Caring is Sharing – Exploiting the Value in Data for Health and Innovation
M. Hägglund et al. (Eds.)

doi:10.3233/SHTI230094

Towards Safe Conversational Agents in Healthcare

Kerstin DENECKE[a,1]
[a] *Bern University of Applied Sciences, Bern, Switzerland*
ORCiD ID: Kerstin DENECKE https://orcid.org/0000-0001-6691-396X

Abstract. Conversational agents (CA) are becoming very popular to deliver digital health interventions. These dialog-based systems are interacting with patients using natural language which might lead to misunderstandings and misinterpretations. To avoid patient harm, safety of health CA has to be ensured. This paper raises awareness on safety when developing and distributing health CA. For this purpose, we identify and describe facets of safety and make recommendations for ensuring safety in health CA. We distinguish three facets of safety: 1) system safety, 2) patient safety, and 3) perceived safety. System safety comprises data security and privacy which has to be considered when selecting technologies and developing the health CA. Patient safety is related to risk monitoring and risk management, to adverse events and content accuracy. Perceived safety concerns a user's perception of the level of danger and user`s level of comfort during the use. The latter can be supported when data security is guaranteed and relevant information on the system and its capabilities are provided.

Keywords. Conversational agent, chatbot, safety, risk management, adverse events

1. Introduction

Supposed to be intuitively used, conversational agents (CA) are applied not only in customer service, but also in healthcare. Application areas include the collection of the medical history, supporting self-management or even delivering mental health interventions. CA can cause harm or death in users when badly designed and users rely upon them as authoritative source of information [1]. In particular unconstrained user input could lead to misinterpretations by the CA which in turn could result in wrong advice. Given technological advances, it is becoming increasingly difficult for users to distinguish CA from humans [2]. Created to simulate human behavior, a health CA is supposed to build a bond of trust to the patient and deals with health data. This creates a huge demand of assessing safety of those applications.

However, we can recognize a lack of research on errors and their impact on patient safety when applying CA in healthcare. A review of Abd-Alrazaq et al. on effectiveness and safety of using chatbots to improve mental health demonstrated that current systems are not seriously assessed towards safety [3]. They found two randomized controlled trials (RCT) out of 12 included in their review that reported on safety. However, the developers of those two systems concluded the systems are safe because no harm,

[1] Corresponding Author: Kerstin Denecke, Bern University of Applied Sciences, Institute for Medical Informatics, Quellgasse 21, 2502 Biel, Switzerland, E-mail: kerstin.denecke@bfh.ch.

distress, adverse events or worsening of the depressive symptoms were reported during the trial that was conducted. We believe that the non-reporting of harm or adverse events is insufficient to be able to conclude about the safety of CA. A more comprehensive analysis is essential. This requires a joint understanding of the facets of safety. Bickmore et al. studied patient and consumer safety risks when using CA (Siri, Alexa, Google Assistant) for gathering medical information [4]. They concluded that "relying on such assistance for actionable medical information represents a safety risk for patients". They found when asking questions that require medical expertise, CA failed more than half of the time and recommended actions could have resulted in harm for the user.

The large research interest in health CA and the missing reporting and assessment of safety aspects related to health CA raises the necessity to define the facets of health CA safety. This paper aims at synthesizing facets of safety in the context of health CA to form a common understanding, to make recommendations for ensuring safety in health CA and shape future research endeavors towards safe health CA.

2. Methods

To identify aspects related to safety of health CA, we were interested in concrete solutions of health CA described in literature. In previous work, we conducted a literature search to identify recent papers published between 2010 and 2022 in which health CA are presented. We searched for relevant scientific papers on PubMed, ACM Digital Library, and IEEE Xplore published between 2010 and 2022 and written in English. To identify appropriate literature, we defined the following search string: *(application OR app OR approach OR implementation) AND (chatbot OR bot OR conversation OR conversational user interface) AND (health OR healthcare)*. Only publications were included that were peer-reviewed conference papers or journal articles of original work. The publication had to present a concrete CA applied in healthcare. We excluded papers not dealing with a concrete healthcare-related CA or only describing the design process, reviews or meta-analyses. The review identified 222 relevant papers. Within these papers, we formed a subset of papers where the full text contained one of the terms *safe* (referring to safety) or *adverse* (referring to adverse event). The resulting 112 documents were manually assessed; they were considered for this paper when they reported on any kind of safety assessment or discussed safety aspects of their solutions. 76 papers contained at least one of the keywords, but used in a different context (e.g., a CA on safer sex or referenced papers with "safe" in the title). They were excluded. 36 papers fulfilled the criteria and were used to extract aspects related to safety were extracted. We aggregated the information, and derived facets of safety and recommendations on how to consider these facets in future health CA developments. The recommendations were derived from the retrieved information and from our experiences on working towards a standard evaluation framework for health CA [5–7].

3. Facets of Safety of Health CA

We can distinguish three facets of safety: 1) system safety, 2) patient safety, and 3) perceived safety. System safety concerns the content that is delivered by the health CA to the user, content reliability, correctness, data quality, data security and data transfer [8–10]. Patient safety concerns the risk that interacting with the health CA could harm a

patient, and perceived safety is the «user's perception of the level of danger and his/her level of comfort during the use" [11]. As follows, we describe the facets and make recommendations to address the safety aspects in the development of CA (Table 1).

In the reviewed papers, system safety primarily concerns data security: Health dialog is characterized by a secure environment. The patient can be sure that her information is only used for the treatment process. This creates a high degree of trust [2]. In case users of a health CA are not confident that data privacy is ensured, they will be less willing to disclose correct information. Health CA must be designed with data privacy in mind, not just for legal requirements but also so that vulnerable patients may develop the confidence necessary to provide personal information.

Table 1. Safety facets with recommendations how to address them during health CA development

Safety facet	Questions to consider	Recommendations
System safety	Is data privacy and data security ensured? Is a data privacy policy available in the health CA?	Robust against cyber attacks Penetration testing Ethical hacking [12] Assessing technical errors (including language understanding and response generation [5]) Providing information on data security and processing including information on use of third part tools
Patient safety	Who is the user? Which safe-critical situations might occur while interacting with the system? Which unexpected consequences might arise (e.g. app-app interactions, adverse events, worsening of symptoms) Who is responsible when a risk occurs? Is only accurate, evidence-based information included in the CA responses and questions? Is the underlying knowledge base evidence-based? Were physicians / healthcare professionals involved in the content development of the health CA? Is there a maintenance process for the information included in the health CA? Is information on the developer or content provider of the health CA provided? Were patient organizations involved in the development of the health CA?	Ensure content accuracy by design Include safety measures for emergencies (e.g. redirect to emergency resources) Clearly describe the limitations of the health CA [13] Assess the adverse events and side effects of the health CA Include automatic risk classifiers (e.g. a self-harm risk classifier [14], or classifier for risks of suicide and violence [15])
Perceived safety	Which personal identifiable information is required and is only this information collected and stored? Does the health CA provides an environment for the user that is safe for disclosing personal information? Did the health CA only provides evidence-based information? Is the information provided in a way that is understandable by the user group of consideration?	Provide information on data use, sharing and storage Provide information on underlying clinical evidence base Provide information on information sources Consider eHealth and health literacy of the user Consider reading level of the user

Two aspects concern patient safety: a) medical safety (does the CA worsen or produce symptoms or diseases in a patient?), and b) emergency safety (Is fast assistance ensured in case of emergencies). In particular when a user is supposed to interact with a health CA for a medium or long-term period, the user might be confronted with a safety-critical situation. When the health CA is autonomous, it has to be ensured that it can recognize such situation and react appropriately [16]. Given the language-based

interaction in the context of health CA, safety problems could result from misunderstandings or misinterpretations of (unconstrained) user input, from inappropriate reaction to unexpected user input, missing knowledge-based interpretation of user input (e.g. drug-drug interactions are not recognized [1]) and missing safety measures when misunderstanding or unexpected user input occurred. To ensure patient safety, mechanisms have to be in place to handle potential health risks such as suicidality, violence or risk of self-harm [17]. This can be realized by including safety plans, generating warnings when a risk is determined, sending appropriate referrals to emergency hotlines, or other contact persons. The content has to be accurate which can be ensured by involving healthcare professionals in the development and relying upon clinical evidence.

Perceived safety can depend on system safety such as data security. Patients reported to feel safe in disclosing information because they have the impression of sharing information with themselves [18]. This is only possible when system safety in general and data security in particular is guaranteed. There are several open issues related to perceived safety: How to measure perceived safety of CA users in healthcare? When do patients feel safe using a CA? What features of health CA are required to feel safe? A user survey could help identifying answers to latter questions.

4. Discussion and Conclusion

In this paper, we described three facets of safety related to health CA: system, patient and perceived safety and we made recommendation how to ensure safety of health CA. There exists a safety event taxonomy, the JCAHO patient safety event taxonomy [19]. This taxonomy describes communication as one of the processes in healthcare that can be faulty or fail and may therefore result in adverse events. It is therefore essential, to study the safety of health CA, since the interaction between a health CA and its users is realized as communication. However, a standardized methodology for assessing safety of health CA along these three facets is still missing. Jang et al. used a questionnaire comprising five aspects to assess Cas` side effects [20]. The aspects are: a disease specific side effect (increase of negative emotional experiences), privacy infringement, sense of alienation from everyday life, violation of therapeutic boundaries, regression of the therapeutic process. Miner et al. studied the CA`s reaction to emergency situations. They analyzed CA`s responses to short emergency messages posted by users [21]. Even though highly relevant, these approaches study only some aspects of the three facets of safety we identified in this work.

The papers reporting on clinical trials with health CA sometimes conclude that the system is “safe to use” since no adverse events were reported by the users [22]. The question arises what is considered as adverse event. Having in mind adverse events in clinical trials, this could be symptoms that occurred during the study period. But when interacting with a machine, additional adverse events may arise, that researchers currently not analyze (e.g., app-app interactions, upcoming addictions to the technology, impact on social activities). To address these issues, we recommend future research on possible adverse events of health CA. This could result in a taxonomy of such events and safety aspects which would contribute to a common view of relevant and possible adverse events. A harmonized safety risks assessment framework could help in improving the assessment of safety risks due to health CA usage. Finally, developing a reporting guideline for safety assessment in health CA would support transparency. From

a patient perspective, an information sheet on possible adverse events and safety risks similar to the package inlet for drugs could help in increase perceived safety and would contribute to an informed patient who is reflecting critically the symptoms that occur and can ask for professional help when necessary.

References

[1] Bickmore T, Trinh H, Asadi R, Olafsson S. Safety First: Conversational Agents for Health Care. In: Moore RJ, Szymanski MH, Arar R, Ren G-J, editors. Studies in Conversational UX Design [Internet]. Cham: Springer International Publishing; 2018 [cited 2022 Nov 6]. p. 33–57. Available from: http://link.springer.com/10.1007/978-3-319-95579-7_3

[2] Bickmore T, Giorgino T. Health dialog systems for patients and consumers. Journal of Biomedical Informatics. 2006;39:556–71.

[3] Abd-Alrazaq AA, et al. Effectiveness and Safety of Using Chatbots to Improve Mental Health: Systematic Review and Meta-Analysis. J Med Internet Res. 2020;22:e16021.

[4] Bickmore TW, Trinh H, Olafsson S, O'Leary TK, Asadi R, Rickles NM, et al. Patient and Consumer Safety Risks When Using Conversational Assistants for Medical Information: An Observational Study of Siri, Alexa, and Google Assistant. J Med Internet Res. 2018;20:e11510.

[5] Denecke K, Abd-Alrazaq A, Househ M, Warren J. Evaluation Metrics for Health Chatbots: A Delphi Study. Methods Inf Med. 2021;60:171–9.

[6] May R, Denecke K. Security, privacy, and healthcare-related conversational agents: a scoping review. Informatics for Health and Social Care. 2022;47:194–210.

[7] Denecke K, May R. Usability Assessment of Conversational Agents in Healthcare: A Literature Review. In: Séroussi B, Weber P, Dhombres F, Grouin C, Liebe J-D, Pelayo S, et al., editors. Studies in Health Technology and Informatics [Internet]. IOS Press; 2022 [cited 2022 Jun 20].

[8] 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. CHItaly 2021: 14th Biannual Conference of the Italian SIGCHI Chapter [Internet]. Bolzano Italy: ACM; 2021 [cited 2022 Jan 13]. p. 1–5.

[9] Tschanz M, et al. Using eMMA to Manage Medication. Computer. 2018;51:18–25.

[10] Ollier J, et al. Elena+ Care for COVID-19, a Pandemic Lifestyle Care Intervention: Intervention Design and Study Protocol. Front Public Health. 2021;9:625640.

[11] Amini R, Lisetti C, Yasavur U, Rishe N. On-Demand Virtual Health Counselor for Delivering Behavior-Change Health Interventions. 2013 IEEE International Conference on Healthcare Informatics. 2013. p. 46–55.

[12] Ganapathy S, et al. Acute paediatrics tele-support for caregivers in Singapore: an initial experience with a prototype Chatbot: UPAL. Singapore Med J. Singapore; 2021;

[13] Mauriello ML, et al. A Suite of Mobile Conversational Agents for Daily Stress Management (Popbots): Mixed Methods Exploratory Study. JMIR Form Res. 2021;5:e25294.

[14] Deshpande S, Warren J. Self-Harm Detection for Mental Health Chatbots. Stud Health Technol Inform. Netherlands; 2021;281:48–52.

[15] Collins C, et al. Covid Connect: Chat-Driven Anonymous Story-Sharing for Peer Support. Designing Interactive Systems Conference. Virtual Event Australia: ACM; 2022. p. 301–18. 5

[16] Tielman M, et al. A Therapy System for Post-Traumatic Stress Disorder Using a Virtual Agent and Virtual Storytelling to Reconstruct Traumatic Memories. Journal of medical systems. 2017;41:125.

[17] Hungerbuehler I, et al. Chatbot-Based Assessment of Employees' Mental Health: Design Process and Pilot Implementation. JMIR Form Res. 2021;5:e21678.

[18] Brandtzaeg PB, Følstad A. Chatbots: Changing User Needs and Motivations. Interactions. New York, NY, USA: Association for Computing Machinery; 2018;25:38–43.

[19] Chang A, Schyve PM, Croteau RJ, O'Leary DS, Loeb JM. The JCAHO patient safety event taxonomy: a standardized terminology and classification schema for near misses and adverse events. Int J Qual Health Care. 2005;17:95–105.

[20] Jang S, Kim J-J, Kim S-J, Hong J, Kim S, Kim E. Mobile app-based chatbot to deliver cognitive behavioral therapy and psychoeducation for adults with attention deficit: A development and feasibility/usability study. Int J Med Inform. Ireland; 2021;150:104440.

[21] Miner AS, et al. Talking to Machines About Personal Mental Health Problems. JAMA. 2017;318:1217.

[22] Maher CA, et al. A Physical Activity and Diet Program Delivered by Artificially Intelligent Virtual Health Coach: Proof-of-Concept Study. JMIR Mhealth Uhealth. 2020;8:e17558.