*pHealth 2022*
*B. Blobel et al. (Eds.)*

*doi:10.3233/SHTI220992*

# What Can We Learn from Quality Requirements in ISO/TS 82304-2 for Evaluating Conversational Agents in Healthcare?

Kerstin DENECKE[a,1,2], Elizabeth M. BORYCKI[b], Andre W. KUSHNIRUK[b]
[a] *Bern University of Applied Sciences, Bern, Switzerland*
[b] *School of Health Information Science, University of Victoria, Victoria, Canada*

**Abstract.** Evaluating conversational agents (CA) that are supposed to be applied in healthcare and ensuring their quality is essential to avoid patient harm. However, most researchers only study usability and use the CA in clinical trials before conducting such careful evaluation. In previous work, consensus on metrics for evaluating healthcare CA have been found. However, the metrics are still too generic to form an evaluation framework. In this work, we try to link the ISO technical specification ISO/TS 82304-2 Quality Requirements for Health and Wellness Apps to the set of metrics to come a step closer towards an evaluation framework. We identify three links between ISO requirements and the set of metrics, namely accessibility, usability, and security. Although the technical specification rather lists aspects to be considered during development instead of concrete metrics for studying the quality, we can link to some aspects that are also of interest for health CA evaluation. For example, measuring the readability for ensuring accessibility or implementing the Web Content Accessibility Guidelines are two aspects of relevance for health CA.

**Keywords.** Conversational agent, chatbot, evaluation, ISO, quality

## 1. Introduction

Conversational agents (CA) become more popular outside medicine for dealing with customer requests, but also for delivering digital health interventions. They are for example applied in mental health for treating patients with posttraumatic stress disorder [1] or for patient education [2]. CA in healthcare differ from customer service or general domain CA. Similar to the physician-doctor conversation, the content of a conversation with a health CA has to be tailored based on the application area, use case, user's context and has to address privacy [3]. CA often process personal identifiable information that are in healthcare settings protected. They are supposed to be used in a care setting; thus, patient harm has to be avoided. To become accepted as treatments, CA are studied in

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

[2] This research was partially funded by Swiss National Science Foundation, IZSEZO_208344

clinical trials regarding efficacy, safety, or cost effectiveness. However, to avoid patient harm, frustration or negative outcomes due to low quality implementations, it is essential that prior to massive investment in clinical trials, a health CA is evaluated to demonstrate that it has an acceptable quality which includes freedom from any of a myriad of possible deficiencies.

Several tools have been suggested for evaluating mobile health apps (e.g. Mobile App Rating Scale [4], Health-ITUES). Hensher et al. developed a mobile app evaluation framework comprising among other things interoperability, technical features and support or developer credibility [5]. However, health CA require additional criteria for evaluation given their focus on communication-based interaction. Evaluation approaches for general domain CA focused on assessing effectiveness, efficiency and usability [6]. To overcome the limitation of a missing evaluation framework for health CA, we identified in previous work metrics for evaluating healthcare CA [7]. To advance this set of metrics to a health CA evaluation framework, the pool of metrics has to be linked to the potential data sources and collection methods, to support formulation of the methodology for evaluation of any specific health CA.

The objective of this work is to come one step closer to an evaluation framework for evaluating conversational agents in healthcare. To achieve this objective, we will study the 2021 released technical specification *ISO/TS 82304-2 Quality requirements for health and wellness apps* towards its applicability to healthcare CA. The concrete focus is on CA that are rule-based, based on written input and output, have a simple personality without any kind of embodiment. They are running on a mobile device, are implemented as stand-alone software and the interaction time is rather short.

## 2. Material and Methods

We go through the quality requirements that are specified in the ISO technical specification ISO/TS 82304-2 and check, whether there are overlaps with the set of global metrics that has been identified in previous work (see Table 1). We are focusing on global metrics, since the ISO specification is considering quality of apps for health and well-being. It is not addressing specific technological implementation features, but the apps as a whole. Therefore, we are convinced, that we will not find any overlaps with metrics for response generation or understanding which are elements of the set of CA evaluation metrics [7]. For all overlaps found, we will discuss the applicability to apps with conversational user interface. In the following, we briefly describe the global evaluation metrics and the ISO/TS 82304-2.

### *2.1. Global Health CA Evaluation Metrics*

A panel of experts working in the field of healthcare CA found consensus regarding metrics deemed relevant for health CA evaluation [7]. This work resulted in 24 metrics comprising 13 global metrics, 8 metrics related to response generation, 3 metrics related to response understanding and 3 metrics related to aesthetics. Definitions for the global metrics are shown in Table 1. We can see that some of the metrics are rather technical (e.g. speed, flexibility in dialogue handling) while others rather focus on the interaction with the user (e.g. ease of use, accessibility).

Table 1: Global metrics with definitions (from [7])

| Metric | Definition |
|---|---|
| Accessibility | Whether all users are able to access an equivalent user experience of the CA |
| Ease of use | Extent to which a person believes that using a particular CA would be effortless |
| Engagement | Whether a user finds value in using a CA and therefore continues using it |
| Classifier performance | How well the algorithm performs in classifying data |
| Speed | How quickly a session / task can be completed using a CA |
| Technical issues | Number of errors or glitches that occur while using a CA |
| Task completion rate | Proportion of tasks successfully completed by the CA |
| Dialogue efficiency | Number of dialogue steps used to complete a task |
| Flexibility in dialogue handling | CA's ability to maintain a conversation and deal with users' generic questions or responses that are more, less, or different than expected |
| Content accuracy | Proportion of responses that are consistent with clinical evidence, also includes correctness of triage and escalation strategies. |
| Context awareness | CA's ability to utilize contextual knowledge to appropriately respond to users |
| Error tolerance | CA's ability to detect and understand misspelled words in users' replies |
| Security | How protected the system is against hack attacks |

### *2.2. Technical specification ISO/TS 82304-2*

The technical specification *ISO/TS 82304-2 Health software — Part 2: Health and wellness apps—Quality and reliability* was first published in August 2020 (https://www.iso.org/standard/78182.html). It builds on guidelines and requirements for apps by many local and national health organizations around the world. Its purpose is to ensure that health and wellness apps are safe, reliable and effective. The guidance provides an internationally-agreed set of specifications to assess the apps. CEN ISO/TS 82304-2 was developed under CEN lead by ISO/TC 215 'Health informatics', in collaboration with IEC/TC 62 'Electrical equipment in medical practice', and adopted by CEN/TC 251 'Health informatics', whose secretariat is held by NEN, the Dutch national standardization committee. The technical specification is intended for use by app manufacturers as well as app assessment organizations in order to communicate the quality and reliability of a health app. It groups quality aspects along 5 categories: product information, healthy and safe, easy to use, secure data and robust build. These five categories bundle a set of subcategories. For example, *healthy and safe* is split up into 5 subcategories: "health requirements", "health risks", "ethics", "health benefit" and "societal benefit". *Easy to use* comprises "accessibility" and "usability", while *secure data* relates to "privacy" and "security". The guideline covers the entire life cycle of an app. We will extract the aspects that are of interest in the phase our evaluation framework is expected to be used, which is before conducting a clinical trial.

## 3. Linking ISO/TS 82304-2 to CA Evaluation Metrics

Our metrics "dialogue efficiency", "task completion rate", "flexibility in dialogue handling", "context awareness" and "classifier performance" are very specific metrics for CA. In contrast, "ease of use", "engagement", "technical issues", "error tolerance", "speed", "security", "accessibility" and "content accuracy" might be of relevance also for health apps without conversational user interface. Out of the latter set, we identified 3 quality requirements in the ISO specification that match with metrics in our set of global metrics: "easy to use" with its subcategories "usability" and "accessibility" as well as "security". These quality requirements are represented by our metrics "ease of use", "security" and "accessibility". In the following, we will discuss what the ISO guideline suggests and what we can conclude from it for health CA.

### 3.1. Accessibility

Related to accessibility, the specification suggests to develop health apps that are age appropriate and compliant with the Web Content Accessibility Guideline 2.1 (WCAG, https://www.w3.org/TR/WCAG21/). More specifically, compliance measures for perception of user interfaces and navigation, for operation and navigation, and for understanding of user interfaces and navigation should be implemented and a mobile health app should be AA or AAA compliant (medium or highest level of conformance). For example, contrast should be considered and zoom functions be available. Considering the WCAG within a CA is clearly of relevance. All elements of a dialogue with a CA must be available to a user according to their available senses (i.e. to address hearing and vision deficits) [8]. This makes it necessary to provide different input and output modalities. The ISO specification also asks for a readability assessment. This would be of particular interest for a health CA since the interaction is basically realized as conversation (written or spoken). The sentences should not be too long and have to be easy understandable. A number of scales and innovative machine learning approaches could also be used for assessing readability level of the CA in order to match the reading level to the level of expected users [9]. The WCAG provides additional aspects of relevance, for example the success criteria 1.3.2 "meaningful sequence" is particularly interesting. In case of a CA this concerns the ordering of a conversation. Beyond, its history should be accessible to the user [8]. There are a number of health literacy and e-health literacy surveys and questionnaires that can be used to assess the literacy level of the CA (and their users) and adjust the wording based on the results of applying those instruments [10,11].

### 3.2. Usability

The eight usability requirements listed in the ISO specification provide guidelines how to ensure usability of health apps. Development should start with a requirement analysis involving future users and should continue involving users throughout the entire development. A user-centered evaluation should be conducted to refine the app design. Measures should be in place to avoid use error. Furthermore, information should be provided to the user on the health app (e.g. functionality, intended use), together with instructions for use. Delivering resources for helping users who experience problems with the app are recommended. Finally, usability should be systematically gathered throughout the entire lifetime of the app. Providing information on the purpose or functionality of a health CA and instructions to use are something, a CA could do at the very beginning of the conversation when introducing itself, its purpose and what the user can expect from it. From our experiences, this could be very helpful for users since otherwise, they don't know what they can ask or write. Providing help in using the agent would require that the user can ask for help at any stage of the conversation and the CA goes back to the initial context after providing help. If such help function is realized this would be one of the benefits of a CA (compared to a normal user interface and even human-human conversation): the CA has no time pressure, can reply to all questions, but should be able at some point to return to the original topic. These aspects clearly have to be considered already in the design phase of a CA and can be checked as part of the usability testing. No concrete metrics or measurement tools are suggested in the ISO specification for studying usability. Related to usability, a study found out that a broad range of tools is used for studying usability in health CA (e.g. System Usability Scale, User Experience Questionnaire), but also individual questionnaires are exploited [12]. Due to the differences in the study designs and assessment tools that have been conducted so far, it is impossible to compare usability among health CA. It was recommended to develop a standardized procedure that can be always applied and which can be enriched by assessments needed for evaluating usability of specific features of a particular health CA. This would make usability test results better understandable and comparable. Nielsen's heuristic evaluation could be applied to the CA interface [13] as well as the cognitive walkthrough [14]. There are now usability testing standards that could be applied to guide analyses of user interactions with CA [15].

### *3.3. Security*

The security requirements of the ISO specification consist of 11 components. They ask for proof of implementation of ISO/IEC 27001 or another recognized standard related to information security management. Information security risks and potential consequences should be assessed. Beyond, a secure by design process is encouraged (e.g. ensuring correct usage of biometric sensors, secure data integration). Measures should be in place to ensure that all third-party software libraries are reliable and maintained. Unauthorized access and modification to the source code should be prevented. An information security policy has to be available; security of the app has to be regularly tested; a process of dealing with security vulnerabilities should be in place, as well as data encryption, user authentication, authorization and session management; finally, standard operating procedures have to be in place for processing personal identifiable information according to the privacy statement.

In principle, these 11 items can be used as checklist when developing health CA. It was already found that data security and privacy are under-researched and remain most often unconsidered in current health CA [16]. Checking whether these 11 items from the ISO specification are fulfilled by a concrete health CA would clearly provide a guidance and ensure information security within a health CA. Consideration must be given to privacy and security so that data would be accessible in emergency situations by health professionals and family members (e.g. during a health crisis) to be able to address the situation appropriately.

## 4. Conclusions

In this work, we linked requirements specified in the ISO/TS 82304-2 to metrics for evaluating health CA. Three links could be identified for the metrics ease of use, accessibility and security. The limited overlap is not surprising given the technical peculiarities of health CA that are reflected in our set of metrics (the interaction aspect). Beyond, the ISO specification targets already software products which we did not had in mind. Our focus was on health CA to be evaluated before conducting a clinical trial. While a few concrete metrics (e.g. readability score) could be derived from the ISO guideline, it rather provides recommendations that might be of relevance during the development of a health app. Some of them are clearly also useful to be considered in health CA development for developing high quality apps (e.g. considering WCAG or user involvement in the development). We will integrate the 11 items on security as a checklist to the health CA evaluation framework. As a next step, we link the other metrics to corresponding data sources and collection methods. In addition, we are exploring development of metrics related to assessing the level of integration of CA with other information systems as well as the patient's overall healthcare system (e.g. stand-alone applications versus CA applications integrated with their healthcare professionals). Further work in developing metrics for assessing the impact of CA on health outcomes is also needed.

## References

[1] Han HJ, Mendu S, Jaworski BK et al. PTSDialogue: Designing a Conversational Agent to Support Individuals with Post-Traumatic Stress Disorder. Adjunct Proceedings of the 2021 ACM International Joint Conference on Pervasive and Ubiquitous Computing New York, NY, USA, 198–203.

[2] May R, Denecke K. Extending Patient Education with CLAIRE: An Interactive Virtual Reality and Voice User Interface Application. In Addressing Global Challenges and Quality Education: 15th European Conference on Technology Enhanced Learning, EC-TEL 2020, Heidelberg, Germany, September 14–18, 2020, Proceedings. Springer-Verlag, Berlin, Heidelberg, 482–486.

[3] Bickmore T, Giorgino T. Health dialog systems for patients and consumers. J Biomed Inform. 2006 Oct;39(5):556–71.

[4] Stoyanov SR, Hides L, Kavanagh DJ et al. Mobile app rating scale: a new tool for assessing the quality of health mobile apps. JMIR mHealth and uHealth 2015; 3: e27-e27. DOI: 10.2196/mhealth.3422.

[5] Hensher M, Cooper P, Dona SWA et al. Scoping review: Development and assessment of evaluation frameworks of mobile health apps for recommendations to consumers. Journal of the American Medical Informatics Association : JAMIA 2021; 28: 1318-1329. DOI: 10.1093/jamia/ocab041

[6] Casas J, Tricot M-O, Khaled OA et al. Trends &amp; Methods in Chatbot Evaluation. Companion Publication of the 2020 International Conference on Multimodal Interaction; 2020; Virtual Event, Netherlands

[7] Denecke K, Abd-Alrazaq A, Househ M, Warren J. Evaluation Metrics for Health Chatbots: A Delphi Study. Methods Inf Med. 2021 Dec;60(5-06):171-179.

[8] Lister K, Coughlan T, Iniesto F et al. Accessible conversational user interfaces: considerations for design. In Proceedings of the 17th International Web for All Conference (W4A '20). Association for Computing Machinery, 2020, pp. 1–11.

[9] Petersen SE, Ostendorf M. A machine learning approach to reading level assessment. Computer speech & language. 2009 Jan 1;23(1):89-106.

[10] Norman CD, Skinner HA. eHEALS: the eHealth literacy scale. Journal of medical Internet research. 2006 Nov 14;8(4):e507.

[11] Kayser L, Karnoe A, Furstrand D, Batterham R, Christensen KB, Elsworth G, Osborne RH. A multidimensional tool based on the eHealth literacy framework: development and initial validity testing of the eHealth literacy questionnaire (eHLQ). Journal of medical Internet research. 2018 Feb 12;20(2):e8371.

[12] Denecke K, May R. Usability Assessment of Conversational Agents in Healthcare: A Literature Review. Stud Health Technol Inform. 2022 May 25;294:169-173.

[13] Nielsen J. Usability engineering. Morgan Kaufmann; 1994 Oct 7.

[14] Mahatody T, Sagar M, Kolski C. State of the art on the cognitive walkthrough method, its variants and evolutions. Intl. Journal of Human–Computer Interaction. 2010 Jul 30;26(8):741-85.

[15] Wichansky AM. Usability testing in 2000 and beyond. Ergonomics. 2000 Jul 1;43(7):998-1006.

[16] May R, Denecke K. Security, privacy, and healthcare-related conversational agents: a scoping review. Inform Health Soc Care. 2022 Apr 3;47(2):194-210.