Review

Evaluating Conversational Agents for Mental Health: Scoping Review of Outcomes and Outcome Measurement Instruments

Ahmad Ishqi Jabir[1,2], BSc, MSc; Laura Martinengo[1], MD, PhD; Xiaowen Lin[1], BA; John Torous[3], MD; Mythily Subramaniam[4,5], PhD; Lorainne Tudor Car[1,6], MD, PhD

[1]Lee Kong Chian School of Medicine, Nanyang Technological University Singapore, Singapore, Singapore

[2]Future Health Technologies, Singapore-ETH Centre, Campus for Research Excellence And Technological Enterprise, Singapore, Singapore

[3]Beth Israel Deaconess Medical Center, Boston, MA, United States

[4]Institute of Mental Health, Singapore, Singapore

[5]Saw Swee Hock School of Public Health, National University of Singapore, Singapore, Singapore

[6]Department of Primary Care and Public Health, School of Public Health, Imperial College London, London, United Kingdom

Corresponding Author:
Lorainne Tudor Car, MD, PhD
Lee Kong Chian School of Medicine
Nanyang Technological University Singapore
11 Mandalay Road, Level 18
Singapore, 308232
Singapore
Phone: 65 69041258
Email: lorainne.tudor.car@ntu.edu.sg

Abstract

Background: Rapid proliferation of mental health interventions delivered through conversational agents (CAs) calls for high-quality evidence to support their implementation and adoption. Selecting appropriate outcomes, instruments for measuring outcomes, and assessment methods are crucial for ensuring that interventions are evaluated effectively and with a high level of quality.

Objective: We aimed to identify the types of outcomes, outcome measurement instruments, and assessment methods used to assess the clinical, user experience, and technical outcomes in studies that evaluated the effectiveness of CA interventions for mental health.

Methods: We undertook a scoping review of the relevant literature to review the types of outcomes, outcome measurement instruments, and assessment methods in studies that evaluated the effectiveness of CA interventions for mental health. We performed a comprehensive search of electronic databases, including PubMed, Cochrane Central Register of Controlled Trials, Embase (Ovid), PsychINFO, and Web of Science, as well as Google Scholar and Google. We included experimental studies evaluating CA mental health interventions. The screening and data extraction were performed independently by 2 review authors in parallel. Descriptive and thematic analyses of the findings were performed.

Results: We included 32 studies that targeted the promotion of mental well-being (17/32, 53%) and the treatment and monitoring of mental health symptoms (21/32, 66%). The studies reported 203 outcome measurement instruments used to measure clinical outcomes (123/203, 60.6%), user experience outcomes (75/203, 36.9%), technical outcomes (2/203, 1.0%), and other outcomes (3/203, 1.5%). Most of the outcome measurement instruments were used in only 1 study (150/203, 73.9%) and were self-reported questionnaires (170/203, 83.7%), and most were delivered electronically via survey platforms (61/203, 30.0%). No validity evidence was cited for more than half of the outcome measurement instruments (107/203, 52.7%), which were largely created or adapted for the study in which they were used (95/107, 88.8%).

Conclusions: The diversity of outcomes and the choice of outcome measurement instruments employed in studies on CAs for mental health point to the need for an established minimum core outcome set and greater use of validated instruments. Future studies should also capitalize on the affordances made available by CAs and smartphones to streamline the evaluation and reduce participants' input burden inherent to self-reporting.

(J Med Internet Res 2023;25:e44548) doi: 10.2196/44548

KEYWORDS

conversational agent; chatbot; mental health; mHealth; mobile health; taxonomy; outcomes; core outcome set

Introduction

Recent technological advances have led to the proliferation of digital interventions, such as conversational agents (CAs), in different areas of health care, including mental health [1]. CAs, also known as chatbots, are multimodal systems that support conversational interactions with users through text, voice, and images [2]. CAs offer scalability and 24-hour availability, which allows timely interventions focusing on management, treatment, prevention of mental health conditions, and improvement of mental well-being. Woebot, for example, is a primarily text-based CA, which provides timely check-ins with users to encourage mood tracking and deliver general psychoeducation based on cognitive behavior therapy and behavior change tools [3,4]. A recent systematic review on the effectiveness of CA-delivered interventions for depression and anxiety showed a significant decrease in depressive symptoms in adults [5]. However, the low quality of overall evidence and limited well-designed randomized controlled trials (RCTs) [5,6] suggest the need to improve the quality of trials further.

Recent reviews on the use of CAs for mental health suggest heterogeneity of the outcome measurements used [6-9]. This issue is not limited to CAs but involves digital health interventions (DHIs) in general [10-12]. For example, studies that evaluated mental health DHIs typically reported user experience, satisfaction, and engagement with the intervention without clear and standardized criteria to evaluate them [6,8,11]. For instance, a study may report subjective feedback from users but not include objective measurements, such as average duration of use and the number of modules completed, to provide a better understanding of the context of use [11,13]. While efforts had been made to set a standardized benchmark for subjectively reported user experience outcomes [14], there are no gold standards that objectively measure these outcomes [8]. This is further hampered by (1) the lack of standardized taxonomy to describe the breadth of measurement instruments available [8,10,11,15] and (2) the use of outcome measurement instruments without validity evidence, which affects the credibility of study findings [16]. Lastly, objective measures to assess the performance of the system are also important [9]. This includes measures that track technical issues, such as system crashes and glitches, to understand if the CA is working well during the intervention. Similar to user experience, subjective measures of technical issues should be explored in conjunction with objective measures. This may include objective counts of error-handling messages sent in addition to user subjective experience of the dialogues or CAs in general [2].

The method of data collection is as important as the outcome measurement instrument used to improve the quality of clinical trials. Traditional methods via pen-and-paper and phone-based surveys can be costly and burdensome to participants and researchers alike [17]. Varied means of data collection approaches are particularly relevant in studies on digital mental health and well-being interventions, which are prone to high dropout rates [18]. This may include innovative ways of data collection, ranging from the integration of web-based survey platforms, such as Qualtrics and Google Form [4], in the system to the collection of passive smartphone sensor information in the form of digital biomarkers [19]. Passive and ongoing data collection also allows for more frequent measurements that can be used to reduce participants' input burden [19]. To improve the transparency and quality of the evaluation and reporting of CA-delivered interventions focusing on mental health and well-being, there is a need to identify the choice of outcomes and outcome measurement instruments used in studies to date. Correspondingly, in this review, we aimed to (1) identify the types of outcome measurement instruments reported in studies assessing the effectiveness of mental health interventions delivered by CAs, (2) identify the data collection methods used (eg, pen-and-paper or technology-assisted methods) and the frequency of data collection in these studies, and (3) determine the prevalence of outcome measurement instruments with validity evidence employed in mental health interventions delivered by CAs.

Methods

Overview

This report follows the Joanna Briggs Institute scoping review guidelines [20] and the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) [21] checklist (Multimedia Appendix 1). The protocol was registered on the Open Science Framework database (protocol ID: DEG4K).

Search Strategy

A search strategy including 63 terms that define or are synonymous with CAs was designed and used in a series of scoping reviews to explore the use of CAs in health care (Multimedia Appendix 2). The search included sources of peer-reviewed research specializing in medical, psychology, engineering, multidisciplinary, and grey literature. The search was performed on April 26, 2021, in the PubMed, Cochrane Central Register of Controlled Trials, PsychINFO, Web of Science, and Embase (Ovid) databases and in the first 10 pages of Google Scholar and Google [22]. These databases were chosen based on our experience in developing similar reviews on CAs in health care and were optimized by a medical librarian for this review [22].

Eligibility Criteria

This scoping review included experimental primary studies, such as RCTs, cluster randomized trials, quasirandomized trials, controlled before-and-after studies, uncontrolled before-and-after studies, interrupted time series, pilot studies, and feasibility studies. Nonexperimental studies, such as observational studies, reviews, qualitative studies, editorials, personal communications, conference abstracts, and articles where the full text was not available, were excluded. We included mental health interventions delivered by CAs, including the promotion of mental well-being, and the prevention and management of

mental health disorders, including but not limited to mood disorders, psychosis, posttraumatic stress disorder, and substance use disorders. We excluded studies that focused primarily on comparing or evaluating specific CA features, those that did not report health outcomes, and those whose dialogue was derived from human operators ("Wizard of Oz").

Within the context of this study, a CA was defined as a human-machine interface that holds human-like synchronous conversation via text, voice, images, video, or multimodal outputs, and autonomously interprets user input via decision trees or complex neural network algorithms [22]. CAs could be preconfigured with a set of predefined responses (rule-based CAs) or enhanced with natural language processing or machine learning (artificial intelligence [AI]-enhanced CAs) [22]. An embodied CA was defined as a CA that includes an avatar with human-like features, which can mimic human movements and facial expressions [22].

Screening, Data Extraction, and Analysis

The title and abstract screening was performed by 2 reviewers (AIJ and XL) independently and in parallel on Covidence [23]. Studies included in this step were uploaded to EndNote X9 (Clarivate) for full-text review, which was performed in parallel by AIJ and XL. Discrepancies among the reviewers were settled via discussions between the reviewers or with input from a third reviewer (LM). The data extraction form was developed by the research team using Microsoft Excel (Microsoft Corp). The data extraction was performed in parallel by AIJ and XL. The form was piloted on 3 studies and then amended based on feedback to better fit the research aims. The extracted data were compared, and disagreements were resolved via discussion or input from LM acting as the arbiter. Data were presented in a diagrammatic or tabular form accompanied by a narrative summary of the findings.

The outcomes were categorized into clinical, technical, and user experience outcomes. Clinical outcomes were defined as "measurable changes in health, function, or quality of life" [22]. These outcomes derive directly or indirectly from the expected mechanisms of the CA-delivered intervention. Clinical outcomes were categorized based on the Core Outcome Measures in Effectiveness Trials (COMET) Initiative's medical research outcome taxonomy comprising 38 categories such as "21: Psychiatric Outcomes," "26: Physical Functioning," and "28: Emotional Functioning" (see Multimedia Appendix 3 for the definitions) [15]. User experience outcomes "encompassed all direct and indirect experiences of the user while interacting with the CA" [22]. These included the subjective self-reported experience of the intervention, such as system usability, satisfaction with the CA, and interviews with users. We also included objective engagement measures, which were further categorized based on a previous systematic review of mobile health (mHealth) interventions for depression [10]. Technical outcomes were measures used to evaluate the performance of the CA itself related to its technical interface, system crashes, and dialogue system, such as chatbot response generation [9]. Unlike previous research [9], we did not consider users' experiences of glitches and errors as technical outcomes. Rather, technical outcomes strictly referred to objective measures of system performance, such as the number of errors from the system log.

The outcome measurement instruments were categorized into those measuring outcomes objectively and subjectively. Objective measures included (1) "sensor data" to monitor human behavior or physiological changes using either external sensors, such as respiratory sensors for breathing rate [24], or smartphone "passive sensing modules," such as gyroscopes, GPS modules, or accelerators [19]; and (2) "objective engagement measures" defined as data captured passively by the system log while the user interacts with the system [8,10]. Subjective measures included those measured with instruments or tools (eg, questionnaires) that involve self-reporting by the participant using either pen-and-paper or digital means. These measurement instruments may assess health-related outcomes, such as symptoms, symptom burden, health-related quality of life, usability, or satisfaction with the system.

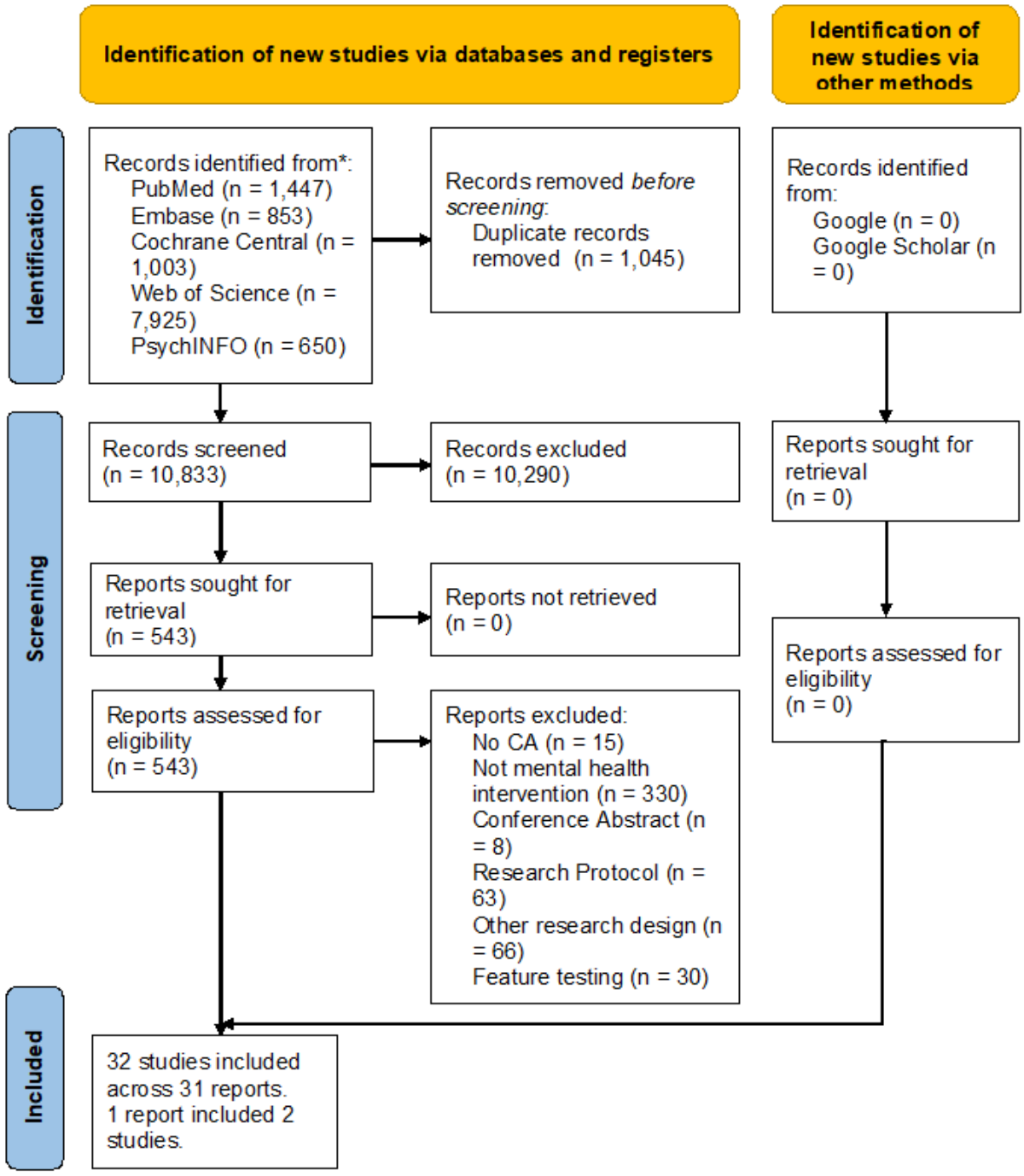
Validity evidence of the outcome measurement instruments was extracted based on the COMET Initiative's taxonomy of measurement properties [25], comprising 3 quality domains: reliability, validity, and responsiveness. The reliability domain includes internal consistency of the items, reliability, and measurement errors not attributed to true changes in the construct measures. The validity domain includes the content, construct, and criterion validity of the instruments. The responsiveness domain covers longitudinal validity or the ability of the instrument to detect change over time. The validity evidence was extracted based on the measurement properties that were reported directly from the studies or referenced by the studies. We recorded the relevant details from the other references, including if the study reported more than one reference.

Results

Overview

The search strategy retrieved 10,833 papers after removing duplicates, of which 543 were eligible for full-text screening, and 31 papers were included. We reported a total of 32 studies as 1 paper included 2 studies. Figure 1 presents the study selection process.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart. CA: conversational agent.



Characteristics of the Included Studies

Of the 32 studies included in this review, 25 (78%) [3,4,13,24,26-45] were published in the past 5 years, with 5 studies [4,13,26,34,35] published in the first quarter of 2021 (Table 1). Most of the studies (16/32, 50%) were from the United States of America [3,4,24,30,32-34,42-49].

Table 1. Characteristics of the included studies.

Study characteristic	Value (N=32), n (%)
Year of publication	
<2017	7 (22)
≥2017	25 (78)
Country	
United States of America	16 (50)
United Kingdom	7 (22)
Germany	1 (3)
Italy	1 (3)
Sweden	1 (3)
Japan	3 (9)
Korea	2 (6)
China	1 (3)
Type of study design	
RCT[a]	13 (41)
Pilot study	8 (25)
Before-and-after study	6 (19)
Feasibility study	3 (9)
Nonrandomized comparison	1 (3)
Crossover RCT	1 (3)
Study duration	
<8 weeks	25 (78)
≥8 weeks	7 (22)
Sample population	
Healthy adults	23 (72)
Inpatients/outpatients	9 (28)
Target clinical outcome	
Treatment and monitoring	21 (66)
Education and training	11 (34)
Target disorder/intervention	
Mental well-being	17 (53)
Co-occurring depression and anxiety	4 (13)
Depression only	3 (9)
Others[b]	8 (25)

[a]RCT: randomized controlled trial.

[b]Height phobia, panic disorder, anxiety only, suicide prevention, gambling disorder, attention-deficit/hyperactivity disorder, irritable bowel syndrome, and substance abuse.

Most studies (22/32, 69%) included at least one comparison group [3,27-39,41-46,50,51]. An RCT design was used in 13 studies (43%) [3,27,28,30-32,37,38,41,42,44,45,50]. Moreover, 6 studies (19%) reported before-and-after trials with no comparison group [4,13,26,47,52]. Most of the studies (23/32, 72%) were conducted on healthy adults [3,4,13,24,26-28,30-32,36,38-43,45,46,48,51,52].

The included studies were primarily focused on promoting mental well-being (17/32, 53%) by offering education and training through psychoeducation, cognitive, or behavioral training, such as mindfulness exercises (6/17, 35%) [24,33,36,42,46,48], or monitoring of well-being indicators, such as daily emotions (11/17, 65%) [13,26-28,30,39-41,45,51,52]. Fifteen studies assessed the

treatment of specific mental health disorders. Among these, 3 assessed depression only [44,49,50], 1 assessed anxiety only [29], and 4 assessed co-occurring depression and anxiety [3,32,43].

Among 21 studies that included a comparison group, most (14/21, 67%) reported that the CAs were more effective than the comparison approach [3,30-32,34-36,39,41,44,46,48-50], 4 (19%) reported mixed findings [33,37,43,51], and 5 (24%) reported no difference between the groups [27,28,38,42,45]. In studies that lasted more than a day but less than 8 weeks and provided attrition data (20/32, 69%), the average attrition rate was 19.16% (range 0%-56%) [3,4,13,31-40,43,45-47,50-52]. Three studies that lasted 4-8 weeks reported an attrition rate of 0% [13,32,52]. Two studies that lasted more than 8 weeks reported attrition rates of 35% [49] and 9% [44]. Multimedia Appendix 4 presents a detailed summary of the included studies.

Most CAs were deployed on a web-based application (7/32, 22%) [27,28,30,39,46,51,52], a standalone smartphone app (7/32, 22%) [4,13,29,34-37], or a laptop/desktop-based program (7/32, 17%) [26,31,43,47,48,50]. Among the 32 studies, 14 (44%) included embodied CAs [24,29-31,39,43-50], 11 (34%) included CAs characterized by avatars [3,4,13,32-35,37,40-42], and 7 (22%) did not specify the type of CA visualization [26-28,36,38,51,52]. Moreover, there were 18 (56%) rule-based CAs [3,24,26,30,33,34,36-38,41,42,44-50] and 14 (44%) AI-enhanced CAs [4,13,27-29,31,32,35,39,40,43,51,52]. The CAs were mostly coach-like (23/32, 72%) [3,4,13,24,29-38,40,43-49], characterized by encouraging, motivating, and nurturing personalities (Table 2). Among the 32 studies, 6 (19%) CAs were presented as a health care professional [26-28,41,50,51], 2 (6%) used informal language and conversed with users like a friend [42,52], and 1 (3%) [39] showed a knowledgeable personality based on content created and informed by medical experts [22].

Table 2. Conversational agent characteristics.

CA[a] characteristic	Value (N=32), n (%)
Type of CA	
ECA[b]	14 (44)
Avatar only	11 (34)
Not specified	7 (22)
Delivery channel	
Web-based application	7 (22)
Standalone smartphone app	7 (22)
Computer/laptop-based program	7 (22)
Messaging app–based approach[c]	6 (19)
Tablet computer	4 (13)
Hybrid approach[d]	1 (4)
Type of CA by dialogue modality	
Rule-based CA	18 (56)
AI[e]-enhanced CA	14 (44)
CA personality	
Coach-like personality	23 (72)
Health care professional–like personality	6 (19)
Informal-like personality	2 (6)
Knowledgeable personality	1 (3)

[a]CA: conversational agent.

[b]ECA: embodied conversational agent.

[c]Facebook, Slack, Telegram, or LINE.

[d]Both standalone and web-based applications.

[e]AI: artificial intelligence.

Types of Outcome Measurement Instruments

In total, there were 203 outcome measurement instruments, of which 149 were used in 1 study only. Sixteen instruments were included more than once using the same version, a translated version, or a shortened version of the instruments (Multimedia Appendix 5). Three of these instruments were reported in 3

separate studies involving the same CA, that is, MYLO [27,28,51].

All the studies included at least one clinical and one user experience outcome measurement instrument, except for 2 studies that only measured clinical outcomes [31,39] (Figure 2). Most of the outcome measurement instruments (123/203, 60.6%) measured clinical outcomes, 36.9% (75/203) measured user experience outcomes, and 1.0% (2/203) included technical outcomes measuring the accuracy of the stress feature detection [52] and the accuracy of the emphatic feedback function [43]. Moreover, 3 (1.5%) outcome measurement instruments were categorized as *others* as they measured the effectiveness of the experiment manipulation unrelated to clinical, user experience, or technical outcomes, such as whether the user paid attention to the experiment manipulation information [30]. Figure 2 describes the numbers and types of outcomes measured by the included studies.

Figure 2. Types of outcomes by the study ID.

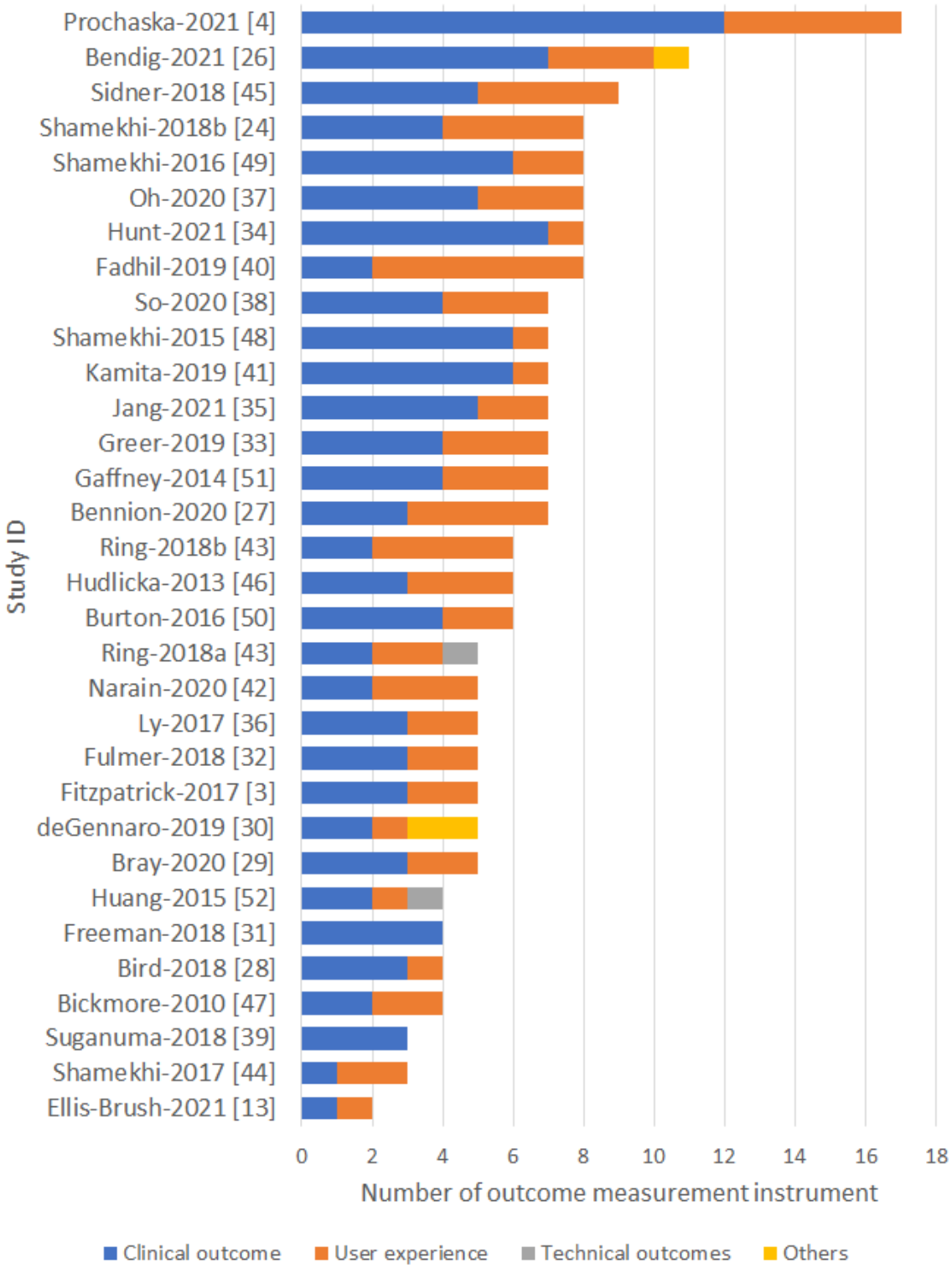


Table 3 maps all the outcome measurement instruments according to clinical, user experience, and technical outcome categories. Based on the COMET Initiative's clinical outcome taxonomy, most of the clinical outcome measurement instruments measured psychiatric outcomes (57/123, 46.3%), followed by emotional functioning/well-being outcomes (31/123, 25.2%) and cognitive functioning outcomes (21/123, 17.1%). Two studies used external sensors to objectively measure physiological changes, specifically the galvanic skin response, to assess physiological arousal [48] and breathing rate [24]. Most of the instruments measuring clinical outcomes (92/123, 74.8%) were based on published literature. A quarter of the studies (31/123, 25.2%) used original tools, and these mostly used 1-item questionnaires to assess emotional functioning/well-being outcomes (12/31, 39%).

Table 3. Frequencies of the categories and subcategories of all outcome measurement instruments.

Categories[a] and subcategories	Value (N=203), n (%[b])
Clinical outcomes[c]	
Psychiatric outcomes	57 (28.1)
Emotional functioning/well-being	32 (15.8)
Cognitive functioning	21 (10.3)
Social functioning	4 (2.0)
Adverse events	4 (2.0)
Delivery of care	2 (1.0)
Physiological data	2 (1.0)
Physical functioning	1 (0.5)
Gastrointestinal outcomes	1 (0.5)
User experience outcomes	
Subjective user experience outcomes	
User experience with the overall system	34 (16.7)
User experience with the CA[d]	15 (7.4)
User attitudes toward technology	6 (3.0)
Objective user engagement measurement[e]	
Total duration of use	10 (4.9)
Interaction with the CA	8 (3.9)
Assessment of active use	7 (3.4)
Total number of sessions	5 (2.4)
Use of specific program features	4 (2.0)
Average duration of the session	4 (2.0)
Completion of a structured module	2 (1.0)
Program use by day or week	2 (1.0)
Adherence to usage instructions	1 (0.5)
Technical outcomes	
Accuracy of the NLP[f] classifier	2 (1.0)
Other outcomes	
Experimental manipulation tests	3 (1.5)

[a]Categories are not mutually exclusive.

[b]The percentages do not add to 100% as some outcomes are mapped to two or more subcategories.

[c]Subcategories are based on the core outcome set taxonomy of clinical outcomes.

[d]CA: conversational agent.

[e]Subcategories are based on a systematic review [10] of engagement with a mobile health intervention for depression.

[f]NLP: natural language processing.

Among the user experience outcomes, the instruments were grouped into 2 major categories: subjective measures of user experience (54/75, 72%) and objective engagement measures via system log data (20/75, 27%). Thirty studies reported subjective measures of user experience. The majority (43/54, 80%) included questionnaires developed by the researchers to measure various aspects of user experience, and 12 studies (12/54, 22%) used validated or previously published instruments. Most studies (34/54, 63%) measured users' experiences of the whole system using validated questionnaires such as the System Usability Scale (SUS) [53]. Almost a third of the studies (15/54, 28%) explored user experience and satisfaction with the CA, including satisfaction and likability of the CA or the user-CA working alliance [54]. Other studies (6/54, 11%) included a questionnaire on users' attitudes toward the technology.

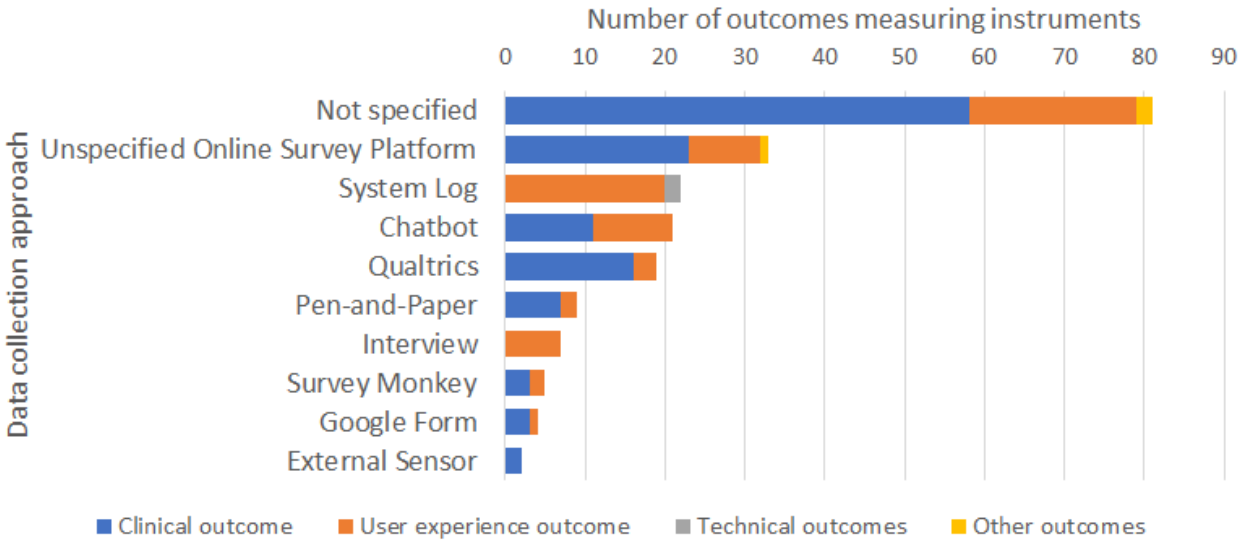
Among the 32 studies, 20 (63%) reported one or more objective engagement measures to collectively describe user engagement with the CA. Multimedia Appendix 6 details the various definitions for the system log data collected by the studies. Most studies (14/20, 70%) reported the total duration of use in general; however, some studies (7/20, 35%) included a specific definition of "active use." This included completing specific tasks, such as at least two user responses via the conversation interface within 5 minutes [33], or completing specific tasks within a session [3,4,35,36,47,50]. Other studies defined engagement based on the number of interactions with the CA [4,32,38,52].

Data Collection Method

Overall, most of the outcomes (170/203, 83.7%) were collected via self-reported questionnaires, with an average of 5 self-reported instruments per study (range 2-16). Almost all the clinical outcomes were collected via self-reported questionnaires (120/123, 97.6%). Three studies measured clinical outcomes using nonquestionnaire-based instruments, including qualitative analysis of participants' conversation logs [51] or external physiological sensors [24,48]. Similarly, most of the user experience outcomes were collected via self-reported questionnaires (47/75, 63%), followed by objective engagement measures (22/75, 28%) and qualitative interview data (7/75, 9%). No study reported any outcomes using a passive sensing module via a smartphone.

One-third of the outcomes (61/203, 30.0%) were collected using a survey platform such as Qualtrics (19/203, 9.4%) (Figure 3). One-tenth of the studies collected data via system logs (22/203, 10.8%) or directly via the CA (21/203, 10.3%). Most studies (81/203, 39.9%), however, did not report the data collection platform.

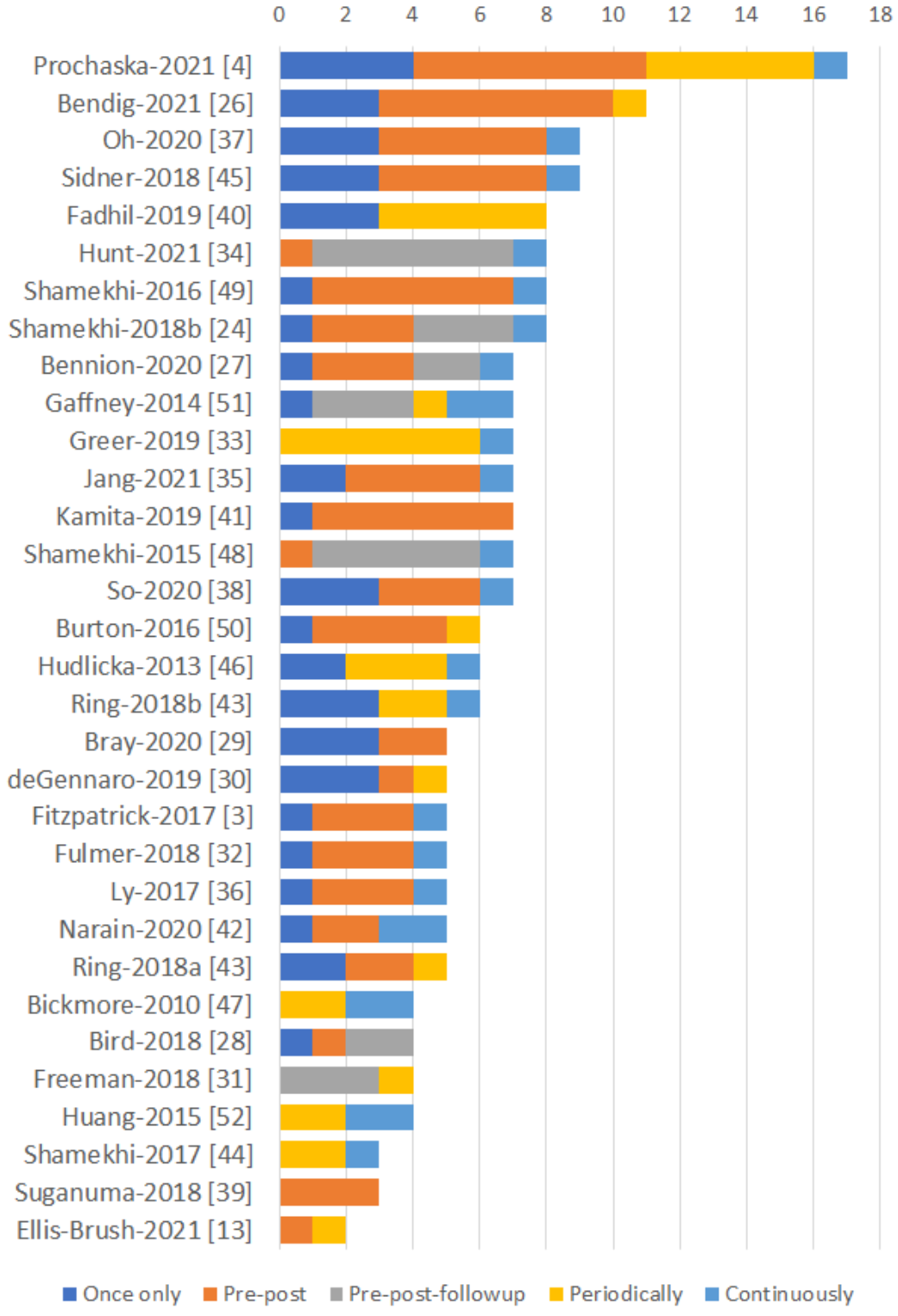
Figure 3. Data collection method employed in studies assessing conversational agent interventions in mental health.



Types of Outcomes and Measurement Time Points

Most of the clinical outcomes were measured twice, that is, at baseline and postintervention (70/123, 56.9%), followed by thrice, that is, at baseline, postintervention, and during the follow-up period typically 2 to 12 weeks after the intervention (20/123, 16.3%). A minority of the clinical outcomes were measured periodically several times during the intervention either daily (4/123, 3.3%) or weekly (7/123, 5.7%). Ecological momentary assessments were used after a specific task to measure the interventions' impact on users' mood [4,26] or to assess adherence to medication or tasks [47]. Two studies included continuous data collection via external physiological sensors [24,48]. Most user experience outcomes were measured once postintervention (33/75, 44%) or continuously during the study using system log data (20/75, 27%). Moreover, 10 (14%) studies used ecological momentary assessments after specific tasks to collect user experiences, typically after a session with the CA (Figure 4).

Figure 4. Frequency of data collection for various measurement time points. Periodically refers to daily, weekly, or posttask data collection. Continuously refers to data collection via sensor or system log data.



Validity Evidence

Validity evidence of the outcome measurement instruments was reported in less than half of the included studies (96/203, 47.3%). Among those without validity evidence, 95 (95/107, 88.8%) were researcher-designed instruments created or adapted for the study. Most clinical outcomes (83/123, 67.5%) included validity evidence, but only 12 user experience outcomes (12/75, 16%) included or cited validity evidence. One of the technical outcomes included validity evidence of the stress detection module [52]. Among the instruments that reported at least one validity evidence, most described or cited reliability statistics (82/96, 85%) or concurrent, convergent, discriminant, or construct validity statistics (91/96, 95%). A minority of the studies (8/97, 8%) cited or included responsiveness or sensitivity statistics.

Association Between Types of Outcomes

One-fifth of the studies (8/32, 25%) analyzed clinical outcomes in association with user experience outcomes [4,13,27,35,37,40,43,51]. Four studies (4/8, 50%) compared psychological symptom outcomes and various user experience outcomes, including SUS, working alliance with CAs, satisfaction, and objective engagement measurements such as duration of use [4,35,37,43]. Three studies (3/8, 38%) compared cognitive functioning, specifically self-efficacy, with user experience [4,13,40]. Two studies (2/8, 25%) compared emotional functioning/well-being and user experience outcomes [27,51]. User experiences were associated with better clinical outcomes in 4 of the studies [27,35,43,51]. Moreover, 3 studies did not find any associations between the 2 types of outcomes [13,37,40], and 1 found that better user experience was related to a reduction in some clinical outcomes but no association with others [4]. All technical outcomes were reported separately.

Discussion

Principal Findings

In this scoping review, we assessed 32 studies reporting outcome measurement instruments used to evaluate CA-delivered mental health interventions, their context of use, the method and frequency of data collection approaches, and the prevalence of validity evidence for the outcome measurement instruments. We identified 203 outcome measurement instruments, out of which 150 were unique instruments that were created or adapted specifically for the study. Most instruments measured clinical outcomes and used self-reported instruments. More than half of the instruments did not report validity evidence. In studies that reported the outcome instruments' validity evidence, reliability and construct validity were the most reported.

Comparison With Prior Work

Our review found heterogeneity in the choice of outcome measurement instruments used. This is similar to other reviews on this topic focusing on CA-delivered mental health interventions [7-9]. A recent review of commonly used measures to assess system usability and user experience further noted a lack of a unified measurement instrument that measures all aspects of usability [55]. Systematic reviews on user adherence and engagement to DHIs also strongly suggested a more unified operationalization of engagement measures and their relationship with other subjectively reported outcomes [10,56]. These results point to a need for a core outcome set specific to understanding and quantifying user experience, usage, and adherence to CAs in mental health. For instance, studies should minimally include SUS or other validated usability questionnaires tailored for CAs [57] as measures of general usability for comparison among different CA systems [58]. This is in line with a recent meta-analysis that suggested the possibility of using the SUS score to benchmark usability across all digital mental health apps [14].

In our review, few studies assessed the relationship between clinical outcomes and other outcomes. A similar review evaluated the engagement with mHealth interventions for depression and found that fewer than half of the reviewed studies assessed the relationship between objective and subjective user experience and clinical outcomes [10]. The relationship between clinical and user experience outcomes is of particular interest in digital mental health interventions. Preliminary evidence suggests an association between better user engagement with the intervention website and a greater reduction in depression and anxiety symptoms [59]. A review has further suggested that a stronger therapeutic alliance with a digital mental health intervention may have an indirect relationship with clinical outcomes [59]. Therapeutic alliance, defined as the therapeutic relationship between the patient and therapist, is fundamental to the success of face-to-face psychological therapy [59,60]. Early evidence suggests that users may develop a therapeutic alliance with the CA [61]. However, factors, such as the quality of the app and user satisfaction, may affect the bond, although the evidence is still limited [62]. Our review, for example, suggested that higher objectively and subjectively measured user experiences with the CA platform were associated with better clinical outcomes in some studies [4,27,35,43,51], but a small number did not find any associations [4,13,37,40]. Future studies should further explore the relationship between clinical and nonclinical outcomes to understand the factors affecting the efficacy of CA-delivered mental health interventions.

Our review also found that most of the outcomes were self-reported electronically via unspecified survey platforms, which may suggest an underutilization of the technological affordances within the digital space. When objectively measured data were reported, they were typically used to understand user engagement with the application, via system log data. A recent review of wearables and smartphone-based passive sensing devices for mental health monitoring suggested innovative ways to incorporate inbuilt smartphone sensors to monitor general well-being and the symptoms of bipolar disorder and depression [19]. These appear to be necessary as half of our included studies measured more than five outcomes using self-reported questionnaires, which may increase participants' input burden over time [63]. Shamekhi and Bickmore [24], for instance, found that the use of passive sensors provided a better user experience compared to having no sensor in a CA-led meditation session. Studies also typically collected self-reported data externally via survey platforms, such as Qualtrics and Survey Monkey, rather than collecting the data directly via the CA. A recent study suggested that a conversational survey collected directly by a CA is a reliable alternative to traditional surveys and may lead to improved response quality [64]. Our findings thus suggest the underutilization of the technological affordances made available by CA systems in terms of the data collection methodology.

Interestingly, our review identified only 2 studies that reported technical outcomes specifically related to the accuracy of the emotion detection modules in their ability to respond to user responses. While one study reported on technical glitches [3], this was mainly in the context of user experience and not for the entire performance of the CA during the study. These technical outcomes might be reported elsewhere as evidenced by a recent review on the complexity of technical outcomes in CA-delivered interventions for mental health [9]. However, without technical outcomes, it is difficult to fully evaluate the

effectiveness of CAs to better understand the various ways that users interact with CAs.

Lastly, our review found that mental health CAs were mostly more effective than the comparison group in the included studies. In addition, the overall mean attrition rate was relatively lower than in other DHI studies [65]. This result is supported by a recent meta-analysis of 11 trials of CA-delivered psychotherapy, which showed significantly improved depressive symptoms among adults [5]. Another meta-analysis of smartphone-delivered mental health interventions further found that the mean study attrition for short-term studies was about 35.5% (95% CI 26.7-45.3) [65], which is higher than the average attrition rate of 19% found in our review and a recently published meta-analysis of CAs for depression and anxiety [5]. The systematic review acknowledged that the findings were still preliminary due to the limitations of the included studies. We hope that future studies will benefit from the recommendations provided in our scoping review by improving the overall quality of evaluation of mental health CAs.

Strengths and Limitations

This scoping review has several strengths. First, we conducted a comprehensive literature search of multiple databases and grey literature sources. We prioritized the sensitivity of our search terms to capture the various representations of CAs used in mental health. Second, unlike other reviews in this area [66,67], our study analyzed all the outcomes included in CA-delivered interventions and provided a more granular mapping of these outcomes. This study, therefore, showcased the possible taxonomy of the outcomes measured in CA-delivered interventions that can be referenced by other researchers in this field.

Our study however has some limitations. First, given the novelty and multidisciplinary nature of the field, some unpublished literature presented at niche conferences and meetings may have been omitted. Second, during the data extraction process, we identified the validity evidence of the employed outcome measurement instruments based on the validation assessments cited in the included studies. Hence, the validity evidence captured here will more accurately reflect the reporting convention but not the actual validity of the instruments included. Some of the outcome measurement instruments used may have the necessary validity evidence but were not cited or reported by the included studies. Third, we used broad inclusion criteria that included studies using less robust experimental designs to provide a snapshot of the current state of the assessment for CA-delivered interventions in mental health.

Recommendations for Future Research

The results from our scoping review suggested the need to standardize the outcome measurement instruments used in CAs for health care and specifically mental health. This may be done via a Delphi study or via existing guidelines [68] to establish the core outcome set directly related to CA functionalities, such as the definition of meaningful engagement with the CA, or determine user attitudes and perceptions toward the CA. This is necessary as engagement and the working alliance with the CA impact the way users interact with it [69]. Our results also suggested the need to include technical outcomes, such as system crashes, out-of-scope questions, and glitches in the dialogues during the implementation. This is to better understand the relationship among the technical issues faced, user experience, and clinical outcomes. Our results showed that most studies did not report the data collection method. We recommend including the data collection method, such as including the online survey platform name, or data collection via embedded programs within the intervention. This may inform future researchers to consider the effectiveness of various evaluation platforms for future interventions. Lastly, researchers may benefit from existing frameworks to guide the incorporation of passive sensing using smartphones or wearables for mental health interventions [70] to better use the technological affordances of CAs.

Conclusion

This review suggests that studies on CA-delivered mental health interventions include a diverse set of clinical, user experience, or user engagement outcomes. Most of the measured outcomes were clinical outcomes, assessed electronically via an unspecified survey platform with uniquely created or adapted measurement instruments that lacked any reference to validity evidence. There is a need for a more consistent approach to the evaluation of these interventions, for example, through the development of guidelines with relevant experts and stakeholders. The review also suggested a greater need to capitalize on the affordances made available by CA systems and smartphones, such as passive sensing modules and conversation-based assessments, to streamline the assessment and reduce participants' input burden when using self-reported instruments.

Acknowledgments

We would like to acknowledge Ms Yasmin Ally (Lee Kong Chian School of Medicine librarian) for her assistance in translating and executing the search strategy. This research is supported by the Singapore Ministry of Education under the Singapore Ministry of Education Academic Research Fund Tier 1 (RG36/20). The research was conducted as part of the Future Health Technologies program, which was established collaboratively between ETH Zurich and the National Research Foundation, Singapore. This research is supported by the National Research Foundation, Prime Minister's Office, Singapore, under its Campus for Research Excellence and Technological Enterprise program.

Data Availability

This scoping review does not contain primary data. The data set analyzed during this study is available from the corresponding author on reasonable request.

Authors' Contributions

LTC conceptualized the study and provided supervision at all steps of the research. LTC and AIJ designed the study. AIJ and XL extracted and conducted the analysis. AIJ wrote the original manuscript. LTC, LM, XL, MS, and JT provided critical review of the manuscript. All authors approved the final version of the manuscript and take accountability for all aspects of the work.

Conflicts of Interest

None declared.

Multimedia Appendix 1

PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) checklist.
[PDF File (Adobe PDF File), 89 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Search strategies.
[DOCX File , 28 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Selected definitions for Core Outcome Measures in Effectiveness Trials (COMET)'s medical research outcome taxonomy.
[DOCX File , 25 KB-Multimedia Appendix 3]

Multimedia Appendix 4

Characteristics of the included studies.
[XLSX File (Microsoft Excel File), 36 KB-Multimedia Appendix 4]

Multimedia Appendix 5

Outcome measures included in more than one study.
[DOCX File , 29 KB-Multimedia Appendix 5]

Multimedia Appendix 6

System log data and definitions provided by the studies.
[DOCX File , 28 KB-Multimedia Appendix 6]

References

1. Laranjo L, Dunn AG, Tong HL, Kocaballi AB, Chen J, Bashir R, et al. Conversational agents in healthcare: a systematic review. J Am Med Inform Assoc 2018 Sep 01;25(9):1248-1258 [FREE Full text] [doi: 10.1093/jamia/ocy072] [Medline: 30010941]
2. Kocaballi AB, Berkovsky S, Quiroz JC, Laranjo L, Tong HL, Rezazadegan D, et al. The Personalization of Conversational Agents in Health Care: Systematic Review. J Med Internet Res 2019 Nov 07;21(11):e15360 [FREE Full text] [doi: 10.2196/15360] [Medline: 31697237]
3. Fitzpatrick KK, Darcy A, Vierhile M. Delivering Cognitive Behavior Therapy to Young Adults With Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial. JMIR Ment Health 2017 Jun 06;4(2):e19 [FREE Full text] [doi: 10.2196/mental.7785] [Medline: 28588005]
4. Prochaska JJ, Vogel EA, Chieng A, Kendra M, Baiocchi M, Pajarito S, et al. A Therapeutic Relational Agent for Reducing Problematic Substance Use (Woebot): Development and Usability Study. J Med Internet Res 2021 Mar 23;23(3):e24850 [FREE Full text] [doi: 10.2196/24850] [Medline: 33755028]
5. Lim SM, Shiau CWC, Cheng LJ, Lau Y. Chatbot-Delivered Psychotherapy for Adults With Depressive and Anxiety Symptoms: A Systematic Review and Meta-Regression. Behav Ther 2022 Mar;53(2):334-347. [doi: 10.1016/j.beth.2021.09.007] [Medline: 35227408]
6. Vaidyam AN, Wisniewski H, Halamka JD, Kashavan MS, Torous JB. Chatbots and Conversational Agents in Mental Health: A Review of the Psychiatric Landscape. Can J Psychiatry 2019 Jul;64(7):456-464 [FREE Full text] [doi: 10.1177/0706743719828977] [Medline: 30897957]
7. Gaffney H, Mansell W, Tai S. Conversational Agents in the Treatment of Mental Health Problems: Mixed-Method Systematic Review. JMIR Ment Health 2019 Oct 18;6(10):e14166 [FREE Full text] [doi: 10.2196/14166] [Medline: 31628789]
8. Vaidyam AN, Linggonegoro D, Torous J. Changes to the Psychiatric Chatbot Landscape: A Systematic Review of Conversational Agents in Serious Mental Illness: Changements du paysage psychiatrique des chatbots: une revue systématique

des agents conversationnels dans la maladie mentale sérieuse. Can J Psychiatry 2021 Apr;66(4):339-348 [FREE Full text] [doi: 10.1177/0706743720966429] [Medline: 33063526]
9. Abd-Alrazaq A, Safi Z, Alajlani M, Warren J, Househ M, Denecke K. Technical Metrics Used to Evaluate Health Care Chatbots: Scoping Review. J Med Internet Res 2020 Jun 05;22(6):e18301 [FREE Full text] [doi: 10.2196/18301] [Medline: 32442157]
10. Molloy A, Anderson PL. Engagement with mobile health interventions for depression: A systematic review. Internet Interv 2021 Dec;26:100454 [FREE Full text] [doi: 10.1016/j.invent.2021.100454] [Medline: 34621626]
11. Ng MM, Firth J, Minen M, Torous J. User Engagement in Mental Health Apps: A Review of Measurement, Reporting, and Validity. Psychiatr Serv 2019 Jul 01;70(7):538-544 [FREE Full text] [doi: 10.1176/appi.ps.201800519] [Medline: 30914003]
12. De Angel V, Lewis S, White K, Oetzmann C, Leightley D, Oprea E, et al. Digital health tools for the passive monitoring of depression: a systematic review of methods. NPJ Digit Med 2022 Jan 11;5(1):3 [FREE Full text] [doi: 10.1038/s41746-021-00548-8] [Medline: 35017634]
13. Ellis-Brush K. Augmenting Coaching Practice through digital methods. International Journal of Evidence Based Coaching and Mentoring 2021;S15:187-197. [doi: 10.24384/er2p-4857]
14. Hyzy M, Bond R, Mulvenna M, Bai L, Dix A, Leigh S, et al. System Usability Scale Benchmarking for Digital Health Apps: Meta-analysis. JMIR Mhealth Uhealth 2022 Aug 18;10(8):e37290 [FREE Full text] [doi: 10.2196/37290] [Medline: 35980732]
15. Dodd S, Clarke M, Becker L, Mavergames C, Fish R, Williamson PR. A taxonomy has been developed for outcomes in medical research to help improve knowledge discovery. J Clin Epidemiol 2018 Apr;96:84-92 [FREE Full text] [doi: 10.1016/j.jclinepi.2017.12.020] [Medline: 29288712]
16. Kane MT. Validating the Interpretations and Uses of Test Scores. Journal of Educational Measurement 2013 Mar 14;50(1):1-73. [doi: 10.1111/jedm.12000]
17. Fei J, Wolff J, Hotard M, Ingham H, Khanna S, Lawrence D, et al. Automated Chat Application Surveys Using Whatsapp: Evidence from Panel Surveys and a Mode Experiment. SSRN Journal 2022:Discussion Paper No. 15263. [doi: 10.2139/ssrn.4114839]
18. Torous J, Lipschitz J, Ng M, Firth J. Dropout rates in clinical trials of smartphone apps for depressive symptoms: A systematic review and meta-analysis. J Affect Disord 2020 Feb 15;263:413-419. [doi: 10.1016/j.jad.2019.11.167] [Medline: 31969272]
19. Sheikh M, Qassem M, Kyriacou PA. Wearable, Environmental, and Smartphone-Based Passive Sensing for Mental Health Monitoring. Front Digit Health 2021;3:662811 [FREE Full text] [doi: 10.3389/fdgth.2021.662811] [Medline: 34713137]
20. Peters MDJ, Godfrey CM, Khalil H, McInerney P, Parker D, Soares CB. Guidance for conducting systematic scoping reviews. Int J Evid Based Healthc 2015 Sep;13(3):141-146. [doi: 10.1097/XEB.0000000000000050] [Medline: 26134548]
21. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. Ann Intern Med 2018 Oct 02;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
22. Tudor Car L, Dhinagaran DA, Kyaw BM, Kowatsch T, Joty S, Theng Y, et al. Conversational Agents in Health Care: Scoping Review and Conceptual Analysis. J Med Internet Res 2020 Aug 07;22(8):e17158 [FREE Full text] [doi: 10.2196/17158] [Medline: 32763886]
23. Covidence. URL: https://www.covidence.org/ [accessed 2022-03-31]
24. Shamekhi A, Bickmore T. Breathe Deep: A Breath-Sensitive Interactive Meditation Coach. In: Proceedings of the 12th EAI International Conference on Pervasive Computing Technologies for Healthcare. 2018 Presented at: 12th EAI International Conference on Pervasive Computing Technologies for Healthcare; May 21-24, 2018; New York, NY, USA p. 108-117. [doi: 10.1145/3240925.3240940]
25. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. Qual Life Res 2010 May;19(4):539-549 [FREE Full text] [doi: 10.1007/s11136-010-9606-8] [Medline: 20169472]
26. Bendig E, Erb B, Meißner D, Bauereiß N, Baumeister H. Feasibility of a Software agent providing a brief Intervention for Self-help to Uplift psychological wellbeing ("SISU"). A single-group pretest-posttest trial investigating the potential of SISU to act as therapeutic agent. Internet Interv 2021 Apr;24:100377 [FREE Full text] [doi: 10.1016/j.invent.2021.100377] [Medline: 33816127]
27. Bennion MR, Hardy GE, Moore RK, Kellett S, Millings A. Usability, Acceptability, and Effectiveness of Web-Based Conversational Agents to Facilitate Problem Solving in Older Adults: Controlled Study. J Med Internet Res 2020 May 27;22(5):e16794 [FREE Full text] [doi: 10.2196/16794] [Medline: 32384055]
28. Bird T, Mansell W, Wright J, Gaffney H, Tai S. Manage Your Life Online: A Web-Based Randomized Controlled Trial Evaluating the Effectiveness of a Problem-Solving Intervention in a Student Sample. Behav Cogn Psychother 2018 Sep;46(5):570-582 [FREE Full text] [doi: 10.1017/S1352465817000820] [Medline: 29366432]

29. Bray L, Sharpe A, Gichuru P, Fortune P, Blake L, Appleton V. The Acceptability and Impact of the Xploro Digital Therapeutic Platform to Inform and Prepare Children for Planned Procedures in a Hospital: Before and After Evaluation Study. J Med Internet Res 2020 Aug 11;22(8):e17367 [FREE Full text] [doi: 10.2196/17367] [Medline: 32780025]
30. de Gennaro M, Krumhuber EG, Lucas G. Effectiveness of an Empathic Chatbot in Combating Adverse Effects of Social Exclusion on Mood. Front Psychol 2019;10:3061 [FREE Full text] [doi: 10.3389/fpsyg.2019.03061] [Medline: 32038415]
31. Freeman D, Haselton P, Freeman J, Spanlang B, Kishore S, Albery E, et al. Automated psychological therapy using immersive virtual reality for treatment of fear of heights: a single-blind, parallel-group, randomised controlled trial. Lancet Psychiatry 2018 Aug;5(8):625-632 [FREE Full text] [doi: 10.1016/S2215-0366(18)30226-8] [Medline: 30007519]
32. Fulmer R, Joerin A, Gentile B, Lakerink L, Rauws M. Using Psychological Artificial Intelligence (Tess) to Relieve Symptoms of Depression and Anxiety: Randomized Controlled Trial. JMIR Ment Health 2018 Dec 13;5(4):e64 [FREE Full text] [doi: 10.2196/mental.9782] [Medline: 30545815]
33. Greer S, Ramo D, Chang Y, Fu M, Moskowitz J, Haritatos J. Use of the Chatbot "Vivibot" to Deliver Positive Psychology Skills and Promote Well-Being Among Young People After Cancer Treatment: Randomized Controlled Feasibility Trial. JMIR Mhealth Uhealth 2019 Oct 31;7(10):e15018 [FREE Full text] [doi: 10.2196/15018] [Medline: 31674920]
34. Hunt M, Miguez S, Dukas B, Onwude O, White S. Efficacy of Zemedy, a Mobile Digital Therapeutic for the Self-management of Irritable Bowel Syndrome: Crossover Randomized Controlled Trial. JMIR Mhealth Uhealth 2021 May 20;9(5):e26152 [FREE Full text] [doi: 10.2196/26152] [Medline: 33872182]
35. Jang S, Kim J, Kim S, Hong J, Kim S, Kim E. Mobile app-based chatbot to deliver cognitive behavioral therapy and psychoeducation for adults with attention deficit: A development and feasibility/usability study. Int J Med Inform 2021 Jun;150:104440. [doi: 10.1016/j.ijmedinf.2021.104440] [Medline: 33799055]
36. Ly KH, Ly A, Andersson G. A fully automated conversational agent for promoting mental well-being: A pilot RCT using mixed methods. Internet Interv 2017 Dec;10:39-46 [FREE Full text] [doi: 10.1016/j.invent.2017.10.002] [Medline: 30135751]
37. Oh J, Jang S, Kim H, Kim J. Efficacy of mobile app-based interactive cognitive behavioral therapy using a chatbot for panic disorder. Int J Med Inform 2020 Aug;140:104171. [doi: 10.1016/j.ijmedinf.2020.104171] [Medline: 32446158]
38. So R, Furukawa TA, Matsushita S, Baba T, Matsuzaki T, Furuno S, et al. Unguided Chatbot-Delivered Cognitive Behavioural Intervention for Problem Gamblers Through Messaging App: A Randomised Controlled Trial. J Gambl Stud 2020 Dec;36(4):1391-1407. [doi: 10.1007/s10899-020-09935-4] [Medline: 32162075]
39. Suganuma S, Sakamoto D, Shimoyama H. An Embodied Conversational Agent for Unguided Internet-Based Cognitive Behavior Therapy in Preventative Mental Health: Feasibility and Acceptability Pilot Trial. JMIR Ment Health 2018 Jul 31;5(3):e10454 [FREE Full text] [doi: 10.2196/10454] [Medline: 30064969]
40. Fadhil A, Wang Y, Reiterer H. Assistive Conversational Agent for Health Coaching: A Validation Study. Methods Inf Med 2019 Jun;58(1):9-23. [doi: 10.1055/s-0039-1688757] [Medline: 31117129]
41. Kamita T, Ito T, Matsumoto A, Munakata T, Inoue T. A Chatbot System for Mental Healthcare Based on SAT Counseling Method. Mobile Information Systems 2019 Mar 03;2019:1-11. [doi: 10.1155/2019/9517321]
42. Narain J, Quach T, Davey M, Park H, Breazeal C, Picard R. Promoting Wellbeing with Sunny, a Chatbot that Facilitates Positive Messages within Social Groups. In: CHI EA '20: Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems. 2020 Presented at: 2020 CHI Conference on Human Factors in Computing Systems; April 25-30, 2020; Honolulu, HI, USA. [doi: 10.1145/3334480.3383062]
43. Ring L. An affect-aware dialogue system for counseling. Northeastern University Library. 2018. URL: https://repository.library.northeastern.edu/files/neu:cj82qk940 [accessed 2023-04-06]
44. Shamekhi A, Bickmore T, Lestoquoy A, Gardiner P. Augmenting Group Medical Visits with Conversational Agents for Stress Management Behavior Change. In: de Vries P, Oinas-Kukkonen H, Siemons L, Beerlage-de Jong N, van Gemert-Pijnen L, editors. Persuasive Technology: Development and Implementation of Personalized Technologies to Change Attitudes and Behaviors. PERSUASIVE 2017. Lecture Notes in Computer Science, vol 10171. Cham: Springer; 2017.
45. Sidner CL, Bickmore T, Nooraie B, Rich C, Ring L, Shayganfar M, et al. Creating New Technologies for Companionable Agents to Support Isolated Older Adults. ACM Trans. Interact. Intell. Syst 2018 Jul 24;8(3):1-27. [doi: 10.1145/3213050]
46. Hudlicka E. Virtual training and coaching of health behavior: example from mindfulness meditation training. Patient Educ Couns 2013 Aug;92(2):160-166 [FREE Full text] [doi: 10.1016/j.pec.2013.05.007] [Medline: 23809167]
47. Bickmore TW, Puskar K, Schlenk EA, Pfeifer LM, Sereika SM. Maintaining reality: Relational agents for antipsychotic medication adherence. Interacting with Computers 2010 Jul;22(4):276-288. [doi: 10.1016/j.intcom.2010.02.001]
48. Shamekhi A, Bickmore T. Breathe with Me: A Virtual Meditation Coach. In: Brinkman WP, Broekens J, Heylen D, editors. Intelligent Virtual Agents. IVA 2015. Lecture Notes in Computer Science, vol 9238. Cham: Springer; 2015:279-282.
49. Shamekhi A, Bickmore T, Lestoquoy A, Negash L, Gardiner P. Blissful Agents: Adjuncts to Group Medical Visits for Chronic Pain and Depression. In: Traum D, Swartout W, Khooshabeh P, Kopp S, Scherer S, Leuski A, editors. Intelligent Virtual Agents. IVA 2016. Lecture Notes in Computer Science, vol 10011. Cham: Springer; 2016:433-437.
50. Burton C, Szentagotai Tatar A, McKinstry B, Matheson C, Matu S, Moldovan R, Help4Mood Consortium. Pilot randomised controlled trial of Help4Mood, an embodied virtual agent-based system to support treatment of depression. J Telemed Telecare 2016 Sep;22(6):348-355. [doi: 10.1177/1357633X15609793] [Medline: 26453910]

51. Gaffney H, Mansell W, Edwards R, Wright J. Manage Your Life Online (MYLO): a pilot trial of a conversational computer-based intervention for problem solving in a student sample. Behav Cogn Psychother 2014 Nov;42(6):731-746. [doi: 10.1017/S135246581300060X] [Medline: 23899405]
52. Huang J, Li Q, Xue Y, Cheng T, Xu S, Jia J, et al. TeenChat: A Chatterbot System for Sensing and Releasing Adolescents' Stress. In: Yin X, Ho K, Zeng D, Aickelin U, Zhou R, Wang H, editors. Health Information Science. HIS 2015. Lecture Notes in Computer Science, vol 9085. Cham: Springer; 2015:133-145.
53. Brooke J. SUS: A 'Quick and Dirty' Usability Scale. In: Jordan PW, Thomas B, McClelland IL, Weerdmeester B, editors. Usability Evaluation In Industry. London: CRC Press; 1996.
54. Bickmore T, Gruber A, Picard R. Establishing the computer-patient working alliance in automated health behavior change interventions. Patient Educ Couns 2005 Oct;59(1):21-30. [doi: 10.1016/j.pec.2004.09.008] [Medline: 16198215]
55. Hodrien A, Fernando T. A Review of Post-Study and Post- Task Subjective Questionnaires to Guide Assessment of System Usability. Journal of Usability Studies 2021;16(3):203-232 [FREE Full text]
56. Sieverink F, Kelders SM, van Gemert-Pijnen JE. Clarifying the Concept of Adherence to eHealth Technology: Systematic Review on When Usage Becomes Adherence. J Med Internet Res 2017 Dec 06;19(12):e402 [FREE Full text] [doi: 10.2196/jmir.8578] [Medline: 29212630]
57. Holmes S, Moorhead A, Bond R, Zheng H, Coates V, Mctear M. Usability testing of a healthcare chatbot: Can we use conventional methods to assess conversational user interfaces? In: ECCE '19: Proceedings of the 31st European Conference on Cognitive Ergonomics. 2019 Presented at: 31st European Conference on Cognitive Ergonomics; September 10-13, 2019; Belfast, United Kingdom. [doi: 10.1145/3335082.3335094]
58. Kocabalil A, Laranjo L, Coiera E. Measuring user experience in conversational interfaces: a comparison of six questionnaires. In: HCI '18: Proceedings of the 32nd International BCS Human Computer Interaction Conference. 2018 Presented at: 32nd International BCS Human Computer Interaction Conference; July 4-6, 2018; Belfast, United Kingdom. [doi: 10.14236/ewic/HCI2018.21]
59. Tremain H, McEnery C, Fletcher K, Murray G. The Therapeutic Alliance in Digital Mental Health Interventions for Serious Mental Illnesses: Narrative Review. JMIR Ment Health 2020 Aug 07;7(8):e17204 [FREE Full text] [doi: 10.2196/17204] [Medline: 32763881]
60. Henson P, Wisniewski H, Hollis C, Keshavan M, Torous J. Digital mental health apps and the therapeutic alliance: initial review. BJPsych Open 2019 Jan;5(1):e15 [FREE Full text] [doi: 10.1192/bjo.2018.86] [Medline: 30762511]
61. Beatty C, Malik T, Meheli S, Sinha C. Evaluating the Therapeutic Alliance With a Free-Text CBT Conversational Agent (Wysa): A Mixed-Methods Study. Front Digit Health 2022;4:847991 [FREE Full text] [doi: 10.3389/fdgth.2022.847991] [Medline: 35480848]
62. Berry K, Salter A, Morris R, James S, Bucci S. Assessing Therapeutic Alliance in the Context of mHealth Interventions for Mental Health Problems: Development of the Mobile Agnew Relationship Measure (mARM) Questionnaire. J Med Internet Res 2018 Apr 19;20(4):e90 [FREE Full text] [doi: 10.2196/jmir.8252] [Medline: 29674307]
63. Druce KL, Dixon WG, McBeth J. Maximizing Engagement in Mobile Health Studies: Lessons Learned and Future Directions. Rheum Dis Clin North Am 2019 May;45(2):159-172 [FREE Full text] [doi: 10.1016/j.rdc.2019.01.004] [Medline: 30952390]
64. Celino I, Re Calegari G. Submitting surveys via a conversational interface: An evaluation of user acceptance and approach effectiveness. International Journal of Human-Computer Studies 2020 Jul;139:102410. [doi: 10.1016/j.ijhcs.2020.102410]
65. Linardon J, Fuller-Tyszkiewicz M. Attrition and adherence in smartphone-delivered interventions for mental health problems: A systematic and meta-analytic review. J Consult Clin Psychol 2020 Jan;88(1):1-13. [doi: 10.1037/ccp0000459] [Medline: 31697093]
66. 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 [FREE Full text] [doi: 10.1016/j.ijmedinf.2019.103978] [Medline: 31622850]
67. Denecke K, May R. Usability Assessment of Conversational Agents in Healthcare: A Literature Review. Stud Health Technol Inform 2022 May 25;294:169-173. [doi: 10.3233/SHTI220431] [Medline: 35612050]
68. Denecke K, Warren J. How to Evaluate Health Applications with Conversational User Interface? Stud Health Technol Inform 2020 Jun 16;270:976-980. [doi: 10.3233/SHTI200307] [Medline: 32570527]
69. Nißen M, Rüegger D, Stieger M, Flückiger C, Allemand M, V Wangenheim F, et al. The Effects of Health Care Chatbot Personas With Different Social Roles on the Client-Chatbot Bond and Usage Intentions: Development of a Design Codebook and Web-Based Study. J Med Internet Res 2022 Apr 27;24(4):e32630 [FREE Full text] [doi: 10.2196/32630] [Medline: 35475761]
70. Mendes JPM, Moura IR, Van de Ven P, Viana D, Silva FJS, Coutinho LR, et al. Sensing Apps and Public Data Sets for Digital Phenotyping of Mental Health: Systematic Review. J Med Internet Res 2022 Feb 17;24(2):e28735 [FREE Full text] [doi: 10.2196/28735] [Medline: 35175202]

Abbreviations

AI: artificial intelligence

CA: conversational agent
COMET: Core Outcome Measures in Effectiveness Trials
DHI: digital health intervention
mHealth: mobile health
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews
RCT: randomized controlled trial
SUS: System Usability Scale

Edited by T Leung; submitted 23.11.22; peer-reviewed by E Kim, L Bai; comments to author 15.02.23; revised version received 01.03.23; accepted 31.03.23; published 19.04.23

Please cite as:
Jabir AI, Martinengo L, Lin X, Torous J, Subramaniam M, Tudor Car L
Evaluating Conversational Agents for Mental Health: Scoping Review of Outcomes and Outcome Measurement Instruments
J Med Internet Res 2023;25:e44548
URL: https://www.jmir.org/2023/1/e44548
doi: 10.2196/44548
PMID: 37074762