*Challenges of Trustable AI and Added-Value on Health*
*B. Séroussi et al. (Eds.)*


*doi:10.3233/SHTI220431*

# Usability Assessment of Conversational Agents in Healthcare: A Literature Review

Kerstin DENECKE[a,1] and Richard MAY[b]
[a] *Bern University of Applied Sciences, Bern, Switzerland*
[b] *Harz University of Applied Sciences, Wernigerode, Germany*

**Abstract.** Conversational agents (CA) are chatbot-based systems supporting the interaction with users through text, speech, or other modalities. They are used in an increasing number of medical use cases. Even though usability is considered a prerequisite for the success of mHealth apps using CA, there is still no standard procedure to study usability of health CA. In this paper, we report the results from a systematic literature review aiming at identifying study designs, tools, and metrics used to assess usability in health CA. We searched three bibliographic databases (PubMed, Scopus, IEEE Xplore) for papers reporting on CA in healthcare to extract information on the usability assessment of those CA. From 273 retrieved results, we included 66 papers for full text review. 34 of them reported on usability assessments. A broad range of tools is used (e.g. SUS, UEQ), but also individual questionnaires are exploited. The examined studies use scenario-based setups but assess also real-world usage. Exploratory setups are rarely reported. Due to the differences in the study designs and assessment tools, it is impossible to compare usability among CA. Thus, we recommend to develop a standardised procedure that can be always applied and which can be enriched by assessments needed for evaluating usability of CA-specific features.

**Keywords.** Conversational agents, usability, evaluation, chatbot, healthcare

## 1. Introduction

With the advent of artificial intelligence and its use for understanding and interpreting speech, conversational agents (CA) and their application in healthcare have gained enormous interest in recent years. CA are chatbot-based systems supporting the interaction with users through text, speech, or other modalities in a variety of medical use cases, such as triage systems [1], for medication management [2], or to recommend ICD-10 codes [3]. CA allow patients to receive immediate response (e.g. when having questions on the medication) or facilitate human-machine interaction. Equipped with empathy features, CA can create a bond of trust with its user which is impossible for other IT systems.

Usability is considered a prerequisite for the success of any mHealth app as it is one of the main indicators for the overall acceptance and success of an application. Moreover, it is a quality attribute that assesses how easy user interfaces are to use [4]. It is typically measured by having a number of test users using the system to perform a prespecified set of tasks. However, despite the extensive research on CA in healthcare, we are not

[1] Corresponding Author, Kerstin Denecke, Institute for Medical Informatics, Bern University of Applied Sciences, Quellgasse 21, 2501 Biel, Switzerland; E-mail: kerstin.denecke@bfh.ch

aware of a systematic overview of CA-related usability assessment. In this paper, we aim to assess recent research on usability assessment in the context of CA, focusing on tools and evaluation metrics used. We will discuss best practices and open research avenues. Finally, a replication package containing lists of all selected publications, including our extraction sheet will be delivered.

## 2. Methodology

To achieve our research objectives, we employed a systematic literature review according to the guidelines of Kitchenham et al. [5]. In the following subsections, we describe the individual steps of our study based on these guidelines.

**Initial Screening.** First, we conducted an initial screening to ensure that there is a relevant body-of-knowledge in the field of CA in healthcare, i.e. there is sufficient research interest. For this purpose, we applied the following search string to the literature databases Pubmed, IEEE Xplore, and Scopus: ("chatbot" OR "conversational agent") AND ("healthcare" OR "health care"). This search string resulted in a total of 720 results (103 on IEEE Xplore, 143 on Pubmed, 474 on Scopus). Thus, we confirmed considerable research interest in CA in healthcare and considered our research objective valuable enough to initiate our detailed systematic literature study.

**Search String.** Second, we performed the refined literature search on Pubmed, IEEE Xplore, and Scopus based on the general findings of our initial screening. These databases include peer-reviewed literature of diverse publishers, which reduced the threat of missing relevant papers and ensured a high publication quality. In this context, we applied the following search string: *("conversational user interface" OR "intelligent agent" OR "conversational agent" OR chatbot) AND (health OR healthcare OR medicine) AND evaluation.* In the Pubmed search string, we resisted on including search terms "(health OR healthcare OR medicine)" since Pubmed basically contains literature from the health domain.

**Selection Criteria (SC):** To identify relevant papers, we used the following criteria:

- **SC1**: The publication is written in English.
- **SC2**: The publication is a peer-reviewed conference paper or journal article.
- **SC3**: The publication has been published between 2012 and 2021.
- **SC4**: The publication is longer than five pages.
- **SC5**: The publication reports on studies dealing with CA.

We intentionally focused on the last decade to cover the most recent research (**SC3**). We argue that relevant findings on CA older than ten years have typically already become established fundamentals or practices. **SC4** was used to ensure a certain quality of the papers, assuming that a publication with a minimum number of pages provides enough details to comprehend the addressed problem. Moreover, we relied on the review of publication venues by the chosen literature databases. This is a well-established adaptation, as we structure previous findings based on different research methods [7]. Regarding **SC5**, review papers as well as papers reporting theoretical frameworks or complete conference proceedings were excluded. Further, we removed publications dealing with embodied CA or systems that send only push up notifications as they are based on other requirements than traditional text- or voice-based CA. Systems not dealing with healthcare were also removed.

**Data Extraction:** To extract relevant data, we defined the following criteria: type of CA, e.g. coaching or informational, input-output type, e.g. text or speech, evaluation

aspect, e.g. usability or efficacy, number of participants, type of participants, e.g. students or patients, metrics used, e.g. SUS or UEQ, tools used, e.g. questionnaire or interview, and evaluation execution, e.g. scenario or exploratory.

**Conduct:** We conducted the literature search on December 12, 2021. Overall, we identified 273 results (24 on IEEE Xplore, 36 on Pubmed, 213 on Scopus). In a first step, both authors manually reviewed the papers by using the collaborative review tool Rayyan QCRI which automatically removed 16 duplicates. Each reviewer examined half of the publications' titles and abstracts resulting in the exclusion of 186 papers including nine additional duplicates. Next, the full texts were downloaded for detailed analyses. However, for three papers the full texts were not accessible, i.e. we considered 84 papers in the full text review. All disagreements between the authors were resolved during discussions until a consensus on a decision was achieved.

## 3. Results

***Data*** assessment***:*** In the assessment phase, 18 papers were removed because they did not fulfil the inclusion criteria (embodied CA, no evaluation results described or CA in domains other than health). 66 papers were finally assessed, and data extracted. The complete reference list and the data extraction sheet is available online[2]. The CA included in our analysis addressed multiple application areas: disease management (asthma, diabetes, chronic pain), intelligent interviewer (family history, PROM), retrieval (for physicians ICD-10 encoding, for patients' information retrieval), mental health (mainly delivering cognitive behavioural therapy for different mental disorders or patient education), medication management.

***CA and its conversation-related characteristics:*** Based on our previous work [6], we assume four different types of CA: informational, coaching, questioning, and monitoring. In the analysed papers, most CA focused on coaching tasks (62.1%). 19.7% of the papers described questioning CA. We also identified that 18.2% of the CA only provide information. 4.6% of the studies presented monitoring CA. Overall, we found out that three CA (4.6%) perform a higher number of different tasks than the other CA, making it impossible to assign them to a single CA type. However, we could not find any relationships between these CA. Regarding the input/output type, all CA are at least based on text. Even 78.8% of the CA are solely based on text. Furthermore, 18.2% of the studies presented CA based on a combination of text and voice. In one paper, a voice user interface without text was described. Another CA combined specific visual elements with text.

***Evaluation aspects.*** About half of the studies considered usability in their CA evaluation (51.5%). A similar number of papers also evaluated user experience, including user satisfaction or human-machine interaction (47.0%). We also identified that 39.4% of the CA were evaluated on technical aspects, including effectiveness, efficiency, performance, or reliability. 16.7% of the papers evaluated the (subjective) effectiveness of the respective CA-related intervention actions on the user. In addition, 9.1% of the studies indicated the technological acceptance of the CA. Overall, 21.2% of the papers reported only one evaluation criterion, such as user experience. However, we assume that these are usually generic terms for several partial evaluation aspects.

[2] https://github.com/Rim007/ReplicationPackage_UsabilityCA

***Characteristics of the usability tests:*** Since our research focuses on the usability evaluation of CA, in the following we will only discuss those papers that actually used usability as an evaluation criterion (n=34). We found out that 11.8% of the usability studies were conducted with 1 to 10 participants. Almost half of all evaluations were performed with 11 to 50 participants (47.0%). Furthermore, 17.7% of the papers presented evaluations with 51 to 100 and 20.6% with more than 100 participants. One paper did not report a number. We could distinguish 5 different groups of participants: doctors / therapists / experts were involved in testing (11.8%), students / staff members (20.6%), random users (20.6%), patients (41.2%) and children/teenagers (5.9%). One paper did not report about the participants (2.9%). One paper involved patients and physicians. 91.1% of the studies exploited questionnaires to assess the usability; 14.7% of the papers reported on interviews and 11.8% analysed the protocols of the conversation to learn about user behaviour and usability. Some of the studies exploited existing usability questionnaires; among them are User experience questionnaire [7-9], (n=3), System Usability Scale [10-12] (n=4), Net promoter score (n=1), Subjective Assessment of System Speech Interfaces [13] (n=1), FEDS [14] ( n=1), User engagement scale [11] (n=4), TRINDI [2] (n=1), PARADISE framework [2] (n=1), ISO 9214 [15,16] (n=2), UTAUT [17] (n=1). When not relying upon existing questionnaires, the researcher developed own surveys, often comprising only few questions (e.g. “is it easy to use?”). The main usability test setup was a real-world application, i.e. the system was used in daily practice by participants (38.2%), or a scenario-based usability test (44.1%). Users were asked to explore the application on their own in 2 studies (5.9%). 4 studies (11.8%) did not provided any information on the setup.

## 4. Discussion and conclusion

Based on the results, we found that most of the examined CA are based on text input/output aiming at coaching their users in a specific domain. On average, around 55 participants were involved in a scenario-based usability evaluation, which indicates quantitative assessment. When evaluating CA, user experience is usually examined besides usability. This shows the strong link between both aspects. However, usability is usually associated with more technical aspects than user experience, which addresses more subjective aspects, such as user satisfaction. Therefore, we assume that there is an insufficient understanding of evaluation criteria, such as the differentiation between usability and user experience. This issue is also supported by the fact that generic terms, such as interaction, were sometimes used as the only, subjective evaluation criteria. This indicates that some studies are less concerned with the evaluation of certain aspects than with the evaluation itself. However, this considerably impairs the comprehensibility of the evaluation method and its results. Only one paper [13] used an assessment tool specifically designed for assessing speech interfaces. The other tools have originally been developed for IT systems in general, not specifically for CA. The comparison of Holmes et al. showed that conventional tools like SUS are not as accurate when applied to CA [18]. We strongly recommend harmonising usability evaluation strategies to create a uniform understanding and basis for usability assessment and thus enable a comparison of evaluation results. For example, an agreed scoring system specifically developed for CA in healthcare to be used in all usability testings would allow comparison. In healthcare, we have to deal with diverse user groups covering multiple social dimensions and diverse levels of cognitive capabilities in a variety of use cases. These specific

requirements cannot be completely addressed by the established assessment tools. Further, the results showed that diversity still remains unconsidered in usability testing. Since communication is language-based with CA, assessing usability in correlation with user's language and communication skills is relevant. To address these limitations, we plan to develop a suitable evaluation tool to provide a uniform understanding and an essential basis for the comprehensible usability evaluation of healthcare-related CA. However, to create this tool, we have to answer the following research questions in future research: Would an exploratory design of a usability test be more appropriate? Would interviews with participants reveal additional challenges? What can we learn from the conversation flow on usability?

## References


[1] Omoregbe NAI, et al.. Text Messaging-Based Medical Diagnosis Using Natural Language Processing and Fuzzy Logic. Dogra A, editor. J Health Eng. 2020 Sep 29;2020:1–14.

[2] Hess GI, Fricker G, Denecke K. Improving and Evaluating eMMA's Communication Skills: A Chatbot for Managing Medication. Stud Health Technol Inform. 2019;259:101–4.

[3] Siangchin N, Samanchuen T. Chatbot Implementation for ICD-10 Recommendation System. In: 2019 Intern Conf on Engineering, Science, and Industrial Applications (ICESI). Tokyo, Japan; 2019. p. 1–6.

[4] Wilson C, editor. User experience re-mastered: your guide to getting the right design. Burlington, MA: Morgan Kaufmann Publishers; 2010. 382 p.

[5] Kitchenham BA, Budgen D, Brereton P. Evidence-based software engineering and systematic reviews. Boca Raton: CRC Press; 2016. 399 p.

[6] May R, Denecke K. Security, privacy, and healthcare-related conversational agents: a scoping review. Inform Health Soc Care. 2021 Oct 7;1–17.

[7] Denecke K, Vaaheesan S, Arulnathan A. A Mental Health Chatbot for Regulating Emotions (SERMO) - Concept and Usability Test. IEEE Trans Emerg Top Comput. 2020;1–1.

[8] Chatzimina M, Koumakis L, Marias K, Tsiknakis M. Employing Conversational Agents in Palliative Care: A Feasibility Study and Preliminary Assessment. In: 2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE). Athens, Greece: IEEE; 2019. p. 489–96.

[9] Chatzimina M, et al. Designing a conversational agent for patients with hematologic malignancies: Usability and Usefulness Study. In: 2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI). Athens, Greece: IEEE; 2021. p. 1–4.

[10] Kadariya D, Venkataramanan R, Yip HY, Kalra M, Thirunarayanan K, Sheth A. kBot: Knowledge-Enabled Personalized Chatbot for Asthma Self-Management. In: 2019 IEEE International Conference on Smart Computing (SMARTCOMP). Washington, DC, USA: IEEE; 2019. p. 138–43.

[11] Oh J, Jang S, Kim H, Kim J-J. Efficacy of mobile app-based interactive cognitive behavioral therapy using a chatbot for panic disorder. Int J Med Inf. 2020 Aug;140:104171.

[12] 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.

[13] Yasavur U, Lisetti C, Rishe N. Let's talk! speaking virtual counselor offers you a brief intervention. J Multimodal User Interfaces. 2014 Oct;8(4):381–98.

[14] Meier P, Beinke JH, Fitte C. FeelFit - Design and Evaluation of a Conversational Agent to Enhance Health Awareness. Int Conf Inf Syst 2019 ICIS 2019. 2019;

[15] Ghaleb M, Almurtadha Y, Algarni F, Abdullah M, Felemban E, M. Alsharafi A, et al. Mining the Chatbot Brain to Improve COVID-19 Bot Response Accuracy. Comput Mater Contin. 2022;70(2):2619–38.

[16] Sia DE, Yu MJ, Daliva JL, Montenegro J, Ong E. Investigating the Acceptability and Perceived Effectiveness of a Chatbot in Helping Students Assess their Well-being. In: Asian CHI Symposium 2021. Yokohama Japan: ACM; 2021. p. 34–40.

[17] Valtolina S, Hu L. Charlie: A chatbot to improve the elderly quality of life and to make them more active to fight their sense of loneliness. In: CHItaly 2021: 14th Biannual Conference of the Italian SIGCHI Chapter. Bolzano Italy: ACM; 2021. p. 1–5.

[18] 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: Proceedings of the 31st European Conference on Cognitive Ergonomics. BELFAST United Kingdom: ACM; 2019. p. 207–14.