "Easy to use" was used to refer to the ease of interaction with TEO, and "usefulness" was used to refer to the perceived usefulness of the app. "Personal usage" was intended to investigate if, in case the conversational agent was available on app stores (iOS or Android), the users would use it (using the question "If TEO was available on the Android/iOS store, would you use/download it for your personal use?"). Statistical analysis with one-way ANOVA was conducted to assess whether there were significant differences between groups for the above variables. No significance was detected. Specific results are reported in Multimedia Appendix 1.

Table 5. Satisfaction and perceived utility of the treatment.

Variable	Group 1[a]	Group 2[b]	Group 3[c]
Satisfaction score, mean (SD)	4.21 (0.89)	4.54 (0.66)	4.29 (0.76)
Usefulness score, mean (SD)	4.21 (0.89)	4.69 (0.63)	4.29 (0.76)

[a]Group 1 received only traditional therapy.

[b]Group 2 received both traditional therapy and the support of a conversational artificial intelligence agent.

[c]Group 3 received only the support of a conversational artificial intelligence agent.

Table 6. Participants' self-assessments of mobile personal health care agent interactions.

Variable	Group 2[a,b]	Group 3[a,c]
Easy to use score, mean (SD)	3.62 (1.04)	3.43 (1.40)
Usefulness score, mean (SD)	3.38 (0.87)	3.29 (1.38)
Personal usage score, mean (SD)	3.77 (1.09)	3.14 (1.77)

[a]All values reported represent the average of the group scores.

[b]Group 2 received both traditional therapy and the support of a conversational artificial intelligence agent.

[c]Group 3 received only the support of a conversational artificial intelligence agent.

Discussion

Principal Findings

Given the small number of subjects per group, the results concerning the differences between groups and between times within each group are discussed. The statistical analysis seemed to show significant differences between groups as follows: at T1, group 2 and group 4 differed in terms of the PSS, and group 1 and group 4 differed in terms of the GAD-7 and PHQ-8 scales. More specifically, group 2 reported higher levels of stress than group 4, group 1 reported higher levels of anxiety than group 4, and group 1 reported higher levels of depression than group 4. Although randomization of the groups was performed (explained in the Participants subsection in the Methods section), the differences at T1 for the GAD-7, PSS, and PHQ-8 scales could be due to a reduced sample size and a nonuniform distribution in the groups of subjects from different geographical zones of Italy as shown in Table 7. Overall, there was a worsening trend in almost all scales across all groups (Tables 2 and 3) between T2 and T3, although it did not appear to be significant. When we compared the interviews with some subjects and the Italian COVID-19 epidemiological statistics, we could observe an increase in COVID-19 positive cases and a general concern arising from the Delta variant of the virus in conjunction when the T3 statistics were collected from the participants after several months of general stability. This may be the reason for the overall deterioration observed from T2 to T3.

Table 7. Distribution of the sample according to the zones of Italy (North, Center, and South).

Zone	Group[a], n (%)				Total (N=45), n (%)
	Group 1 (n=15)	Group 2 (n=12)	Group 3 (n=8)	Group 4 (n=10)	
North	1 (6.7)	4 (33.3)	2 (25.0)	0 (0.0)	7 (15.6)
Center	12 (80.0)	6 (50.0)	6 (75.0)	9 (90.0)	33 (73.3)
South	2 (13.3)	2 (16.7)	0 (0.0)	1 (10.0)	5 (11.1)

[a]Group 1 received only traditional therapy; group 2 received both traditional therapy and the support of a conversational artificial intelligence agent; group 3 received only the support of a conversational artificial intelligence agent; and group 4 did not receive any treatment (control group).

Analysis conducted separately within each group showed that there were many significant differences between times in group 2. Altogether, group 2 seemed to show improvements in the PSS, GSI, PST, PSDI, obsessiveness-compulsiveness, interpersonal hypersensitivity, depression, hostility, and psychoticism scores (Table 2), as well as the task-oriented,

mental health, and physical health scores (Table 3). Despite significant improvements in group 2, for the PSS, there was significant worsening between T2 and T3, and there was an increase in stress at T3, although it was lower than that at T1. There was also worsening of psychological symptoms generally related to stress for the mental health scale of the OSI questionnaire (see Table 3 above). Despite significant worsening of the PSS and mental health (OSI) scores from T2 to T3, which generally detect similar symptoms of stress, it emerged that the physical health (OSI) score improved. This could indicate that subjects in group 2 were more susceptible to sudden changes, that is, increased cases of the contagious disease at T3 in their residential areas (in particular in Northern Italy) that could have increased psychological stress. However, the results for other scales suggest that the participants living in that geographical area could sufficiently cope with increased COVID-19–related worries, without developing higher levels of stress-related physical symptoms.

Group 3 reported significant improvements in the scores of the interpersonal hypersensitivity and PAR scales of the SCL-90-R questionnaire between T2 and T3 (see Table 2). Group 1 reported improved PAR (Table 2), logic (Table 3), and GAD-7 (Table 4) scores between times. Furthermore, group 4 showed a significant improvement in the PAR score between times (Table 2). In group 3, several individuals withdrew their consent to participate. Among them, 2 withdrew their consent owing to organizational complications that emerged and 2 withdrew their consent owing to very high expectations of the conversational AI agent that were not maintained. They judged that it was a waste of time to participate in this research compared with the perceived benefits.

The feedback questionnaires administered to understand the users' experiences showed that group 2 experienced higher levels of satisfaction and perceived the usefulness of the received treatment more than the other groups (ie, psychological support and use of the mobile health [mHealth] agent), as shown in Table 5. Moreover, comparing groups 2 and 3, participants in group 2 showed a greater ease of interaction with TEO, and they found it more useful than those in group 3 (ie, the group that interacted with the conversational AI agent without human psychological support). Indeed, the "personal usage" scores revealed a greater inclination of group 2 participants to use TEO.

A few specific questions were administered per group to explore some expectations. In group 1, the aim was to understand whether users would accept or find useful the use of a mHealth app together with traditional treatment. The results showed a score of 3.07. In group 3, the aim was to understand whether users would accept or find useful the use of a mHealth app together with traditional treatment. The results showed a score of 4.57. Overall, positive expectations related to combining traditional treatment with a mHealth app were found, considering the fact that participants in group 3, who used the app, had higher expectations. In group 2, the aim was to understand not the expectations but how effectively, for the subjects, the use of the mHealth app facilitated traditional treatment. The results showed a score of 3.93.

Altogether, these results suggest that a psychological treatment, characterized by the presence of human contact, along with a conversational mHealth agent would improve the impact of treatment in terms of satisfaction and usefulness.

A further aspect to be considered in the evaluation of these results is that this experiment was performed during the third wave of the COVID-19 pandemic in Italy, as mentioned in the Methods section. Our results indicate that although this event had an impact on the levels of stress and on the general psychological well-being of the participants, the observed and perceived improvements were maintained over time in terms of the reduction of physical stress–related symptoms.

Limitations

Following the recruitment process, the number of active participants involved in this study was small, and this may weaken the inferences and conclusions. Besides, although the recruitment campaign was conducted through social media platforms to reach out to all Italian regions, the majority of our participants were from the center of Italy. The participants were mainly women, and the fact that women tend to seek psychological help more often than men has been studied previously [23,24]. Nevertheless, the observed gender predominance weakens the generalization of the drawn inferences for both genders.

Conclusions

The aim of this study was to evaluate the possible improvements related to the introduction of an AI-based mHealth app in psychological interventions aiming to reduce stress-related physical and psychological symptoms in aging workers. We administered different standard psychological tests to measure the levels of perceived stress, generalized anxiety, and depression, along with other psychological dimensions. We could not observe statistically significant differences between the participants who used the mHealth app alone and those who used it within the traditional setting of psychological treatment. On the contrary, we could observe significant within-group differences, with improvements in subjects who received treatment. Moreover, we observed greater levels of satisfaction and subjective perception of usefulness in participants who were supported by a human therapist as well as the mHealth conversational agent.

Acknowledgments

The research leading to the present results has received funding from the European Union H2020 Programme under grant agreement 826266: COADAPT.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary results.
[DOCX File , 26 KB - mental_v9i9e38067_app1.docx]

Multimedia Appendix 2
CONSORT-eHEALTH checklist (V 1.6.1).
[PDF File (Adobe PDF File), 1219 KB - mental_v9i9e38067_app2.pdf]

References

1. Del Castillo AP. Occupational safety and health in the EU: back to basics. In: Vanhercke B, Natali D, Bouget D, editors. Social policy in the European Union: state of play 2016. Brussels: European Trade Union Institute and European Social Observatory; 2017:131-155.
2. Varianou-Mikellidou C, Boustras G, Dimopoulos C, Wybo J, Guldenmund FW, Nicolaidou O, et al. Occupational health and safety management in the context of an ageing workforce. Safety Science 2019 Jul;116:231-244. [doi: 10.1016/j.ssci.2019.03.009]
3. Lecca L, Campagna M, Portoghese I, Galletta M, Mucci N, Meloni M, et al. Work related stress, well-being and cardiovascular risk among flight logistic workers: An observational study. Int J Environ Res Public Health 2018 Sep 07;15(9):1952 [FREE Full text] [doi: 10.3390/ijerph15091952] [Medline: 30205457]
4. Götz S, Hoven H, Müller A, Dragano N, Wahrendorf M. Age differences in the association between stressful work and sickness absence among full-time employed workers: evidence from the German socio-economic panel. Int Arch Occup Environ Health 2018 May 28;91(4):479-496 [FREE Full text] [doi: 10.1007/s00420-018-1298-3] [Medline: 29487994]
5. Sara J, Prasad M, Eleid M, Zhang M, Widmer R, Lerman A. Association between work‐related stress and coronary heart disease: A review of prospective studies through the job strain, effort‐reward balance, and organizational justice models. JAHA 2018 May;7(9):a. [doi: 10.1161/jaha.117.008073]
6. Pieper C, Schröer S, Eilerts AL. Evidence of workplace interventions-A systematic review of systematic reviews. Int J Environ Res Public Health 2019 Sep 23;16(19):3553 [FREE Full text] [doi: 10.3390/ijerph16193553] [Medline: 31547516]
7. Carolan S, Harris PR, Cavanagh K. Improving employee well-being and effectiveness: Systematic review and meta-analysis of web-based psychological interventions delivered in the workplace. J Med Internet Res 2017 Jul 26;19(7):e271 [FREE Full text] [doi: 10.2196/jmir.7583] [Medline: 28747293]
8. Abd-Alrazaq AA, Alajlani M, Ali N, Denecke K, Bewick BM, Househ M. Perceptions and opinions of patients about mental health chatbots: Scoping review. J Med Internet Res 2021 Jan 13;23(1):e17828 [FREE Full text] [doi: 10.2196/17828] [Medline: 33439133]
9. Molodynski A, McLellan A, Craig T, Bhugra D. What does COVID mean for UK mental health care? Int J Soc Psychiatry 2021 Nov 10;67(7):823-825 [FREE Full text] [doi: 10.1177/0020764020932592] [Medline: 32517530]
10. Rossi R, Socci V, Talevi D, Mensi S, Niolu C, Pacitti F, et al. COVID-19 pandemic and lockdown measures impact on mental health among the general population in Italy. Front Psychiatry 2020 Aug 7;11:790 [FREE Full text] [doi: 10.3389/fpsyt.2020.00790] [Medline: 32848952]
11. Abd-Alrazaq AA, Alajlani M, Alalwan AA, Bewick BM, Gardner P, Househ M. An overview of the features of chatbots in mental health: A scoping review. Int J Med Inform 2019 Dec;132:103978. [doi: 10.1016/j.ijmedinf.2019.103978] [Medline: 31622850]
12. Danieli M, Ciulli T, Mousavi SM, Riccardi G. A conversational artificial intelligence agent for a mental health care app: Evaluation study of its participatory design. JMIR Form Res 2021 Dec 01;5(12):e30053 [FREE Full text] [doi: 10.2196/30053] [Medline: 34855607]
13. Berwick DM. Choices for the "new normal". JAMA 2020 Jun 02;323(21):2125-2126. [doi: 10.1001/jama.2020.6949] [Medline: 32364589]
14. Kroenke K, Spitzer RL. The PHQ-9: A new depression diagnostic and severity measure. Psychiatric Annals 2002 Sep 01;32(9):509-515. [doi: 10.3928/0048-5713-20020901-06]
15. Kroenke K, Spitzer RL, Williams JB, Monahan PO, Löwe B. Anxiety disorders in primary care: Prevalence, impairment, comorbidity, and detection. Ann Intern Med 2007 Mar 06;146(5):317. [doi: 10.7326/0003-4819-146-5-200703060-00004]
16. Facebook users in Italy January 2021. NapoleonCat Statistics. URL: https://napoleoncat.com/stats/facebook-users-in-italy/2021/01/ [accessed 2022-03-08]
17. CO-ADAPT. URL: https://www.co-adapt.it/ [accessed 2022-08-24]
18. Conti S, Bonazzi S, Laiacona M, Masina M, Coralli MV. Montreal Cognitive Assessment (MoCA)-Italian version: regression based norms and equivalent scores. Neurol Sci 2015 Feb 20;36(2):209-214. [doi: 10.1007/s10072-014-1921-3] [Medline: 25139107]

19. Riccardi G. Towards healthcare personal agents. In: RFMIR '14: Proceedings of the 2014 Workshop on Roadmapping the Future of Multimodal Interaction Research including Business Opportunities and Challenges. 2014 Presented at: 2014 Workshop on Roadmapping the Future of Multimodal Interaction Research including Business Opportunities and Challenges; November 16, 2014; Istanbul, Turkey p. 53-56. [doi: 10.1145/2666253.2666266]
20. Mousavi SM, Cervone A, Danieli M, Riccardi G. Would you like to tell me more? Generating a corpus of psychotherapy dialogues. In: Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations. 2021 Presented at: Second Workshop on Natural Language Processing for Medical Conversations; June 2021; Online p. 1-9. [doi: 10.18653/v1/2021.nlpmc-1.1]
21. Mousavi SM, Negro R, Riccardi G. An Unsupervised Approach to Extract Life-Events from Personal Narratives in the Mental Health Domain. CEUR. URL: http://ceur-ws.org/Vol-3033/paper12.pdf [accessed 2022-08-23]
22. Sullivan GM, Artino AR. Analyzing and interpreting data from Likert-type scales. J Grad Med Educ 2013 Dec;5(4):541-542 [FREE Full text] [doi: 10.4300/JGME-5-4-18] [Medline: 24454995]
23. Sagar-Ouriaghli I, Godfrey E, Bridge L, Meade L, Brown JSL. Improving mental health service utilization among men: A systematic review and synthesis of behavior change techniques within interventions targeting help-seeking. Am J Mens Health 2019 Oct 27;13(3):1557988319857009-1557988319851286 [FREE Full text] [doi: 10.1177/1557988319857009] [Medline: 31184251]
24. Liddon L, Kingerlee R, Barry JA. Gender differences in preferences for psychological treatment, coping strategies, and triggers to help-seeking. Br J Clin Psychol 2018 Mar 09;57(1):42-58. [doi: 10.1111/bjc.12147] [Medline: 28691375]

Abbreviations

ABC: Activation, Belief, and Consequence
AI: artificial intelligence
CBT: cognitive behavioral therapy
GAD-7: General Anxiety Disorders-7
GSI: Global Severity Index
mHealth: mobile health
m-PHA: mobile personal health care agent
OSI: Occupational Stress Indicator
PAR: paranoid ideation
PHQ-8: Patient Health Questionnaire-8
PSDI: Positive Symptom Distress Index
PSS: Perceived Stress Scale
PST: Positive Symptom Total
RCT: randomized controlled trial
SCL-90-R: Symptom Checklist 90 Revised
SE: standard error
TEO: Therapy Empowerment Opportunity

Edited by J Torous; submitted 17.03.22; peer-reviewed by K Uludag, M Birk, H Tanaka, M Rampioni; comments to author 09.06.22; revised version received 21.07.22; accepted 23.07.22; published 23.09.22.

Please cite as:
Danieli M, Ciulli T, Mousavi SM, Silvestri G, Barbato S, Di Natale L, Riccardi G
Assessing the Impact of Conversational Artificial Intelligence in the Treatment of Stress and Anxiety in Aging Adults: Randomized Controlled Trial
JMIR Ment Health 2022;9(9):e38067
URL: https://mental.jmir.org/2022/9/e38067
doi:10.2196/38067
PMID:

Original Paper

The Use of Automated Machine Translation to Translate Figurative Language in a Clinical Setting: Analysis of a Convenience Sample of Patients Drawn From a Randomized Controlled Trial

Hailee Tougas[1], MD; Steven Chan[2], MBA, MD; Tara Shahrvini[1], BA, BS; Alvaro Gonzalez[1], MA; Ruth Chun Reyes[1], BA; Michelle Burke Parish[1], PhD; Peter Yellowlees[1], MBBS, MD

[1]Department of Psychiatry and Behavioral Sciences, University of California, Davis, Sacramento, CA, United States

[2]Department of Psychiatry and Behavioral Sciences, Stanford University School of Medicine, Stanford, CA, United States

Corresponding Author:
Hailee Tougas, MD
Department of Psychiatry and Behavioral Sciences
University of California, Davis
2230 Stockton Blvd
Sacramento, CA, 95817
United States
Phone: 1 916 734 3574
Email: htougs@gmail.com

Abstract

Background: Patients with limited English proficiency frequently receive substandard health care. Asynchronous telepsychiatry (ATP) has been established as a clinically valid method for psychiatric assessments. The addition of automated speech recognition (ASR) and automated machine translation (AMT) technologies to asynchronous telepsychiatry may be a viable artificial intelligence (AI)–language interpretation option.

Objective: This project measures the frequency and accuracy of the translation of figurative language devices (FLDs) and patient word count per minute, in a subset of psychiatric interviews from a larger trial, as an approximation to patient speech complexity and quantity in clinical encounters that require interpretation.

Methods: A total of 6 patients were selected from the original trial, where they had undergone 2 assessments, once by an English-speaking psychiatrist through a Spanish-speaking human interpreter and once in Spanish by a trained mental health interviewer-researcher with AI interpretation. 3 (50%) of the 6 selected patients were interviewed via videoconferencing because of the COVID-19 pandemic. Interview transcripts were created by automated speech recognition with manual corrections for transcriptional accuracy and assessment for *translational* accuracy of FLDs.

Results: AI-interpreted interviews were found to have a significant increase in the use of FLDs and patient word count per minute. Both human and AI-interpreted FLDs were frequently translated inaccurately, however FLD translation may be more accurate on videoconferencing.

Conclusions: AI interpretation is currently not sufficiently accurate for use in clinical settings. However, this study suggests that alternatives to human interpretation are needed to circumvent modifications to patients' speech. While AI interpretation technologies are being further developed, using videoconferencing for human interpreting may be more accurate than in-person interpreting.

Trial Registration: ClinicalTrials.gov NCT03538860; https://clinicaltrials.gov/ct2/show/NCT03538860

(JMIR Ment Health 2022;9(9):e39556) doi:10.2196/39556

KEYWORDS

telepsychiatry; automated machine translation; language barriers; psychiatry; assessment; automated translation; automated; translation; artificial intelligence; AI; speech recognition; limited English proficiency; LEP; asynchronous telepsychiatry; ATP; automated speech recognition; ASR; AMT; figurative language device; FLD; language concordant; language discordant; AI interpretation

Introduction

The most recent US Census Bureau investigation records that nearly 26 million individuals older than 5 years are considered of limited English proficiency (LEP), with a reduced ability to speak, write, or read English [1]. In the United States, over 16 million Spanish-speaking individuals are classified as having LEP [1]. Among Latino immigrants, those with LEP are less likely to receive psychiatric health care as compared to those with English proficiency (EP) [2,3]. Federal and state policies have been created to reduce language barriers to health care and mandate that interpreter services be available to all LEP individuals [4,5]. Human interpreters are considered the *gold standard* to provide linguistically and culturally competent health care to patients with LEP, leading to improvements in patient comprehension and satisfaction, clinical outcomes, and health care use [6]. However, the usage rate of these services remains low, as less than 20% of clinical encounters for patients with LEP use interpreting services, often due to time constraints for clinical encounters [7].

Currently, artificial intelligence (AI) interpretation technologies have already been implemented in a variety of industries as either a replacement for or augmentation to human interpretation [8]. Health care, however, has been slow to apply AI technologies. Moreover, there are limited published applications of AI interpretation in health care, despite promising early results for the use of AI interpretation for the translation of written text, including public health information and electronic health records [6,8,9]. Notably, a paucity of information exists on the application of AI interpretation in health care to *spoken* rather than written text.

Most publications regarding clinical interpretation focus on ways to optimize the experience of using an interpreter, and there are various guidelines that suggest strategies to best integrate the interpreter into the encounter [10]. It is frequently advised to use simplified speech, with pauses between sentences to allow for sentence-by-sentence translation. Some published simplifications include shortening of phrases as well as avoidance of complex language, including idiomatic expressions, jargon, and humor [10]. The extent to which patients condense and simplify their speech when using an interpreter is yet to be evaluated.

This paper describes the results of a cross-sectional study to evaluate the translational accuracy of a novel AI interpretation technological tool composed of dual automated speech recognition (ASR) and automated machine translation (AMT) function. *ATP App* was developed by the University of California, Davis team to transcribe and translate psychiatric interviews with Spanish-speaking patients who have LEP. When assessing translational accuracy, it is important to be aware that mistakes can occur at both the ASR transcription and the AMT translation stages of AI interpretation. A separate paper further describing the accuracy of the AI interpretation has been prepared (Chan S et al, unpublished data, 2021). This study focuses specifically on the ability of *ATP App* to translate complex, figurative language devices (FLDs) such as metaphors, similes, and euphemisms [11]. To maintain the original meaning of these devices, the technology must be capable of recognizing that a literal, word-for-word translation does not always confer semantic equivalence between a phrase in Spanish and English [12]. As such, the translation of FLDs is a complex task, but one that would be required of AI interpretation in its application to real-world patient-provider conversations.

This study also aimed to quantify the extent to which the use of an interpreter affects patient speech quantity, measured by patient word count per minute; it also aimed to understand whether patient speech quantity differed between in-person or videoconferencing environments, the latter being required during the COVID-19 pandemic [13]. As such, we hoped to objectively quantify some of the time and language content barriers that physicians and patients face when using interpreting services.

Methods

Ethics Approval

This study was nested within a larger clinical trial approved by the University of California, Davis Institutional Review Board (IRB reference number: 1131922; trial registration number: NCT0358860) [14].

Participant Selection

The original study recruited Hispanic individuals with significant LEP from mental health and primary care clinics. All participants were aged 18 or older and screened as likely to have either a nonurgent psychiatric disorder, namely mood, anxiety or substance use disorders, or a chronic medical condition. Exclusion criteria included suicidal ideation or plans, significant cognitive deficits, and those otherwise deemed inappropriate for participation by their primary care provider or psychiatrist.

A total of 6 patients with psychiatric disorders were randomly selected from the original study of 114 patients. The first 3 (50%) patients were recruited prior to the COVID-19 pandemic, and the second 3 (50%) patients were recruited after the start of the pandemic. This allowed us to assess if the transition to a web-based, Zoom platform would impact AI interpretation.

Interview Format

The participants underwent 2 methods of psychiatric assessments. Method A represented the current gold standard of interviews of patients with LEP, whereby the Spanish-speaking patient was interviewed by an English-speaking psychiatrist, and the interview was interpreted by a human, English-Spanish interpreter. This method is the language-discordant format, with the provider and patient speaking different languages. Method B represented the novel, asynchronous telepsychiatry (ATP), AI interpretation format whereby the Spanish-speaking patient was interviewed by a Spanish-speaking researcher-interviewer, who was trained to administer psychiatric interviews. These interviews were video and audio recorded and subsequently transcribed and translated into English with subtitles added to the video file. The files were then sent to an English-speaking psychiatrist for diagnosis and treatment plan recommendations. Asynchronous telepsychiatry, without the added component of language interpretation, has already been established as a clinically valid

method for psychiatric assessments [15]. Transcription and translation were carried out via a novel, cloud-based, dual ASR and AMT app already developed by the research team, entitled *ATP App*. The videos were later viewed by the psychiatrist. This method is the language-concordant format, with the researcher-interviewer and patient speaking the same language. Of note, although it is common practice for human interpreters to *set the stage* and ask participants to simplify or shorten their speech to facilitate ease of interpretation, we specifically did not ask the participants to modify their speech in any way. This allowed us to analyze the natural speech of the encounters for both methods [9]. All interviews in both methods were video and audio recorded.

Transcription and Translation

Transcripts for both methods were generated from the video/audio recording of each interview. These transcripts were initially generated automatically and were subsequently verified for accuracy and edited by 2 bilingual researchers. The verification process was a labor-intensive process, requiring each reviewer to replay the file multiple times to add, remove, and replace words. The process of transcript verification required approximately 4 minutes of editing per 1 minute of the interview (Chan S et al, unpublished data, 2021). Instances of use of FLDs spoken by the patient were then separately marked by 2 bilingual researchers. There is a wide variety of FLDs (eg, similes, metaphors, irony, idiomatic expressions, and euphemisms), all of which apply language in a nonliteral manner to add connotation [11]. Table 1 presents examples for some common types of FLDs. FLDs used by the interviewers were excluded from analysis to control for natural variation in the style of speech used by the interviewers.

Table 1. Example figurative language devices.

Figurative language device subtype	Example in Spanish	Correct translation into English	Literal translation into English
Metaphor	Eso se me está llenando el cerebro.	This is overwhelming me.	This is filling my brain.
Idiomatic expression	Me hacen bien pesado.	It's been very hard.	They make me very heavy.
Simile	Me siento que no sirvo para nada.	I feel like I'm worthless.	I feel like I don't serve for anything.
Personification	Se me despega mi cabeza.	I lose my mind.	I peel away my head.
Euphemism	Me sentía yo más decaída.	I felt more down.	I felt more droopy.
Hyperbole or exaggeration	No me muero de hambre.	I'm not going to starve to death.	I'm not going to die from hunger.

Accuracy of transcription and translation of each FLD was independently determined by 2 bilingual researchers. If an FLD was categorized as an *inaccurate transcription*, the FLD was marked as "transcript inaccurate," and no subsequent analysis of translation was made, as translation is dependent on accurate transcription. If an FLD was categorized as an *accurate transcription*, the FLD was then subdivided into either an *accurate* or an *inaccurate translation*.

To analyze the quantity of patient speech, separate subtranscripts were created of only the patients' speech to obtain a patient word count. This word count was then divided by the minutes of the interview, to control for varying lengths of interviews. The number of instances of FLDs was divided by the number of minutes of the interview to control for the varying lengths of patient interviews.

The primary statistical analysis compared FLD frequency per minute, patient word count per minute, and percentage of accurate translation of FLDs between Method A and Method B for each patient. Analysis was performed using Microsoft Excel with paired sample two-sided *t* tests. The secondary statistical analysis compared only the percentage of accurate translation of FLDs as stratified into the in-person, pre–COVID-19 group for patients 1-3, and the Zoom format, post–COVID-19 group for patients 4-6. $P<.05$ was used to determine significance for all analyses.

Results

The study included 4 (67%) female and 2 (33%) male participants, with an age range of 42-71 years and an average age of 53 years; 4 (67%) participants were born in Mexico, 1 (17%) in Costa Rica, and 1 (17%) in Guatemala.

Figure 1 details the results of the three primary comparisons between each method—the frequency of figurative language devices as measured by number of FLDs per minute, the patient word count per minute, and the percentage of accurate translation as measured by number of correctly translated FLDs per total number of FLDs. There was a significant increase in the per-minute frequency of FLDs using AI interpretation (mean 0.61, SD 0.26) compared to using the human interpreter: mean 0.2, SD 0.1; $t_5=-4.58$, $P\leq.05$. There was a significant increase in the per-minute patient word count using AI interpretation (mean 90, SD 24.4) as compared to using the human interpreter: mean 45.8, SD 16.8; $t_5=-7.7$, $P\leq.05$. There was an insignificant decrease in the mean percentage of accurate translation of FLDs using AI interpretation (mean 0.3, SD 0.18) compared to using the human interpreter: mean 0.52, SD 0.29; $t_5=1.59$, $P=.17$.

Figure 1. Frequency of figurative language devices, patient word count per minute, and percentage of accurate translation per method and patient. FLDs: figurative language devices.

Patient	Method	# of FLDs per Minute	Word Count per Minute	% Accurate Translation of FLDs
1	A	0.25	64.7	0.6
2	A	0.29	41.3	0.45
3	A	0.3	66.6	0.47
4	A	0.13	40.4	0.8
5	A	0.21	38.9	0.8
6	A	0.03	22.9	0
	Mean	0.2	45.8	0.52
	SD	0.1	16.8	0.29
1	B	1.04	108.3	0.18
2	B	0.46	59.4	0.25
3	B	0.69	121.9	0.17
4	B	0.58	93.2	0.41
5	B	0.64	93.2	0.36
6	B	0.26	64.1	0.44
	Mean	0.61	90	0.3
	SD	0.26	24.4	0.18

Secondary comparisons were made to assess for possible differences in the percentage of accurate translation of FLDs for the interviews that were performed in person, prior to the COVID-19 pandemic, compared to those that were obtained over Zoom, after the COVID-19 pandemic. There was an insignificant increase in the accuracy of both methods for the Zoom format (mean 0.47, SD 0.3) as compared to the in-person format: mean 0.35, SD 0.17; t_5=–0.95, P=.39. When broken down separately by method, however, there was a near significant increase in the accuracy of AI interpretation for the Zoom format (mean 0.4, SD 0.04) as compared to the in-person format: mean 0.2, SD 0.04; t_2=–4.02, P=.06. There was an insignificant increase in the accuracy of human interpretation for the Zoom format (mean 0.53, SD 0.46) compared to the in-person format: mean 0.51, SD 0.08; t_2=–0.1, P=.92.

Discussion

Principal Findings

This study looked at the linguistic differences in psychiatric interviews of Spanish-speaking patients with LEP. The results demonstrate that the patients' speech differs significantly. Method A in the presence of a human interpreter showed fewer instances of FLDs, compared with Method B with language-concordant interviews augmented with AI interpretation. Additionally, in Method A, patients spoke with a lower word count per minute compared to Method B, with an average of half as many words per minute in the presence of a human interpreter. There was no statistically significant change in these results when using videoconferencing, compared to in-person consultations, although the interpreting accuracy over videoconferencing was higher for both methods.

Our findings aligned with our expectation that patient speech becomes simplified and truncated when using a human interpreter. This simplification aligns with many published guidelines and articles that detail best practices for use of human interpreting services, which often encourage a reduction in the use of idiomatic speech, as well as a simplification of sentence structure [10]. Within the specialty of psychiatry, diagnosis and treatment decisions are heavily reliant on the verbal history conveyed to the provider [16]. Our results suggest that the history provided using a human interpreter will likely differ and could represent a less comprehensive picture of the patient's psychopathology. Of note, human interpreting services guidelines are generally geared toward providers rather than patients, and the patients included in our study would likely not have read these guidelines prior to the study. Instead, we propose that there is an innate tendency for the patients to simplify their speech when having to pause between sentences to allow for translation. Additionally, the use of a human interpreter has previously been associated with a reduced number of follow-up appointments, reduced patient and provider satisfaction, and an increased likelihood of not asking the questions that the patient wanted to ask [17-19].

Moreover, the results of our study demonstrate that the use of an in-person human interpreter (Method A) is currently more accurate than AI interpretation (Method B) regarding the translation of FLDs. The aggregate translational accuracy for human interpreters was 52% versus 30% for AI interpretation (P>.05), suggesting that both methods lend themselves to a high degree of inaccuracy when translating FLDs. Of note, a sizable contribution to the inaccuracy of translation by the AMT starts from an inaccurate transcription of the conversation, suggesting that improvements in audio recording and transcription would increase the translational accuracy of the AI interpretation.

Finally, our results show that the transition of interviews from in-person to the web-based, Zoom format in response to the COVID-19 pandemic led to a higher, but statistically insignificant percentage of translational accuracy of FLDs, suggesting that both human-interpretation and AI interpretation technologies can be adapted to accommodate the movement away from in-person psychiatric evaluations. The aggregate translational accuracy of Method A is 50% in-person vs 53% over Zoom, and the aggregate translational accuracy of Method B is 20% in-person vs 40% over Zoom. This difference appears to stem from an improvement in *transcriptional* accuracy on the Zoom format, likely seen because interview participants took longer pauses after speaking and spoke in shorter phrases over the Zoom format.

Limitations

There are several limitations that we have identified in this study. First, the study is limited due to the small panel of patient interviews that are included. The decision to analyze a limited

subset of 12 patient interviews from the initial cohort of approximately 200 patient interviews was made due to the significant time required to both generate transcriptions for the in-person Method A and to verify the machine-generated transcripts for accuracy for Method B. Expanding the sample size of the included patient interviews is possible in the future using our database of recorded interviews but will be time consuming. This study is additionally limited by the wide variety of types of FLDs used in the interview discourse. Some devices, such as idioms and metaphors, are clear to delineate from nonfigurative speech. For example, the following patient statement, "estoy viendo una luz al final del túnel" ("I am seeing a light at the end of the tunnel") is clear to recognize as a figurative language device; it is well understood that the patient is not actually seeing a light, but rather that they are using an idiom that is in common use in both the English and the Spanish languages. By contrast, some of the types of devices that are used less frequently (eg, personification and euphemism) are more subtle. For example, the following patient statement, "la enfermedad me hizo traermelo para acá" (or "the sickness made me bring him too") is less obvious to recognize as figurative language, whereby her depression ("the sickness") is personified to have forced the patient to do something.

Highlights

- Patients with LEP frequently receive substandard health care because of language communication difficulties. Medical interpreters are often in short supply and commonly lengthen the time and simplify the language of medical interviews.
- A combination of ASR and AMT technologies have been developed as a method of AI interpretation. We applied these to ATP consultations as we believe AI interpretation may be a way of improving psychiatric interviews across languages compared with interviews mediated through human interpretation.
- In this study, the number of FLDs, the translation accuracy of figurative language, and the patient word counts were compared as proxies for interview complexity and volume. We found in the AI interpretation model that word counts were greater, and FLDs were more common but less accurately translated than in the human interpreter model.

Conclusion

Going forward, technological improvements of AI interpretation from both the transcription component and the translation component will be required for ATP interviews to be conducted in languages other than English. The field of AI interpretation has made substantial progress within the past decade with the transition from statistical machine translation to neural machine translation [20]; we expect that AI interpretation will continue to expand and improve in the coming years and to eventually be at least as accurate as professional interpreters, allowing it to be introduced into regular clinical use. As our patient population in the United States continues to diversify, it will be important to further develop novel technological approaches to circumvent the time limitations and simplification of speech that are currently seen with human interpretation. Further studies of the accuracy of interpretation over videoconferencing compared with in-person interpreting are required.

Acknowledgments

This study was funded by the Agency for Healthcare Research and Quality (R01 HS024949).

Conflicts of Interest

PY receives book royalties from the American Psychiatric Association. In 2022, SC has performed contract consulting for University of California, Davis, and owns <1% of stock in Orbit Health Telepsychiatry and Doximity. From 2017-2022, SC has taught and is financially compensated by North American Center for Continuing Medical Education, LLC.

References

1. Batalova J, Zong J. Language diversity and English proficiency in the United States. Migration Policy Institute. 2016. URL: https://www.migrationpolicy.org/article/language-diversity-and-english-proficiency-united-states-2015 [accessed 2022-08-19]
2. Kim G, Aguado Loi CX, Chiriboga DA, Jang Y, Parmelee P, Allen RS. Limited English proficiency as a barrier to mental health service use: a study of Latino and Asian immigrants with psychiatric disorders. J Psychiatr Res 2011 Jan;45(1):104-110. [doi: 10.1016/j.jpsychires.2010.04.031] [Medline: 20537658]
3. Limited English Proficiency (LEP). HHS.gov. URL: https://www.hhs.gov/civil-rights/for-individuals/special-topics/limited-english-proficiency/index.html [accessed 2021-01-15]
4. Section 1557: Ensuring meaningful access for individuals with limited English proficiency. HHS.gov. URL: https://www.hhs.gov/civil-rights/for-individuals/section-1557/fs-limited-english-proficiency/index.html [accessed 2021-01-15]
5. Ryan C. Language Use in the United States: 2011. Census.gov. 2013 Aug. URL: https://www2.census.gov/library/publications/2013/acs/acs-22/acs-22.pdf [accessed 2021-01-15]
6. Karliner LS, Jacobs EA, Chen AH, Mutha S. Do professional interpreters improve clinical care for patients with limited English proficiency? A systematic review of the literature. Health Serv Res 2007 Apr;42(2):727-754 [FREE Full text] [doi: 10.1111/j.1475-6773.2006.00629.x] [Medline: 17362215]
7. Hsieh E. Not just "getting by": factors influencing providers' choice of interpreters. J Gen Intern Med 2015 Jan 23;30(1):75-82 [FREE Full text] [doi: 10.1007/s11606-014-3066-8] [Medline: 25338731]

8. Kirchhoff K, Turner AM, Axelrod A, Saavedra F. Application of statistical machine translation to public health information: a feasibility study. J Am Med Inform Assoc 2011 Jul 01;18(4):473-478 [FREE Full text] [doi: 10.1136/amiajnl-2011-000176] [Medline: 21498805]
9. Soto X, Perez-de-Viñaspre O, Labaka G, Oronoz M. Neural machine translation of clinical texts between long distance languages. J Am Med Inform Assoc 2019 Dec 01;26(12):1478-1487 [FREE Full text] [doi: 10.1093/jamia/ocz110] [Medline: 31334764]
10. Juckett G, Unger K. Appropriate use of medical interpreters. Am Fam Physician 2014 Oct 01;90(7):476-480. [Medline: 25369625]
11. Crystal D. A Dictionary of Linguistics and Phonetics. Cambridge, MA, US: Basil Blackwell Ltd; 1991.
12. Saygin A. Processing figurative language in a multilingual task: Translation, transfer and metaphor. Corpus Linguistics 2001:1-7.
13. Yellowlees P, Nakagawa K, Pakyurek M, Hanson A, Elder J, Kales HC. Rapid Conversion of an Outpatient Psychiatric Clinic to a 100% Virtual Telepsychiatry Clinic in Response to COVID-19. Psychiatr Serv 2020 Jul 01;71(7):749-752. [doi: 10.1176/appi.ps.202000230] [Medline: 32460683]
14. University of California, Davis, Agency for Healthcare Research and Quality (AHRQ). Validation of an Automated Online Language Interpreting Tool - Phase Two. ClinicalTrials.gov. 2017. URL: https://clinicaltrials.gov/ct2/show/NCT03538860 [accessed 2020-11-19]
15. O'Keefe M, White K, Jennings JC. Asynchronous telepsychiatry: A systematic review. J Telemed Telecare 2019 Jul 29;27(3):137-145. [doi: 10.1177/1357633x19867189]
16. Bauer AM, Alegría M. Impact of patient language proficiency and interpreter service use on the quality of psychiatric care: a systematic review. Psychiatr Serv 2010 Aug;61(8):765-773 [FREE Full text] [doi: 10.1176/ps.2010.61.8.765] [Medline: 20675834]
17. Green AR, Ngo-Metzger Q, Legedza ATR, Massagli MP, Phillips RS, Iezzoni LI. Interpreter services, language concordance, and health care quality. Experiences of Asian Americans with limited English proficiency. J Gen Intern Med 2005 Nov;20(11):1050-1056 [FREE Full text] [doi: 10.1111/j.1525-1497.2005.0223.x] [Medline: 16307633]
18. Chang DF, Hsieh E, Somerville WB, Dimond J, Thomas M, Nicasio A, et al. Rethinking Interpreter Functions in Mental Health Services. Psychiatr Serv 2021 Mar 01;72(3):353-357. [doi: 10.1176/appi.ps.202000085] [Medline: 32988324]
19. Ngo-Metzger Q, Sorkin DH, Phillips RS, Greenfield S, Massagli MP, Clarridge B, et al. Providing high-quality care for limited English proficient patients: the importance of language concordance and interpreter use. J Gen Intern Med 2007 Nov 24;22 Suppl 2(S2):324-330 [FREE Full text] [doi: 10.1007/s11606-007-0340-z] [Medline: 17957419]
20. Wang H, Wu H, He Z, Huang L, Ward Church K. Progress in Machine Translation. Engineering. Preprint posted online July 14, 2021. [doi: 10.1016/j.eng.2021.03.023]

Abbreviations

AI: artificial intelligence
AMT: automated machine translation
ASR: automated speech recognition
ATP: asynchronous telepsychiatry
FLD: figurative language device
LEP: limited English proficiency

Edited by J Torous; submitted 13.05.22; peer-reviewed by A Naser, S Markham; comments to author 27.06.22; revised version received 10.07.22; accepted 10.07.22; published 06.09.22.

Please cite as:
Tougas H, Chan S, Shahrvini T, Gonzalez A, Chun Reyes R, Burke Parish M, Yellowlees P
The Use of Automated Machine Translation to Translate Figurative Language in a Clinical Setting: Analysis of a Convenience Sample of Patients Drawn From a Randomized Controlled Trial
JMIR Ment Health 2022;9(9):e39556
URL: https://mental.jmir.org/2022/9/e39556
doi:10.2196/39556
PMID:36066959

Publisher:
JMIR Publications
130 Queens Quay East.
Toronto, ON, M5A 3Y5
Phone: (+1) 416-583-2040
Email: support@jmir.org