Review

Technical Metrics Used to Evaluate Health Care Chatbots: Scoping Review

Alaa Abd-Alrazaq[1], BSc, MSc, PhD; Zeineb Safi[1], BSc, MSc; Mohannad Alajlani[2], BSc, MSc, MBA, PhD; Jim Warren[3], BS, PhD; Mowafa Househ[1], BCom, MEng, PhD; Kerstin Denecke[4], Dr rer nat

[1]College of Science and Engineering, Hamad Bin Khalifa University, Doha, Qatar

[2]Institute of Digital Healthcare, University of Warwick, Coventry, United Kingdom

[3]School of Computer Science, University of Auckland, Auckland, New Zealand

[4]Institute for Medical Informatics, Bern University of Applied Sciences, Bern, Switzerland

Corresponding Author:
Kerstin Denecke, Dr rer nat
Institute for Medical Informatics
Bern University of Applied Sciences
Quellgasse 21
2502 Biel
Bern
Switzerland
Phone: 41 76 409 97 61
Email: kerstin.denecke@bfh.ch

Abstract

Background: Dialog agents (chatbots) have a long history of application in health care, where they have been used for tasks such as supporting patient self-management and providing counseling. Their use is expected to grow with increasing demands on health systems and improving artificial intelligence (AI) capability. Approaches to the evaluation of health care chatbots, however, appear to be diverse and haphazard, resulting in a potential barrier to the advancement of the field.

Objective: This study aims to identify the technical (nonclinical) metrics used by previous studies to evaluate health care chatbots.

Methods: Studies were identified by searching 7 bibliographic databases (eg, MEDLINE and PsycINFO) in addition to conducting backward and forward reference list checking of the included studies and relevant reviews. The studies were independently selected by two reviewers who then extracted data from the included studies. Extracted data were synthesized narratively by grouping the identified metrics into categories based on the aspect of chatbots that the metrics evaluated.

Results: Of the 1498 citations retrieved, 65 studies were included in this review. Chatbots were evaluated using 27 technical metrics, which were related to chatbots as a whole (eg, usability, classifier performance, speed), response generation (eg, comprehensibility, realism, repetitiveness), response understanding (eg, chatbot understanding as assessed by users, word error rate, concept error rate), and esthetics (eg, appearance of the virtual agent, background color, and content).

Conclusions: The technical metrics of health chatbot studies were diverse, with survey designs and global usability metrics dominating. The lack of standardization and paucity of objective measures make it difficult to compare the performance of health chatbots and could inhibit advancement of the field. We suggest that researchers more frequently include metrics computed from conversation logs. In addition, we recommend the development of a framework of technical metrics with recommendations for specific circumstances for their inclusion in chatbot studies.

(J Med Internet Res 2020;22(6):e18301) doi: 10.2196/18301

KEYWORDS

chatbots; conversational agents; health care; evaluation; metrics

Introduction

Background

The potential of human-computer dialog to provide health care benefits has been perceived for many decades. In 1966, Weizenbaum's ELIZA system caught the public imagination with its imitation of a psychotherapist through the relatively simple linguistic token manipulation possible at the time [1]. From the mid-1990s, a family of interventions based on automated telephone sessions (telephone-linked care) demonstrated effectiveness in promoting health adherence across a range of behaviors including medication, diet, and physical activity [2]. As mobile phones have become commonplace, a range of SMS text messaging–based interventions have been developed and trialed, with particular success in smoking cessation [3]. At the same time, internet/web-based interventions have shown the ability to promote positive health behavior change [4,5], and the interaction components associated with users sticking with an internet intervention are increasingly well understood and include the inclusion of dialog elements [6].

With the advent of smartphones, the distribution of highly interactive chatbots has been greatly facilitated, particularly with the ubiquitous use of app stores and wide installation of chat apps that can include chatbots, notably Facebook Messenger. Chatbots, as with other electronic health (eHealth) interventions, offer scalability and 24-hour availability to plug gaps in unmet health needs. For example, Woebot delivers cognitive behavior therapy and has been tested with students with depression [7]. The students who used Woebot significantly reduced their symptoms of depression over the study period as measured by the depression questionnaire PHQ-9, while those in the information control group (who instead read a self-help book) did not [7]. In recent years, artificial intelligence (AI) based on deep learning has created waves with its ability to outperform physicians at some diagnostic tasks [8,9]. XiaoIce is a social chatbot that emphasizes emotional connection and it has communicated with over 660 million active users since its launch in 2014 [10]; its performance shows that deep learning can be successfully applied to meaningful dialog tasks. Combining the factors of ease-of-distribution, successful applications, and AI methods to improve health chatbot performance, it is reasonable to expect health chatbots in increasing numbers and variety to take an increasingly serious role in the health care system.

Research Problem and Aim

To be an evidence-based discipline requires measurement of performance. The impact of health chatbots on clinical outcomes is the ultimate measure of success. For example, did the condition (eg, depression, diabetes) improve to a statistically significant degree on an accepted measure (eg, PHQ-9 [11] or hemoglobin A1c [12], respectively), as compared to a control group? Such studies, however, may require large sample sizes to detect the effect and provide relatively little insight into the mechanism by which the chatbot achieves the change; additionally, studies may provide particularly little insight if the result is negative.

As an alternative and useful precursor to clinical outcome metrics, technical metrics concern the performance of the chatbot itself (eg, did participants feel that it was usable, give appropriate responses, and understand their input?). Appropriateness refers to the relevance of the provided information in addressing the problem prompted [13]. Furthermore, this includes more objective measures of the chatbot interaction, such as the number of conversational turns taken in a session or time taken, and measures that require some interpretation but are still well-defined, such as task completion. These technical measures offer a potential method for comparison of health chatbots and for understanding the use and performance of a chatbot to decide if it is working well enough to warrant the time and expense of a trial to measure clinical outcomes.

Previously, we had introduced a framework for evaluation measures of health chatbots to provide guidance to developers [14]. The framework development, however, was based on a relatively informal process vulnerable to the authors' biases in terms of what studies were considered in its formulation. Therefore, the aim of this study is to use a rigorous review methodology to identify the technical metrics used by previous studies to evaluate health care chatbots. The final goal of these efforts is to be able to make recommendations for an evaluation framework for health chatbots.

Methods

Overview

To achieve the aforementioned objective, a scoping review was conducted. To conduct a transparent and replicable review, we followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) Extension for Scoping Reviews (PRISMA-ScR) guidelines [15]. This research was conducted by an interdisciplinary team of researchers with backgrounds in nursing, computer science, and mental health applications.

Search Strategy

Search Sources

For this review, we searched the following bibliographic databases November 1-3, 2019: MEDLINE (via EBSCO), EMBASE (Excerpta Medica Database; via Ovid), PsycINFO (via Ovid), CINAHL (Cumulative Index of Nursing and Allied Health Literature; via EBSCO), IEEE (Institute of Electrical and Electronics Engineers) Xplore, ACM (Association for Computing Machinery) Digital Library, and Google Scholar. We screened only the first 100 hits retrieved by searching Google Scholar, as it usually retrieves several thousand references ordered by their relevance to the search topic. We checked the reference list of the included studies to identify further studies relevant to the current review (ie, backward reference list checking). Additionally, we used the "cited by" function available in Google Scholar to find and screen studies that cited the included studies (ie, forward reference list checking).

Search Terms

The search terms were derived from previously published literature and the opinions of informatics experts. For health-related databases, we used search terms related to the intervention of interest (eg, chatbot, conversational agent, and chat-bot). In addition to intervention-related terms, we used terms related to the context (eg, health, disease, and medical) for non–health-related databases (eg, IEEE and ACM digital library). Multimedia Appendix 1 details the search strings used for searching each electronic database.

Study Eligibility Criteria

The intervention of interest in this review was chatbots that are aimed at delivering health care services to patients. Chatbots implemented in stand-alone software or web-based platforms were included. However, we excluded chatbots operated by a human (Wizard-of-Oz) or integrated into robotics, serious games, SMS text messaging, or telephone systems. To be included, studies had to report a technical evaluation of a chatbot (eg, usability, classifier performance, and word error rate). We included peer-reviewed articles, dissertations, and conference proceedings, and we excluded reviews, proposals, editorials, and conference abstracts. This review included studies written in the English language only. No restrictions were considered regarding the study design, study setting, year of publication, and country of publication.

Study Selection

Authors MA and ZS independently screened the titles and abstracts of all retrieved references and then independently read the full texts of included studies. Any disagreements between the two reviewers were resolved by AA. We assessed the intercoder agreement by calculating Cohen, which was 0.82 for screening titles and abstracts and 0.91 for reading full texts, indicating a very good agreement [16].

Data Extraction

To conduct a reliable and accurate extraction of data from the included studies, a data extraction form was developed and piloted using 8 included studies (Multimedia Appendix 2). The data extraction process was independently conducted by two reviewers (MA and ZS) and a third reviewer (AA) resolved any disagreements. Intercoder agreement between the reviewers was good (Cohen κ=0.67).

Data Synthesis

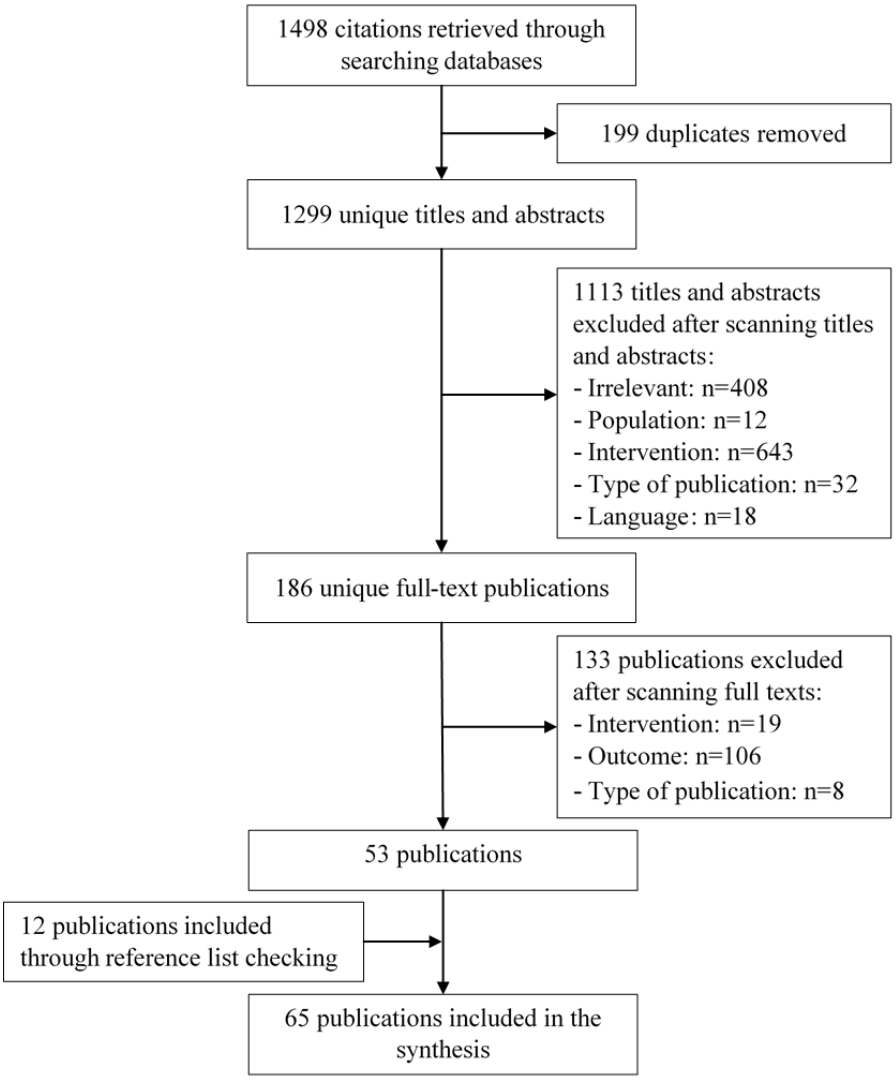
A narrative approach was used to synthesize the extracted data. After identifying all technical metrics used by the included studies to evaluate chatbots, we divided them into 4 categories based on the aspect of chatbots that the metrics evaluate. The 4 categories were formed after a discussion by the authors in which consensus was reached. For each metric, we identified how the studies measured it. Data synthesis was managed using Microsoft Excel (Microsoft Corporation).

Results

Search Results

By searching the 7 electronic databases, 1498 citations were retrieved. After removing 199 (13.3%) duplicates of these citations, 1299 (86.7%) titles and abstracts were screened. The screening process resulted in excluding 1113 (74.3%) titles and abstracts due to several reasons detailed in Figure 1. When we read the full text of the remaining 186 (12.4%) citations, a further 133 (8.9%) citations were excluded (Figure 1). In total, 12 studies were found by backward and forward reference checking. We included 65 studies in the review.

Figure 1. Flowchart of the study selection process.



Description of Included Studies

Characteristics of the included studies are detailed in Table 1. Cross-sectional survey was the most commonly used study design (n=41, 63%). About 57% (n=37) of the included studies were published as journal articles. Half of the studies (n=33, 51%) were conducted in the United States. Approximately 70% (n=45) of the studies were published between 2015 and 2019.

The sample size was reported in 61 studies, and 38 studies (62%) had 50 or fewer participants. In 44 studies, the age of participants was reported; the mean age of participants was 39 years, with a range of 13-79 years. Sex of participants was reported in 54 studies, where the mean percentage of males was 48.1%. Of the 62 studies that reported participants' health conditions, 34 (54.8%) studies recruited participants from a clinical population (ie, those with health issues). Participants were recruited from clinical settings (n=30, 49.2%), community (n=20, 32.8%), and educational settings (n=18, 29.5%). Metadata and population characteristics of each included study are presented in Multimedia Appendix 3.

Chatbots were used for self-management (n=17, 26.2%), therapeutic purposes (n=12, 18.5%), counselling (n=12, 18.5%), education (n=10, 15.4%), screening (n=9, 13.8%), training (n=7, 10.8%), and diagnosing (n=3, 4.6%). Although chatbots were implemented in stand-alone software in about 62% (n=40) of studies, chatbots were implemented in web-based platforms in the remaining studies (n=25, 39%). Chatbot responses were generated based on predefined rules, machine learning approaches, or both methods (hybrid) in 82% (N=53), 17% (n=11), and 2% (n=1) of the included studies, respectively. In the majority of studies (n=58, 89%), chatbots led the dialogue. In about 62% (n=40) of studies, users interacted with chatbots only by typing in their utterances (texts). The most common modalities used by chatbots to interact with users were a combination of text, voice, and nonverbal language (ie, facial expression and body language; n=21, 32%), text only (n=20, 31%), and a combination of voice and nonverbal language (n=19, 29%). The most common disorders targeted by chatbots were any health condition (n=20, 31%) and depression (n=15, 23%). Multimedia Appendix 4 displays characteristics of the intervention in each included study.

Table 1. Characteristics of the included studies (N=65).

Parameters and characteristics	Studies, n (%)[a]
Study metadata	
Study design	
Survey	41 (63)
Quasi-experiment	11 (17)
Randomized controlled trial	13 (20)
Type of publication	
Journal article	37 (57)
Conference proceeding	25 (39)
Thesis	3 (5)
Country	
United States	33 (51)
France	5 (8)
Netherlands	3 (5)
Japan	3 (5)
Australia	3 (5)
Italy	2 (3)
Switzerland and Netherlands	2 (3)
Finland	1 (2)
Sweden	1 (2)
Turkey	1 (2)
United Kingdom	1 (2)
Switzerland & Germany	1 (2)
Mexico	1 (2)
Spain	1 (2)
Global population	1 (2)
Romania, Spain and Scotland	1 (2)
Philippines	1 (2)
Switzerland	1 (2)
New Zealand	1 (2)
Spain and New Zealand	1 (2)
South Africa	1 (2)
Year of publication	
Before 2010	3 (5)
2010-2014	17 (26)
2015-2019	45 (70)
Population characteristics	
Sample size[b]	
≤50	38 (62)
51-100	11 (18)
101-200	9 (15)
>200	3 (5)

Parameters and characteristics	Studies, n (%)[a]
Age (years)[c]	
Mean (range)	39 (13-79)
Sex (%)[e]	
Male	48.1
Health condition[f]	
Clinical sample	34 (55)
Nonclinical sample	28 (45)
Setting[g,h]	
Clinical	30 (50)
Community	20 (33)
Educational	18 (30)
Intervention characteristics	
Purpose[i]	
Self-management	17 (26)
Therapy	12 (19)
Counselling	12 (19)
Education	10 (15)
Screening	9 (14)
Training	7 (11)
Diagnosing	3 (5)
Platform	
Stand-alone software	40 (62)
Web-based	25 (39)
Response generation	
Rule-based	53 (82)
Artificial intelligence	11 (17)
Hybrid	1 (2)
Dialogue initiative	
Chatbot	58 (89)
Users	4 (6)
Both	3 (5)
Input modality	
Text	40 (62)
Voice	9 (14)
Voice and nonverbal	8 (12)
Text and voice	6 (9)
Text and nonverbal	2 (3)
Output modality	
Text, voice and nonverbal	21 (32)
Text	20 (31)
Voice and nonverbal	19 (29)

Parameters and characteristics	Studies, n (%)[a]
Text & voice	4 (6)
Voice	1 (2)
Targeted disorders[j]	
Any health condition	20 (31)
Depression	15 (23)
Autism	5 (8)
Anxiety	5 (8)
Substance use disorder	5 (8)
Posttraumatic stress disorder	5 (8)
Mental disorders	3 (5)
Sexually transmitted diseases	3 (5)
Sleep disorders	2 (3)
Diabetes	2 (3)
Alzheimer	1 (2)
Asthma	1 (2)
Cervical cancer	1 (2)
Dementia	1 (2)
Schizophrenia	1 (2)
Stress	1 (2)
Genetic variants	1 (2)
Cognitive impairment	1 (2)
Atrial fibrillation	1 (2)

[a]Percentages were rounded and may not sum to 100.

[b]Sample size was reported in 61 studies.

[c]Mean age was reported in 44 studies.

[d]N/A: not applicable.

[e]Sex was reported in 54 studies.

[f]Sample type was reported in 62 studies.

[g]Setting was reported in 61 studies.

[h]Numbers do not add up as several chatbots focused on more than one health condition.

[i]Numbers do not add up as several chatbots have more than one purpose.

[j]Numbers do not add up as several chatbots target more than one health condition.

Results of Studies

Overview

The included studies evaluated chatbots using many technical metrics, which were categorized into 4 main groups: metrics related to chatbots as a whole (global metrics), metrics related to response generation, metrics related to response understanding, and metrics related to esthetics. More details about metrics are presented in the following sections.

Global Metrics

The included studies evaluated chatbots as a whole using the following metrics: usability, classifier performance, speed, technical issues, intelligence, task completion rate, dialogue efficiency, dialogue handling, context awareness, and error management.

Usability of chatbots was assessed in 37 (56.9%) studies [17-53]. Usability was evaluated using a single question in a self-administrated questionnaire [17,20-25,29-31,33,34,36,37,40,42,44,45,47-51,53], multiple questions in a self-administrated questionnaire [28,41,43], a specific questionnaire (eg, system usability scale [SUS] questionnaire) [18,26,27,32,35,38,39,46,52], or observation [19].

Classifier performance of chatbots was evaluated in 8 (12.3%) studies [54-61]. Many metrics were used to measure the classifier performance, namely: area under curve [54,55,60,61], accuracy [56-58,61], sensitivity [55,57,59,60], specificity

[55,57,59,60], positive predictive value [55,57,60], and negative predictive value [55,60]. The speed of chatbots was examined in 4 studies [29,35,53,62]. The speed was evaluated using a single question in a self-administrated questionnaire [29,35], multiple questions in a self-administrated questionnaire [53], and interviews [62].

Technical issues (eg, errors/glitches) in chatbots were examined in 4 studies (6.2%) [7,36,51,63]. Technical issues were assessed through interviews [7,51,63], a single question in a self-administrated questionnaire [36], and checking staff logs [51]. In addition, 3 studies assessed the intelligence of chatbots using either multiple questions in a self-administrated questionnaire [41,64] or a single question in a self-administrated questionnaire [27]. In 3 studies, the task completion rate was examined by checking the conversation logs [38,53,65].

Of the reviewed studies, 2 (3.1%) studies examined chatbot flexibility in dialogue handling (eg, its ability to maintain a conversation and deal with users' generic questions or responses that require more, less, or different information than was requested) using interviews [27] and multiple questions in a self-administrated questionnaire [38]. Dialogue efficiency of chatbots, which refers to the number of dialogue steps required to finish a task, was assessed in 1 study by reviewing transcribed conversation logs [38]. The same study examined the chatbot's context awareness (ie, its ability to utilize contextual knowledge to appropriately respond to users) using multiple questions in a self-administrated questionnaire [38]. Error management, which refers to a chatbot's ability to detect and understand misspelled words in users' replies (eg, "anious" instead of anxious), was examined in only 1 study [27].

Metrics Related to Response Generation

The following metrics were utilized by the included studies to evaluate response generation by chatbots: appropriateness of responses, comprehensibility, realism, speed of response, empathy, repetitiveness, clarity of speech, and linguistic accuracy.

Of the reviewed studies, 15 (23.1%) examined the appropriateness and adequacy of verbal [18,19,27,28,31,34,38,39,51,58,66-69] and nonverbal responses of chatbots [32]. Appropriateness of responses was assessed using interviews [18,19,31,34,51,66,68], a single question in self-administrated questionnaire [27,32,39,67], conversation logs [38,58,69], and multiple questions in self-administrated questionnaire [28].

Comprehensibility of responses, which refers to the degree to which a chatbot generates responses understandable by users, was evaluated by 14 (21.5%) studies [20,23,31,34,36,39,42,44,51,52,59,60,63,69]. Comprehensibility of responses was evaluated using a single question in a self-administrated questionnaire [20,23,31,36,39,42,44,52,59,60,63,69] and interviews [34,51].

In total, 14 (21.5%) studies assessed how human-like chatbots are (realism) [17,18,21,34,39,41,46,50,63,66,68,70-72]. Realism of chatbots was examined in terms of verbal responses only [17,21,34,39,46,63,68,70], nonverbal responses only [66], or both verbal and nonverbal responses [18,41,50,71,72]. The included studies evaluated realism using a single question in a self-administrated questionnaire [17,18,21,39,46,50,63,70], multiple questions in a self-administrated questionnaire [41,72], and interviews [18,34,66,68,71].

Altogether, 11 (16.9%) studies assessed the speed of a chatbot's responses [18,19,28,30,34,36,38,68-70,73]. The speed of responses was examined using a single question in a self-administrated questionnaire [18,30,36,69,70,73], interviews [19,34,68], multiple questions in a self-administrated questionnaire [53], and system logs [38]. Empathy of a chatbot, which refers to its ability to show empathy to users, was examined in 10 studies [7,35,41,42,64,66,67,71,73,74]. Those studies evaluated empathy using a single question in a self-administrated questionnaire [7,35,41,42,67,71,73], interviews [66,71], and multiple questions in a self-administrated questionnaire [64].

Repetitiveness of a chatbot's responses was examined in 9 (13.8%) studies [7,20,27,53,57,66,73,75,76]. Repetitiveness of responses was evaluated using a single question in a self-administrated questionnaire [7,20,27,53,57,73] and interviews [66,75,76]. We found that 6 (9.2%) studies evaluated clarity or quality of speech using either interviews [51,62,75] or a single question in a self-administrated questionnaire [27,69,77]. Linguistic accuracy of a chatbot's responses was evaluated in 2 (3.7%) studies using a single question in a self-administrated questionnaire [31,35].

Metrics Related to Response Understanding

The included studies evaluated chatbot understanding of users' responses using the following metrics: understanding as assessed by users, word error rate, concept error rate, and attention estimator errors.

Chatbot understanding, which refers to a chatbot's ability to adequately understand the verbal and nonverbal responses of users, was assessed by 20 (30.8%) studies [7,18,20,23,27,32,33,36,39,41,42,53,57,59,63,64,68,73,78,79]. Of those studies, 2 studies assessed understanding of both verbal and nonverbal responses [18,79], 1 study assessed understanding of nonverbal responses only [32], and the remaining studies assessed understanding of verbal responses only. The understanding of responses was evaluated using multiple questions in a self-administrated questionnaire in 4 studies [42,64,78,79], interviews in 2 studies [18,68], and a single question in a self-administrated questionnaire in the remaining studies.

Word error rate, which assesses the performance of speech recognition in chatbots, was examined in 2 (3.7%) studies using conversational logs [65,69]. Concept error rate, which depends on the correct recognition of the semantic result of a user utterance, was evaluated in 1 study by checking conversational logs [65]. Attention estimation, which refers to a chatbot's ability to determine whether the user is gazing toward the screen or away from it, was examined in 1 study by checking conversational logs [69].

Metrics Related to Esthetics

The included studies evaluated the esthetics of chatbots using the following metrics: appearance of the virtual agent, background color and content, font type and size, button color, shape, icon, and background color contrast.

In total, 5 (7.7%) studies assessed the appearance of the virtual agent using a single question in a self-administrated questionnaire [69,77,80], interviews [51], and focus groups [45]. In addition, 1 (1.5%) study evaluated background color, color contrast, and content; font type and size; and button color, shape, and icon using a survey [80].

Discussion

Principal Findings

It became clear that there is currently no standard method in use to evaluate health chatbots. Most aspects are studied using self-administered questionnaires or user interviews. Common metrics are response speed, word error rate, concept error rate, dialogue efficiency, attention estimation, and task completion. Various studies assessed different aspects of chatbots, complicating direct comparison. Although some of this variation may be due to the individual characteristics of chatbot implementations and their distinct use cases, it is difficult to see why metrics such as appropriateness of responses, comprehensibility, realism, speed of response, empathy and repetitiveness are each only applicable to a small percentage of cases. Further, objective quantitative metrics (eg, those based on log reviews) were comparatively rarely used in the reported studies. We thus suggest continuing research and development toward an evaluation framework for technical metrics with recommendations for specific circumstances for their inclusion in chatbot studies.

Jadeja et al [81] introduced 4 dimensions for chatbot evaluations: the information retrieval (IR) perspective, the user experience (UX) perspective, the linguistic perspective, and the AI (human-likeness) perspective. In earlier work [14], we adapted and broadened this categorization, modifying the IR perspective to a task-oriented perspective since health chatbots are not necessarily designed only to retrieve information; additionally, we included system quality and health care quality perspectives. Excluding the health care quality perspective, which is outside the definition of technical metrics, the findings of this scoping review show that all these dimensions are indeed represented in health chatbot evaluations. Rather, the issue appears to be the inconsistency in what is measured and how, along with the skew toward self-reporting and the UX perspective. Additional work is still required to come up with standard metrics and corresponding assessment tools specifically addressing quality in health chatbots.

We found usability to be the most commonly assessed aspect of health chatbots. The system usability scale (SUS [82,83]) is a well-established usability instrument that we observed was used repeatedly, although it was not used in the majority of the studies assessing usability; in many cases, a single survey question was used instead. The SUS is nonproprietary, technology-agnostic, and designed to support comparison across products [82]. As such, global assessment of the user experience of health chatbots could be enhanced in quality and comparability by researchers standardizing on inclusion of the SUS in their evaluations. However, studies by Holmes et al [84] showed that conventional methods for assessing usability and user experience may not be as accurate when applied to health chatbots. As such, there remains research to be done toward appropriate metrics for health chatbots.

Conversational-turns per session (CPS) has been suggested as a success metric for social chatbots as exemplified by XiaoIce [85]. Although the aims for health chatbots are not identical to those of social chatbots, if CPS gains acceptance as a standard measure in the social chatbot domain, it would make it a leading candidate for a standard measure to include in health chatbot evaluations to assess their social engagement dimension. An alternative or supplementary measure related to the social dimension would be to have users score the chatbot on empathy; however, CPS has the advantage of being an objective and quantitative measure. Other objective and quantitative measures such as interaction time or time on task could be alternatives to CPS, but might be less representative of engagement than CPS if for instance the user was multitasking chatbot interaction with other tasks. Beyond social engagement, task completion (often assessed by analyzing conversation logs) is another promising global measure.

A further area for standardization would be in the quality of responses. We observed response generation to be widely measured but in very diverse ways. Emergence of standard measures for response generation and understanding would greatly advance the comparability of studies. Development of validated instruments in this area would be a useful contribution to chatbot research.

We commend the inclusion of classifier performance in health chatbot studies where this is applicable and practical to assess. It could be less meaningful to compare raw performance (eg, as area under the curve) across domains due to differences in difficulty; ideally, chatbot performance would be compared to the performance of a human expert for the task at hand. Further, we perceive the opportunity for a progression of performance measures in health chatbot studies as a given product gains maturity. Good early-stage metrics would be those that assess response quality and response understanding to establish that the product is working well. Subsequent experiments can advance the assessment of self-reported usability and metrics of social engagement. Where applicable, classifier performance can round out technical performance evaluation to establish whether trials to assess clinical outcomes are warranted.

Strengths and Limitations

Strengths

This study is the first review that summaries the technical metrics used by previous studies to evaluate health care chatbots. This helps readers explore how chatbots were evaluated in health care. Given that this review was executed and reported in line with PRISMA-ScR guidelines [1], it could be considered a high-quality review.

To retrieve as many relevant studies as possible, the most commonly used databases in the fields of health and information technology were searched. Further, we searched Google Scholar and conducted backward and forward reference list checking to retrieve gray literature and minimize the risk of publication bias.

As two reviewers independently selected the studies and extracted the data, the selection bias in this review was minimal. This review can be considered a comprehensive review given that we did not apply restrictions regarding the study design, study setting, year of publication, and country of publication.

Laranjo et al [86] reviewed the characteristics, current applications and evaluation measures of health chatbots. In contrast to our work, they did not solely concentrate on the technical metrics used for chatbot evaluations. The metrics they reviewed included task completion or word accuracy. In contrast to Laranjo et al [86], who included only 17 papers reporting on 14 different conversational agents, our work is more comprehensive as it included 65 publications. Further, we had a different research question in mind while conducting the review.

Limitations

This review focused on chatbots that are aimed at delivering health care services to patients and work on stand-alone software and web browsers; it excluded chatbots that used robotics, serious games, SMS text messaging, Wizard-of-Oz, and telephones. Thus, this review did not include many technical metrics used to evaluate chatbots for other users (eg, physicians, nurses, and caregivers), in other fields (eg, business and education), or with alternative modes of delivery (eg, SMS text messaging, Wizard-of-Oz, and telephones). The abovementioned restrictions were applied by previous reviews about chatbots as these features are not part of ordinary chatbots [87-90].

Due to practical constraints, we could not search interdisciplinary databases (eg, Web of Science and ProQuest), conduct a manual search, or contact experts. Further, the search in this review was restricted to English-language studies. Accordingly, it is likely that this review missed some studies.

Conclusion

From this review, we perceive the need for health chatbot evaluators to consider measurements across a range of aspects in any given study or study series, including usability, social experience, response quality, and, where applicable, classifier performance. The establishment of standard measures would greatly enhance comparability across studies with the SUS and CPS as leading candidates for usability and social experience, respectively. It would be ideal to develop guidelines for health chatbot evaluators indicating what should be measured and at what stages in product development. Development of validated measurement instruments in this domain is sparse and such instruments would benefit the field, especially for response quality metrics.

Authors' Contributions

AA developed the protocol and conducted the search with guidance from and under the supervision of KD and MH. Study selection and data extraction were carried out independently by MA and ZS. AA solved any disagreements between the two reviewers. AA synthesized the data. AA and KD drafted the manuscript, and it was revised critically for important intellectual content by all authors. KD and JW reviewed the related literature and interpreted the results. All authors approved the manuscript for publication and agree to be accountable for all aspects of the work.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Search string.
[DOCX File , 19 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Data extraction form.
[DOCX File , 17 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Metadata and population characteristics of each included study.
[DOCX File , 24 KB-Multimedia Appendix 3]

Multimedia Appendix 4

Characteristics of the intervention in each included study.
[DOCX File , 23 KB-Multimedia Appendix 4]

References

1. Weizenbaum J. ELIZA-a computer program for the study of natural language communication between man and machine. Commun ACM 1966;9(1):36-45. [doi: 10.1145/365153.365168]
2. Friedman R. Automated Telephone Conversations to Assess Health Behavior and Deliver Behavioral Interventions. Journal of Medical Systems 1998 Apr 01;22(2):95-102. [doi: 10.1023/A:1022695119046]
3. Whittaker R, McRobbie H, Bullen C, Rodgers A, Gu Y. Mobile phone-based interventions for smoking cessation. Cochrane Database Syst Rev 2016;4:CD006611. [doi: 10.1002/14651858.CD006611.pub4] [Medline: 27060875]
4. Neve M, Morgan PJ, Jones PR, Collins CE. Effectiveness of web-based interventions in achieving weight loss and weight loss maintenance in overweight and obese adults: a systematic review with meta-analysis. Obes Rev 2010 Apr;11(4):306-321. [doi: 10.1111/j.1467-789X.2009.00646.x] [Medline: 19754633]
5. Webb TL, Joseph J, Yardley L, Michie S. Using the internet to promote health behavior change: a systematic review and meta-analysis of the impact of theoretical basis, use of behavior change techniques, and mode of delivery on efficacy. J Med Internet Res 2010;12(1):e4 [FREE Full text] [doi: 10.2196/jmir.1376] [Medline: 20164043]
6. Kelders SM, Kok RN, Ossebaard HC, Van Gemert-Pijnen JEWC. Persuasive system design does matter: a systematic review of adherence to web-based interventions. J Med Internet Res 2012 Nov 14;14(6):e152 [FREE Full text] [doi: 10.2196/jmir.2104] [Medline: 23151820]
7. 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]
8. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. JAMA 2016 Dec 13;316(22):2402-2410. [doi: 10.1001/jama.2016.17216] [Medline: 27898976]
9. Stead WW. Clinical Implications and Challenges of Artificial Intelligence and Deep Learning. JAMA 2018 Sep 18;320(11):1107-1108. [doi: 10.1001/jama.2018.11029] [Medline: 30178025]
10. Zhou L, Gao J, Li D, Shum H. The Design and Implementation of XiaoIce, an Empathetic Social Chatbot. Computational Linguistics 2020 Mar;46(1):53-93. [doi: 10.1162/coli_a_00368]
11. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001 Sep;16(9):606-613 [FREE Full text] [Medline: 11556941]
12. Conlin PR, Colburn J, Aron D, Pries RM, Tschanz MP, Pogach L. Synopsis of the 2017 U.S. Department of Veterans Affairs/U.S. Department of Defense Clinical Practice Guideline: Management of Type 2 Diabetes Mellitus. Ann Intern Med 2017 Nov 07;167(9):655-663. [doi: 10.7326/M17-1362] [Medline: 29059687]
13. Kocaballi A, Quiroz J, Rezazadegan D, Berkovsky S, Magrabi F, Coiera E. Responses of Conversational Agents to Health and Lifestyle Prompts: Investigation of Appropriateness and Presentation Structures. J Med Internet Res 2020;22(2):e15823. [doi: 10.2196/preprints.15823]
14. Denecke K, Warren J. How to Evaluate Health Applications with Conversational User Interface? Studies in Health Technology and Informatics 2020:1-9.
15. 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. [doi: 10.7326/M18-0850] [Medline: 30178033]
16. Altman D. Practical statistics for medical research. London: Chapman & Hall/CRC; 1990.
17. Abdullah AS, Gaehde S, Bickmore T. A Tablet Based Embodied Conversational Agent to Promote Smoking Cessation among Veterans: A Feasibility Study. J Epidemiol Glob Health 2018 Dec;8(3-4):225-230. [doi: 10.2991/j.jegh.2018.08.104] [Medline: 30864768]
18. Ali M, Rasazi Z, Mamun A, Langevin R, Rawassizadeh R, Schubert L. A Virtual Conversational Agent for Teens with Autism: Experimental Results and Design Lessons. arXiv 2018:1-13.
19. Beiley M. Mental Health And Wellness Chatbot. United States: The University of Arizona; 2019.
20. Bickmore T, Caruso L, Clough-Gorr K. Acceptance and usability of a relational agent interface by urban older adults. In: CHI'05 extended abstracts on Human factors in computing systems: ACM. 2005 Presented at: CHI EA '05; April; Portland, OR, USA p. 1212-1215. [doi: 10.1145/1056808.1056879]
21. Bickmore TW, Pfeifer L, Jack BW. Taking the time to care. 2009 Presented at: Proceedings of the 27th international conference on Human factors in computing systems - CHI 09; 2009; Boston. [doi: 10.1145/1518701.1518891]
22. Bickmore TW, Mitchell SE, Jack BW, Paasche-Orlow MK, Pfeifer LM, Odonnell J. Response to a Relational Agent by Hospital Patients with Depressive Symptoms. Interact Comput 2010 Jul 01;22(4):289-298 [FREE Full text] [doi: 10.1016/j.intcom.2009.12.001] [Medline: 20628581]
23. Bickmore TW, Pfeifer LM, Byron D, Forsythe S, Henault LE, Jack BW, et al. Usability of conversational agents by patients with inadequate health literacy: evidence from two clinical trials. J Health Commun 2010;15 Suppl 2:197-210. [doi: 10.1080/10810730.2010.499991] [Medline: 20845204]
24. 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]

25. Bickmore TW, Schulman D, Sidner C. Automated interventions for multiple health behaviors using conversational agents. Patient Educ Couns 2013 Aug;92(2):142-148 [FREE Full text] [doi: 10.1016/j.pec.2013.05.011] [Medline: 23763983]
26. Bresó A, Martínez-Miranda J, Botella C, Baños RM, García-Gómez JM. Usability and acceptability assessment of an empathic virtual agent to prevent major depression. Expert Systems 2016 May 25;33(4):297-312. [doi: 10.1111/EXSY.12151] [Medline: 2016]
27. Cameron G, Cameron D, Megaw G, Bond R, Mulvenna M, O'Neill S. Assessing the Usability of a Chatbot for Mental Health Care. Internet Science 2019;Cham:121-132. [doi: 10.1007/978-3-030-17705-8_11]
28. Comendador BEV, Francisco BMB, Medenilla JS, Nacion SMT, Serac TBE. Pharmabot: A Pediatric Generic Medicine Consultant Chatbot. JOACE 2015;3(2):137-140. [doi: 10.12720/joace.3.2.137-140]
29. Crutzen R, Peters GY, Portugal SD, Fisser EM, Grolleman JJ. An artificially intelligent chat agent that answers adolescents' questions related to sex, drugs, and alcohol: an exploratory study. J Adolesc Health 2011 May;48(5):514-519. [doi: 10.1016/j.jadohealth.2010.09.002] [Medline: 21501812]
30. Demirci H. User experience over time with conversational agents case study of woebot on supporting subjective well-being. Turkey: Middle East Technical University; 2018.
31. Denecke K, Hochreutener S, Pöpel A, May R. Self-Anamnesis with a Conversational User Interface: Concept and Usability Study. Methods Inf Med 2019 Mar 15;57(05/06):243-252. [doi: 10.1055/s-0038-1675822]
32. DeVault D, Artstein R, Benn G, Dey T, Fast E, Gainer A, et al. SimSensei kiosk: a virtual human interviewer for healthcare decision support. 2014 Presented at: International Foundation for Autonomous Agents and Multiagent Systems; May 2014; Paris p. 1061-1068.
33. Dworkin MS, Lee S, Chakraborty A, Monahan C, Hightow-Weidman L, Garofalo R, et al. Acceptability, Feasibility, and Preliminary Efficacy of a Theory-Based Relational Embodied Conversational Agent Mobile Phone Intervention to Promote HIV Medication Adherence in Young HIV-Positive African American MSM. AIDS Educ Prev 2019 Feb;31(1):17-37. [doi: 10.1521/aeap.2019.31.1.17] [Medline: 30742481]
34. Elmasri D, Maeder A. A Conversational Agent for an Online Mental Health Intervention. In: Brain Informatics and Health: International Conference. 2016 Presented at: Brain Informatics and Health 2016; October 13-16; Omaha, NE, USA p. 243-251. [doi: 10.1007/978-3-319-47103-7_24]
35. Fadhil A, Schiavo G, Villafiorita A, Fondazione B. OlloBot-Towards A Text-Based Arabic Health Conversational Agentvaluation and Results. 2019 Presented at: Proceedings of Recent Advances in Natural Language Processing; 2019; Varna p. 295-303. [doi: 10.26615/978-954-452-056-4_034]
36. Griol D, Callejas Z. Mobile Conversational Agents for Context-Aware Care Applications. Cogn Comput 2015 Aug 21;8(2):336-356. [doi: 10.1007/s12559-015-9352-x]
37. Hanke S, Sandner E, Kadyrov S, Stainer-Hochgatterer A. Daily life support at home through a virtual support partner. 2016 Presented at: 2nd IET International Conference on Technologies for Active and Assisted Living (TechAAL); 2016; London p. 1-7. [doi: 10.1049/ic.2016.0058]
38. Hess G, Fricker G, Denecke K. Improving and Evaluating eMMA?s Communication Skills: A Chatbot for Managing Medication. Studies in health technology and informatics 2019;259:101-104.
39. Kadariya D, Venkataramanan R, Yip H, Kalra M, Thirunarayanan K, Sheth A. kBot: Knowledge-Enabled Personalized Chatbot for Asthma Self-Management. 2019 Presented at: IEEE International Conference on Smart Computing (SMARTCOMP); 2019; Washington p. 138-143. [doi: 10.1109/smartcomp.2019.00043]
40. Kowatsch T, Otto L, Harperink S, Cotti A, Schlieter H. A design and evaluation framework for digital health interventions. IT - Information Technology 2019;61(5-6):253-263. [doi: 10.1515/itit-2019-0019]
41. Lisetti C, Amini R, Yasavur U, Rishe N. I Can Help You Change! An Empathic Virtual Agent Delivers Behavior Change Health Interventions. ACM Trans Manage Inf Syst 2013 Dec 01;4(4):1-28. [doi: 10.1145/2544103]
42. Magnani JW, Schlusser CL, Kimani E, Rollman BL, Paasche-Orlow MK, Bickmore TW. The Atrial Fibrillation Health Literacy Information Technology System: Pilot Assessment. JMIR Cardio 2017;1(2):e7 [FREE Full text] [doi: 10.2196/cardio.8543] [Medline: 29473644]
43. Micoulaud-Franchi J, Sagaspe P, de SE, Bioulac S, Sauteraud A, Philip P. Acceptability of Embodied Conversational Agent in a Health Care Context. Intelligent Virtual Agents 2016:416-419. [doi: 10.1007/978-3-319-47665-0_45]
44. Milne M, Luerssen M, Lewis T, Leibbrandt R, Powers D. Development of a virtual agent based social tutor for children with autism spectrum disorders. The 2010 International Joint Conference on Neural Networks (IJCNN). 2010 Presented at: The 2010 International Joint Conference on Neural Networks (IJCNN); 2010 July 2010; Barcelona p. 18-23. [doi: 10.1109/ijcnn.2010.5596584]
45. Schmidlen T, Schwartz M, DiLoreto K, Kirchner HL, Sturm AC. Patient assessment of chatbots for the scalable delivery of genetic counseling. J Genet Couns 2019 Dec;28(6):1166-1177. [doi: 10.1002/jgc4.1169] [Medline: 31549758]
46. Schroeder J, Wilkes C, Rowan K, Toledo A, Paradiso A, Czerwinski M. Pocket Skills: A Conversational Mobile Web App To Support Dialectical Behavioral Therapy. Canada: ACM; 2018 Presented at: The 2018 CHI Conference on Human Factors in Computing Systems; 2018; Montreal p. 1-15. [doi: 10.1145/3173574.3173972]

47. Smith MJ, Ginger EJ, Wright M, Wright K, Boteler Humm L, Olsen D, et al. Virtual reality job interview training for individuals with psychiatric disabilities. J Nerv Ment Dis 2014 Sep;202(9):659-667 [FREE Full text] [doi: 10.1097/NMD.0000000000000187] [Medline: 25099298]
48. Smith MJ, Ginger EJ, Wright K, Wright MA, Taylor JL, Humm LB, et al. Virtual reality job interview training in adults with autism spectrum disorder. J Autism Dev Disord 2014 Oct;44(10):2450-2463 [FREE Full text] [doi: 10.1007/s10803-014-2113-y] [Medline: 24803366]
49. Smith MJ, Humm LB, Fleming MF, Jordan N, Wright MA, Ginger EJ, et al. Virtual Reality Job Interview Training for Veterans with Posttraumatic Stress Disorder. J Vocat Rehabil 2015;42(3):271-279 [FREE Full text] [doi: 10.3233/JVR-150748] [Medline: 27721645]
50. Tanaka H, Sakti S, Neubig G, Toda T, Negoro H, Iwasaka H. Automated Social Skills Trainer. 2015 Presented at: The 20th International Conference on Intelligent User Interfaces; 2015; Atlanta. [doi: 10.1145/2678025.2701368]
51. Thompson D, Callender C, Gonynor C, Cullen KW, Redondo MJ, Butler A, et al. Using Relational Agents to Promote Family Communication Around Type 1 Diabetes Self-Management in the Diabetes Family Teamwork Online Intervention: Longitudinal Pilot Study. J Med Internet Res 2019 Sep 13;21(9):e15318 [FREE Full text] [doi: 10.2196/15318] [Medline: 31538940]
52. Tielman ML, Neerincx MA, Bidarra R, Kybartas B, Brinkman W. A Therapy System for Post-Traumatic Stress Disorder Using a Virtual Agent and Virtual Storytelling to Reconstruct Traumatic Memories. J Med Syst 2017 Aug;41(8):125 [FREE Full text] [doi: 10.1007/s10916-017-0771-y] [Medline: 28699083]
53. Yasavur U, Lisetti C, Rishe N. Let's talk! speaking virtual counselor offers you a brief intervention. J Multimodal User Interfaces 2014 Sep 5;8(4):381-398. [doi: 10.1007/s12193-014-0169-9]
54. Amato F, Marrone S, Moscato V, Piantadosi G, Picariello A, Sansone C. Chatbots Meet eHealth: Automatizing Healthcare. 2017 Presented at: Workshop on Artificial Intelligence with Application in Health; 2017; Bari p. 1-10.
55. Auriacombe M, Moriceau S, Serre F, Denis C, Micoulaud-Franchi J, de Sevin E, et al. Development and validation of a virtual agent to screen tobacco and alcohol use disorders. Drug and Alcohol Dependence 2018 Dec;193(6):1-6. [doi: 10.1016/J.drugalcdep.2018.08.025] [Medline: 2018]
56. Ghosh S, Bhatia S, Bhatia A. Quro: Facilitating User Symptom Check Using a Personalised Chatbot-Oriented Dialogue System. Stud Health Technol Inform 2018;252:51-56. [Medline: 30040682]
57. Inkster B, Sarda S, Subramanian V. An Empathy-Driven, Conversational Artificial Intelligence Agent (Wysa) for Digital Mental Well-Being: Real-World Data Evaluation Mixed-Methods Study. JMIR mHealth uHealth 2018 Nov 23;6(11):e12106 [FREE Full text] [doi: 10.2196/12106] [Medline: 30470676]
58. Ni L, Lu C, Liu N, Liu J. MANDY: Towards a Smart Primary Care Chatbot Application. Knowledge and Systems Sciences 2017:38-52. [doi: 10.1007/978-981-10-6989-5_4]
59. Philip P, Bioulac S, Sauteraud A, Chaufton C, Olive J. Could a Virtual Human Be Used to Explore Excessive Daytime Sleepiness in Patients? Presence: Teleoperators and Virtual Environments 2014 Nov;23(4):369-376. [doi: 10.1162/pres_a_00197]
60. Philip P, Micoulaud-Franchi J, Sagaspe P, Sevin ED, Olive J, Bioulac S, et al. Virtual human as a new diagnostic tool, a proof of concept study in the field of major depressive disorders. Sci Rep 2017 Feb 16;7:42656 [FREE Full text] [doi: 10.1038/srep42656] [Medline: 28205601]
61. Tanaka H, Adachi H, Ukita N, Ikeda M, Kazui H, Kudo T, et al. Detecting Dementia Through Interactive Computer Avatars. IEEE J Transl Eng Health Med 2017;5:1-11. [doi: 10.1109/jtehm.2017.2752152]
62. Gardiner PM, McCue KD, Negash LM, Cheng T, White LF, Yinusa-Nyahkoon L, et al. Engaging women with an embodied conversational agent to deliver mindfulness and lifestyle recommendations: A feasibility randomized control trial. Patient Educ Couns 2017 Sep;100(9):1720-1729. [doi: 10.1016/j.pec.2017.04.015] [Medline: 28495391]
63. Razavi S, Ali M, Smith T, Schubert L, Hoque M. The LISSA virtual human and ASD teens: An overview of initial experiments. 2016 Presented at: International Conference on Intelligent Virtual Agents; 2016; Los Angeles p. 460-463. [doi: 10.1007/978-3-319-47665-0_55]
64. Liu B, Sundar SS. Should Machines Express Sympathy and Empathy? Experiments with a Health Advice Chatbot. Cyberpsychol Behav Soc Netw 2018 Oct;21(10):625-636. [doi: 10.1089/cyber.2018.0110] [Medline: 30334655]
65. Turunen M, Hakulinen J, Ståhl O, Gambäck B, Hansen P, Rodríguez Gancedo MC, et al. Multimodal and mobile conversational Health and Fitness Companions. Computer Speech & Language 2011 Apr;25(2):192-209. [doi: 10.1016/j.csl.2010.04.004]
66. 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]
67. Swartout W, Artstein R, Forbell E, Foutz S, Lane HC, Lange B, et al. Virtual Humans for Learning. AIMag 2013 Dec 15;34(4):13. [doi: 10.1609/aimag.v34i4.2487]
68. van Heerden A, Ntinga X, Vilakazi K. The potential of conversational agents to provide a rapid HIV counseling and testing services. 2017 Presented at: International Conference on the Frontiers and Advances in Data Science; 2017; Xi'an p. 80-85. [doi: 10.1109/fads.2017.8253198]

69. Wargnier P, Benveniste S, Jouvelot P, Rigaud A. Usability assessment of interaction management support in LOUISE, an ECA-based user interface for elders with cognitive impairment. TAD 2018 Nov 26;30(3):105-126. [doi: 10.3233/tad-180189]
70. Kang J, Wei L. "Give Me the Support I Want!": The Effect of Matching an Embodied Conversational Agent's Social Support to Users' Social Support Needs in Fostering Positive User-Agent Interaction. 2018 Presented at: The 6th International Conference on Human-Agent Interaction; 2018; Southampton p. 106-113. [doi: 10.1145/3284432.3284462]
71. Olafsson S, O'Leary T, Bickmore T. Coerced Change-talk with Conversational Agents Promotes Confidence in Behavior Change. 2019 Presented at: The 13th EAI International Conference on Pervasive Computing Technologies for Healthcare; 2019; Trento p. 31-40. [doi: 10.1145/3329189.3329202]
72. Tielman ML, Neerincx MA, van Meggelen M, Franken I, Brinkman W. How should a virtual agent present psychoeducation? Influence of verbal and textual presentation on adherence. Technol Health Care 2017 Dec 04;25(6):1081-1096 [FREE Full text] [doi: 10.3233/THC-170899] [Medline: 28800346]
73. 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]
74. Martínez-Miranda J, Bresó A, García-Gómez J. Look on the bright side: a model of cognitive change in virtual agents. 2014 Presented at: International Conference on Intelligent Virtual Agents; 2014; Boston p. 285-294. [doi: 10.1007/978-3-319-09767-1_37]
75. Bickmore T. Relational agents affecting change through human-computer relationships. USA: Massachusetts Institute of Technology; 2003:1-284.
76. Ly KH, Ly A, Andersson G. A fully automated conversational agent for promoting mental well-being: A pilot RCT using mixed methods. Internet Interventions 2017 Dec;10:39-46. [doi: 10.1016/j.invent.2017.10.002]
77. Martínez-Miranda J, Martínez A, Ramos R, Aguilar H, Jiménez L, Arias H, et al. Assessment of users' acceptability of a mobile-based embodied conversational agent for the prevention and detection of suicidal behaviour. J Med Syst 2019 Jun 25;43(8):246. [doi: 10.1007/s10916-019-1387-1] [Medline: 31240494]
78. Pinto MD, Greenblatt AM, Hickman RL, Rice HM, Thomas TL, Clochesy JM. Assessing the Critical Parameters of eSMART-MH: A Promising Avatar-Based Digital Therapeutic Intervention to Reduce Depressive Symptoms. Perspect Psychiatr Care 2016 Jul;52(3):157-168. [doi: 10.1111/ppc.12112] [Medline: 25800698]
79. Yokotani K, Takagi G, Wakashima K. Advantages of virtual agents over clinical psychologists during comprehensive mental health interviews using a mixed methods design. Computers in Human Behavior 2018 Aug;85(6):135-145. [doi: 10.1016/j.chb.2018.03.045] [Medline: 2018]
80. Wu Y, Samant D, Squibbs K, Chaet A, Morshedi B, Barnes L. Design of Interactive Cancer Education Technology for Latina Farmworkers. 2014 Apr Presented at: The 2014 IEEE International Symposium on Software Reliability Engineering Workshops; 2014; Washington URL: http://europepmc.org/abstract/MED/29978858 [doi: 10.1109/SIEDS.2014.6829908]
81. Jadeja M, Varia N. Perspectives for evaluating conversational AI. arXiv preprint 2017:1-6.
82. Bangor A, Kortum PT, Miller JT. An Empirical Evaluation of the System Usability Scale. International Journal of Human-Computer Interaction 2008 Jul 30;24(6):574-594. [doi: 10.1080/10447310802205776]
83. Brooke J. SUS-A quick and dirty usability scale. In: Jordan P, Thomas B, McClelland I, Weerdmeester B, editors. Usability evaluation in industry. London: CRC Press; 1996:4-7.
84. Holmes S, Moorhead A, Bond R, Zheng H, Coates V, McTear M. Can we use conventional methods to assess conversational user interfaces? Usability testing of a healthcare chatbot; 2019 Presented at: The 31st European Conference on Cognitive Ergonomics; 2019; Belfast p. 207-2014. [doi: 10.1145/3335082.3335094]
85. Shum H, He X, Li D. From Eliza to XiaoIce: challenges and opportunities with social chatbots. Frontiers Inf Technol Electronic Eng 2018 Jan 8;19(1):10-26. [doi: 10.1631/fitee.1700826]
86. 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]
87. 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. International Journal of Medical Informatics 2019 Dec;132:103978. [doi: 10.1016/j.ijmedinf.2019.103978]
88. Abd-alrazaq A, Rababeh A, Alajlani M, Bewick B, Househ M. The effectiveness and safety of using chatbots to improve mental health: A systematic review and meta-analysis. Journal of Medical Internet Research. Forthcoming 2020:2020 (forthcoming). [doi: 10.2196/16021]
89. 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. [doi: 10.1177/0706743719828977] [Medline: 30897957]
90. Provoost S, Lau HM, Ruwaard J, Riper H. Embodied Conversational Agents in Clinical Psychology: A Scoping Review. J Med Internet Res 2017 May 09;19(5):e151 [FREE Full text] [doi: 10.2196/jmir.6553] [Medline: 28487267]

Abbreviations

ACM: Association for Computing Machinery
AI: artificial intelligence
CINAHL: Cumulative Index of Nursing and Allied Health Literature
CPS: conversational-turns per session
eHealth: electronic health
EMBASE: Excerpta Medica Database
IEEE: Institute of Electrical and Electronics Engineers
IR: information retrieval
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses-Extension for Scoping Reviews
UX: user experience

Edited by C Lovis; submitted 18.02.20; peer-reviewed by DU Iqbal, E Bellei; comments to author 08.04.20; revised version received 13.04.20; accepted 15.04.20; published 05.06.20

Please cite as:
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;22(6):e18301
URL: http://www.jmir.org/2020/6/e18301/
doi: 10.2196/18301
PMID: