Digital Personalized Health and Medicine
L.B. Pape-Haugaard et al. (Eds.)

How to Evaluate Health Applications with Conversational User Interface?

Kerstin DENECKE[a, 1] and Jim WARREN[b]
[a] *Bern University of Applied Sciences, Bern, Switzerland*
[b] *University of Auckland, Auckland, New Zealand*

Abstract. Application of conversational user interfaces (CUI) or chatbots to healthcare is gaining interest fueled by the rising power of artificial intelligence, increasing popularity of mobile health applications and the desire for engagement and usability. While their use is mainly justified by increasing adherence to mobile health applications and facilitating interactions with the system, the question arises: How can such systems be evaluated in a reliable manner? This paper introduces an evaluation framework for health systems whose core interaction principle is a CUI. We derive quality dimensions and attributes by collecting relevant evaluation aspects from applications that have been developed in previous work and from literature on health chatbots. The collected aspects are aggregated into six thematic categories for chatbot quality, including user experience, linguistic, task-oriented and artificial intelligence perspectives, but also healthcare quality and system quality perspectives. The framework is intended to support developers and researchers in the domain of chatbots in healthcare in selecting relevant quality attributes to be assessed before their systems are distributed to patients.

Keywords. Evaluation, Health application, Chatbot, Conversational user interface, Chatbot evaluation, Natural language understanding

1. Introduction

Mobile health applications are increasingly used by patients to collect health data, and in this way to continuously monitor personal health and to get support from a virtual personal health coach throughout the day. To realize virtual health coaches or to provide mobile health interventions, conversational user interfaces (CUI) have been gaining in interest with mobile health application developers in recent years [1]. A CUI-based system is a computer program that interacts with users using natural language (written or spoken). The aim of such a system is to simulate a human conversation. To reduce system complexity, the user input is often restricted to selecting specific predefined items (e.g. choosing options as replies). A minority of CUI-based systems allow unconstrained natural language input. Some systems use embodied avatars, while others reduce the conversation to an exchange of text messages. Among healthcare chatbots, we can recognize different application areas or scopes of use. Therapeutic or counselling chatbots provide some specific therapy such as cognitive behavior therapy (CBT) [2]. Disease or medication management chatbots support the user in managing medications, provide knowledge on medication or a disease, remind on intake, explain interactions or

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

contraindications etc. (e.g. eMMA [3]). Educational applications can be stand-alone or integrated in an application with additional scope of use (e.g. psychoeducation is integrated in a CBT chatbot). Other chatbots are used for screening or collecting the medical history (e.g. Ana [4]). Several applications exist that collect symptoms from a user to make a triage. Finally, chatbots could help in retrieving information such as enabling physicians in getting the relevant information from the electronic health record of a specific patient. Depending on the scope of use of a chatbot, different quality attributes gain in importance when judging the quality. While thousands of health chatbots can be downloaded from app stores, evaluation results are reported only rarely [1]. If at all, only selected aspects are evaluated with results often not comparable due to different methodologies, evaluation aspects and metrics used [1]. In this paper, we address the question: What are the relevant aspects for evaluating healthcare chatbots? The main objective of this work is to provide guidance on relevant dimensions for evaluating health chatbots.

2. Methodology

We are focusing on mobile health applications for patients with CUI that allow unrestricted natural language user input. In previous work, we developed four such applications: eMMA – a medication management assistant [3], CLAIRE – a mobile application with virtual reality and voice user interface to educate patients [5], Ana – a system that collects the music biography as a starting point for music therapy [4] and SERMO – a mental health chatbot that helps in regulating emotions. In previous work, these chatbots have been evaluated using different methods and from various perspectives. However, their clinical effectiveness has not yet been studied. From those works, we collect evaluation aspects. Additionally, we reviewed articles on healthcare chatbots retrieved by a PubMed search with keywords "chatbot, conversational user interface, chatbot, CUI". According to a snowball approach, we added additional works starting from the retrieved results. To develop the framework, we grouped the attributes reported on healthcare chatbots and derived evaluation dimensions.

3. Results

Six perspectives emerge as important for our framework, including user experience, linguistic, task-oriented and artificial intelligence perspectives, as well as healthcare quality and system quality perspectives (Fig. 1).

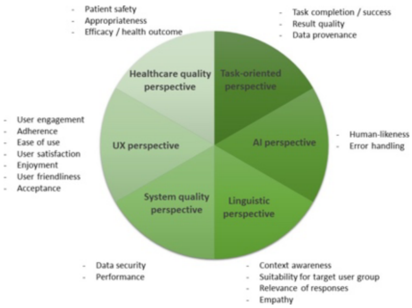

Figure 1 Evaluation framework. Quality attributes fall into six thematic categories

3.1. Task-oriented perspective

The task-oriented perspective of a chatbot evaluation assesses the capabilities of a chatbot to perform a specific task such as retrieving information, collecting specific information from a patient or predicting a diagnosis etc. It considers the degree of *task completion* or *task success* and the *result quality*. For example, a chatbot that is supposed to make a triage based on the symptoms described by a user should be assessed with respect to 1) the completeness of collected data relevant for triage and 2) the quality of the classification based on the symptoms. As such, the underlying algorithm should minimize triage classification errors and an appropriate statistic should be selected and measured (e.g. as accuracy, if the triage classes are relatively equal in frequency and in harm associated with confusion of one class for another). As a further requirement to avoid patient harm, the underlying knowledge base of a healthcare chatbot has to be evidence-based and relevant. This issue is included in the framework by the aspect *data provenance* within the task-oriented perspective. An underlying knowledge base also has to be complete with respect to the task that a chatbot is supposed to do.

3.2. Artificial intelligence (AI) perspective

The AI perspective studies to what extent the chatbot is capable of acting like a human being, e.g. in terms of problem solving or influencing a user as well as the dialogue efficiency (Note obviously, this isn't the full scope of AI, but we use 'AI perspective' as a convenient label for this focus on achieving human-like/anthropomor- phic behavior). In principle, human-likeness can be studied with the Turing test. Since the Turing test cannot measure emotional engagement with users, the metric conversation turns per session has been introduced for success of social chatbots [6]. However, in healthcare applications, it is important that users are not deceived into over- reliance on a chatbot and led to failure in recognizing its limitations due to its human- like dialog. To address this issue, it could be assessed whether the chatbot implements mechanisms to determine its own intellectual limits, so that it can forward a patient to a human healthcare provider to avoid patient harm. Error handling is another aspect of relevance for healthcare chatbots within the AI perspective. It concerns the capability of a bot to react on unexpected user input or even missing data points.

3.3. System quality perspective

System quality in healthcare chatbots should consider data security and performance (e.g. answering time). Unlike patient-doctor encounters, where patient privacy and confidentiality are protected, healthcare chatbots often do not yet consider these aspects. Some of these systems even run on social media platforms such as Facebook messenger where the use of collected data is unknown to the user or captured in data policies that are long and difficult to understand and to assess. This means the data could be sold, traded or marketed by the distributor of a chatbot. To address this issue, we included data security into the framework. In existing evaluations, aspects related to system quality have not yet been reported. They include among other things that privacy policy must be provided that is application-specific and easily accessible, ideally within the application; and the treatment of confidential data of the chatbot user must be described in detail. Some of the relevant aspects can be assessed using a checklist such as to which third parties data are transmitted, and where data are stored. Further, health chatbots must comply with existing regulations such as the General Data Protection Regulation and Medical Device Regulation.

3.4. Linguistic perspective

Evaluation of a chatbot from the linguistic perspective concerns the effectiveness of a conversation. Relevant aspects include response relevance [7], context-awareness or overall dialogue quality given divergent user input. The latter addresses the fact that the language of bot responses should be suitable for the target user group, i.e. a chatbot designed for interacting with children must use different language than one for adults. Patient-doctor encounters are ideally characterized by empathy, particularly for counseling applications. Thus, a health chatbot should express *empathy* in its dialog, and this is an aspect that should be evaluated for example by adopting the interpersonal communication competence scale [8].

3.5. UX Perspective

The UX perspective evaluates an application from a human factors or usability point of view, i.e. the feasibility of a chatbot solving specific tasks. It is mainly reflecting usability issues, for example adherence, user friendliness, ease of use, appropriateness [4], user engagement [2], user satisfaction, enjoyment. The indicator *acceptance* is a new aspect that has been added to our framework and could not be found in existing healthcare chatbot evaluations. For example, acceptance could be assessed in terms of a classic technology acceptance model (TAM) [9] which combines a user's perception of ease of use along with their perception of usefulness of the technology.

3.6. Healthcare quality perspective

The healthcare quality perspective addresses patient safety, appropriateness, and efficacy or health outcome. *Patient safety* concerns evaluations studying whether the use of the chatbot might create patient harm or risks for patients. For example, a medication assistant chatbot should provide the correct dosage of the medication to be taken. *Appropriateness* of using a healthcare chatbot assesses whether it is appropriate to deliver a certain healthcare service by means of a chatbot, e.g. whether it is appropriate to deliver cognitive behavior therapy using a chatbot for a particular patient. Finally, *health outcome or efficacy* has to be assessed aligned with practice of Evidence-Based Medicine (EBM). Then the ideal is 'Level 1' evidence as produced by randomized controlled trials (RCTs). To conduct an RCT requires measurement of a relevant validated health outcome as the dependent variable, and random assignment of subjects from the target population to the health chatbot or an appropriate 'control'. How to quantify the health outcome of a mobile application depends on the medical condition and treatment it is supposed to support (e.g. quantify efficacy of a diabetes management app by comparing the HbA1c value, or scoring systems such as Patient health questionnaire PHQ-9).

4. Discussion

This paper introduced an evaluation framework for health systems whose core interaction principle is a CUI. Jadeja et al. [10] distinguished four perspectives for evaluating general domain chatbots: Information retrieval (IR) perspective, UX perspective, linguistic perspective, and AI perspective. We adapted this categorization by broadening the scope of the IR perspective to a task-oriented perspective. Furthermore, we included two dimensions that have not yet been reported explicitly for chatbot evaluation: the system quality perspective and the healthcare quality perspective. Low quality healthcare chatbots could readily harm their users in myriad ways - such as divulging confidential data, delaying available treatment, or recommending ineffective or directly contraindicated treatment - which must be avoided. Our framework suggests attributes to be assessed for a health chatbot. We still have to assess whether relevant evaluation aspects are included related to the Medical Devices Act that have to be considered for chatbots that makes recommendations for drug administration. What is still missing are experiences on the judgement, i.e. when can we consider a health chatbot to be good or appropriate to deliver health support for real patients? Depending on the scope of use of a chatbot, some of the quality attributes suggested in the framework might become irrelevant while others gain in importance. As a next step, we start a scoping review to fortify our framework. Afterwards, a Delphi study will be conducted to collect input on how to weight the different criteria. The ultimate goal of these efforts is to provide means that help in evaluating healthcare chatbots and to ensure that only validated, evidence-based, evaluated applications will be adopted by app stores or distributed by trustworthy distributors of health applications. Experiences with implementing the framework for chatbot evaluation has to be gained. For enabling comparison among healthcare chatbots with respect to quality, agreement upon standardized metrics for the single dimensions would be helpful.

References

[1] L. Laranjo, A.|G Dunn, H.L. Tong, et al., Conversational agents in healthcare: a systematic review. *Journal of the American Medical Informatics Association* 25(9) (2019), 1248-1258.

[2] B. Inkster, S. Sarda, V. Subramanian, An empathy-driven, conversational artificial intelligence agent (wysa) for digital mental well-being: Real-world data evaluation mixed-methods study. *JMIR Mhealth Uhealth* 6(11), e12106 (Nov 2018).

[3] M. Tschanz, T.L. Dorner, J. Holm, K. Denecke, Using eMMA to manage medication. *IEEE Computer* 51(8) (2018), 18-25.

[4] K. Denecke, S. Lutz Hochreutener, A. Pöpel, R. May, Self-anamnesis with a conversational user interface: Concept and usability study. *Methods Inf Med* 57(05/06) (2018), 243-252

[5] K. Denecke, M. Tschanz, T.L. Dorner, R. May, Intelligent Conversational Agents in Healthcare: Hype or Hope? *Stud Health Technol Inform.* 259 (2019), 77-84.

[6] H. Shum, X. He, D. Li, From Eliza to Xiaoice: Challenges and opportunities with social chatbots. *CoRR abs/1801.01957* (2018), http://arxiv.org/abs/1801.01957

[7] E. Ruane, T. Faure, R. Smith, et al., Botest: A framework to test the quality of conversational agents using divergent input examples. In: *Proc. of the 23rd International Conference on Intelligent User Interfaces Companion.* IUI '18 Companion, ACM, New York, NY, USA (2018), 64:1-64:2.

[8] R.B. Rubin, M.M. Martin, Development of a measure of interpersonal communication competence. *Communication Research Reports*, 11(1), (1994), 33-44.

[9] F.D. Davis, Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Q* 13(3), (1989), 319-340.

[10] M. Jadeja, N. Varia, Perspectives for evaluating conversational AI. *arXiv preprint arXiv:1709.04734*. 2017 Sep 14.