PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Protocol for the Development of the Chatbot Assessment Reporting Tool (CHART) for Clinical Advice
AUTHORS	CHART Collaborative, The; Huo, Bright

VERSION 1 – REVIEW

REVIEWER	Kohlboeck, Gabriele Sandoz GmbH, Biopharmaceuticals MS&T
REVIEW RETURNED	14-Nov-2023

GENERAL COMMENTS	Review: Protocol for the Development of the Chatbot Assessment Reporting Tool (CHART) for Clinical Advice: The authors work on a protocol for medical chatbot assessment research which is a very urgent topic, as this application can be easily used by the patients given the shortage of medical professionals, and is considered to be a new genre of medical research that might revolutionize medical advice. Chatbots provide answers to lay prompts, which have a great potential to help patients in all kind of medical specialties who might seek urgent medical advice, but have no personal access to the care they would require from physicians, doctors, psychologists or psychiatrists etc. However, as the authors already have stated these chatbots need to be very carefully programmed to meet certain standards and guidelines, which seem currently not to exist. However, there are some points that the authors should consider in their protocol to make sure chatbots will be a safe medical application to get medical advice for lay people in the future. Some of the points mentioned here might be too early to consider (e.g. the patients perspective), but might be essential at a later point: Major points: 1. Patients use this application for their own treatment, so there lies a heavy responsibility on the chatbot developers and programmers to meet requirements of medical interventions and applications in terms of safety and efficacy. Therefore, the protocol for this artificial intelligence application in health care must also comply with existing guidelines for clinical research, such as International Council for Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH) guidelines or Declaration of Helsinki as a statement of ethical principles for medical research. How do the authors ensure that principle safety principles and ethics as a first step in medical research or clinical studies are included? This has not been described sufficiently yet.

	2. Phase one p. 10 identification of checklist items with EQUATOR network: how do the authors make sure that all existing worldwide guidelines written in English are included? The equator network only refer to UK, France, and China guidelines. Why they have not included standards, norms and guidelines from the US or Europe? a. The authors state that Equator identified guidelines are further checked according to their PubMed references. Why not use PubMed for searching for chatbot guidelines as a first instance then? The authors should describe why the search in Equator is required. Pubmed and other medical databases might also list all relevant studies with chatbot research or guidelines. b. In this section it is confusing whether the authors search for guidelines on chatbot research or if they are searching for chatbot research per se (refers to p. 10, particularly to lines 32-50). If there are no existing guidelines why search for them? The authors should specify whether they are searching for medical guidelines in general or for guidelines, norms, standards and specifications for developing chatbots. c. p. 11 line 13-16: It is not clear to me how out of the guidelines from the EQUATOR network the checklist items are identified. Please give a short summary on how this is planned to do. Minor Points: 3. Phase two scoping review: Objective beginning line 28: a. Performance of LLMs: The authors should go more into detail here which performance of LLMs is meant here: for the LLM itself (size, speed) or its outcome for the patient (e.g. safe medical information?). b. It would also be of interest to know whether for the design and performance of the chatbot the perspective of the lay people (patients) is considered here: how accessible is the chatbot to patients having difficulties with IT applications? Does the design of the chatbot allow access for blind people and how are elderly patients having problems with their sight considered? How is human interface device design in general considered? The definition of a human interface device or HID might also apply to chatbots: HID is a type of computer device usually used by humans that takes input from humans and gives output to humans. 4. Phase Three: The authors should clarify here whether they want to develop a guideline for studies assessing chatbot performance or a guideline for the (safety) use of medical chatbots. a. Expert panel p. 14: I would suggest to include more than two patient partners, as at least at its final step, safety chatbots are used by patients. Therefore, a diverse patient's perspective in terms of accessibility and usability needs to be considered here. b. P. 18 line 4-10: identify authors in the scoping review...why is this not stated in its respective section, i.e. the second phase? Here should be made clear that the chatbot study authors do evaluate their own study based on the checklist provided by the authors of the protocol. Would this be step 4 "evaluation or validation" then? Thank you for the very well written research protocol, and I consider it as a kind of pioneer work in its field. I am sure it is very interesting to read for a wide audience to get knowledge of medical chatbots and its potential use for lay people.

REVIEWER	Lim, Gilbert

	National University of Singapore, School of Computing
REVIEW RETURNED	21-Nov-2023

GENERAL COMMENTS	This manuscript presents a protocol for the development of the Chatbot Assessment Reporting Tool (CHART) reporting guideline. This is a timely proposal given the recent surge of interest in LLMs and chatbots. Some issues might be considered: 1. Following editorial guidelines, proposed dates for the various phases of the study might be provided. 2. For Phase One, the use of only the EQUATOR network for initial identification of reporting guidelines might be justified. Moreover, the subsequent literature search does not appear to be as extensively described/defined as for Phase Two. This might be addressed. 3. For Phase Two, it is stated that a separate protocol presents the search strategy and other details of the scoping review (the sentence also contains a repeated "in a separate protocol" at the end). The nature of this separate protocol might be described further - is it still under development, in which case its development and due date might be provided, or has it been completed/published, in which case it might be cited. 4. For the Advisory Committee makeup in Phase Three, it is stated that a snowballing method will be used to identify members. This snowballing method might be described further, in particular justifying its ability to select a sufficiently-comprehensive committee. 5. In the Expert Panel subsection, a threshold of 80% consensus is stated for grouping purposes. The determination of this threshold (e.g. ad-hoc, by convention, etc.) might be briefly explained. 6. It is stated that the duration of both Synchronous Consensus Meetings will be at most four hours each. It might be clarified as to whether eight hours total is expected to be sufficient to reach a conclusion, and also what contingencies are in place if not.

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1

Dr. Gabriele Kohlboeck, Sandoz GmbH

The authors work on a protocol for medical chatbot assessment research which is a very urgent topic, as this application can be easily used by the patients given the shortage of medical professionals, and is considered to be a new genre of medical research that might revolutionize medical advice. Chatbots provide answers to lay prompts, which have a great potential to help patients in all kind of

medical specialties who might seek urgent medical advice, but have no personal access to the care they would require from physicians, doctors, psychologists or psychiatrists etc. However, as the authors already have stated these chatbots need to be very carefully programmed to meet certain standards and guidelines, which seem currently not to exist.

However, there are some points that the authors should consider in their protocol to make sure chatbots will be a safe medical application to get medical advice for lay people in the future. Some of the points mentioned here might be too early to consider (e.g. the patients perspective), but might be essential at a later point:

Major points:

1. Patients use this application for their own treatment, so there lies a heavy responsibility on the chatbot developers and programmers to meet requirements of medical interventions and applications in terms of safety and efficacy.

Therefore, the protocol for this artificial intelligence application in health care must also comply with existing guidelines for clinical research, such as International Council for Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH) guidelines or Declaration of Helsinki as a statement of ethical principles for medical research.

How do the authors ensure that principle safety principles and ethics as a first step in medical research or clinical studies are included? This has not been described sufficiently yet.

- We thank reviewer #1 for this salient point. We agree that ethical principles should be considered in this work.
- We have reviewed the guidelines suggested by reviewer #1. We find the Declaration of Helsinki particularly relevant for our work.
    a. We have created an ethics subheading outlining that ethics approval was submitted to HiREB at our institution and deemed not to require their review, and that we will adhere to the Declaration of Helsinki principles.
- We have recruited an ethicist and regulatory expert to join our team We will also be meeting separately with them to capture key ethics & safety principles for discussion in preparation for our consensus meeting, added to our protocol as follows:
    a. “In preparation for the next phase, the steering committee will meet with an ethicist and regulatory expert to review draft checklist items from the Delphi process to revise or add key principles of ethics & safety for discussion during the consensus meeting."

2. Phase one p. 10 identification of checklist items with EQUATOR network: how do the authors make sure that all existing worldwide guidelines written in English are included? The equator network only refer to UK, France, and China guidelines. Why they have not included standards, norms and guidelines from the US or Europe?

- We agree that a comprehensive search should be conducted.
- It is not the case that the library only contains guidance from those 5 countries, but rather the EQUATOR Network is an international initiative established in 2006 that collates and hosts a comprehensive library of reporting guidelines and includes guidance irrespective of country or language for inclusion in the EQUATOR library.
- Note that the EQUATOR Network just comprises 5 centres in the UK, France, Canada, China and Australia, and the library of reporting guidelines is hosted by the UK EQUATOR Centre, but that inclusion of reporting guidelines are not restricted to those 5 countries.
- Reporting guidelines are also rarely from a single country and are typically the product of multi-institutional, international collaborations.

a. The authors state that Equator identified guidelines are further checked according to their PubMed references. Why not use PubMed for searching for chatbot guidelines as a first instance then? The authors should describe why the search in Equator is required. Pubmed and other medical databases might also list all relevant studies with chatbot research or guidelines.

- As we were interested in checklists that meet rigorous standards, the search through the EQUATOR network was added to explicitly outline a systematic, comprehensive approach to identifying high-quality reporting guidelines.
- As noted above the EQUATOR library is a comprehensive database of reporting guidelines (www.equator-network.org).

b. In this section it is confusing whether the authors search for guidelines on chatbot research or if they are searching for chatbot research per se (refers to p. 10, particularly to lines 32-50). If there are no existing guidelines why search for them? The authors should specify whether they are searching for medical guidelines in general or for guidelines, norms, standards and specifications for developing chatbots.

- We thank the reviewer for this point.
- Though we anticipated that no reporting guidelines exist, due diligence was required to perform a comprehensive search to identify completed or in-progress reporting guidelines in this area.
- For clarity, the objective under "PHASE ONE" now states that our aim is "to identify checklist items used in previous reporting guidelines and identify related reporting standards for studies assessing the ability of LLMs to provide clinical advice."
- We also now clarify the broader scope of LLMs applicable to studies that would use this reporting guideline in a paragraph prior to "PHASE ONE:"
    a. "This reporting guideline will emphasize transparent reporting standards for studies evaluating the performance of LLMs when providing clinical advice to patients and clinicians. It will apply to LLM-linked chatbots, but also LLMs more broadly. It will also apply to studies using both traditional and multimodal LLMs."

c. p. 11 line 13-16: It is not clear to me how out of the guidelines from the EQUATOR network the checklist items are identified. Please give a short summary on how this is planned to do.

- We have outlined that we will use the "search" feature and toggle through all study types. Any relevant reporting guidelines will be reviewed, including their checklists.

Minor Points:

3. Phase two scoping review: Objective beginning line 28:

a. Performance of LLMs: The authors should go more into detail here which performance of LLMs is meant here: for the LLM itself (size, speed) or its outcome for the patient (e.g. safe medical information?).

- We thank the reviewer for the opportunity to add clarity.
- We have added details regarding some of the outcomes we will extract in "PHASE TWO." In our scoping review, we are interested in both examples illustrated above – the LLM characteristics reported by Chatbot Assessment Studies, but primarily how the appropriateness of the advice for patients and clinicians was assessed, as reported in Chatbot Assessment Studies.
- See response to point 4 below.

b. It would also be of interest to know whether for the design and performance of the chatbot the perspective of the lay people (patients) is considered here: how accessible is the chatbot to patients

having difficulties with IT applications? Does the design of the chatbot allow access for blind people and how are elderly patients having problems with their sight considered? How is human interface device design in general considered? The definition of a human interface device or HID might also apply to chatbots: HID is a type of computer device usually used by humans that takes input from humans and gives output to humans.

- We agree with reviewer #1 that patient input is necessary.
    a. We will recruit numerous patients to participate in the Delphi consensus process, while two patient partners will participate in the synchronous consensus meetings.
- Though the scope of our reporting guideline will be focused on developing transparent reporting standards for studies assessing the clinical performance of LLMs, rather than the development of the LLM itself, discussions for candidate checklist items will be held in the context of the needs and accessibility of various patient populations with respect to interacting with LLMs including but not limited to, the visually impaired and the elderly.

4. Phase Three: The authors should clarify here whether they want to develop a guideline for studies assessing chatbot performance or a guideline for the (safety) use of medical chatbots.

- We clarify that we are developing a reporting guideline (distinct from a clinical practice guideline) for studies assessing clinical chatbot performance.
- The studies that use our reporting tool will be evaluating the performance of the chatbot, secondarily the chatbot's technical performance, but by far most importantly with respect to the accuracy of the chatbot's summary of evidence or recommendations.
- If the evaluations are well done, they will faithfully report on whether the chatbots have provided accurate information and sound recommendations (and thus promote patient safety) or inaccurate information and misleading recommendations (in which case they may endanger patient safety).
- Our reporting tool will promote optimal chatbot evaluations and thus may indirectly promote patient safety.

a. Expert panel p. 14: I would suggest to include more than two patient partners, as at least at its final step, safety chatbots are used by patients. Therefore, a diverse patient's perspective in terms of accessibility and usability needs to be considered here.

- We agree that patient input is vital. The CHART study will include multiple patient partners in the Delphi consensus process as well as two patient partners in the panel consensus meetings.
- However, we have the challenge of maintaining a feasible and efficient panel size with adequate representation of other key stakeholders such as statisticians, research methodologists, reporting checklist developers, NLP researchers, journal editors, chatbot researchers, ethicists, regulatory experts, and policy experts.
- To achieve the goals of this representation, and a functional and diverse panel, we will maintain two patient partners in the panel.
- However, we have outlined an explicit approach to recruiting patients to participate in the Delphi process under "*Advisory Committee & Delphi*" consisting of both public and internal calls through our affiliate journals, and through snowballing via our panel, including our patient partner members.

b. P. 18 line 4-10: identify authors in the scoping review...why is this not stated in its respective section, i.e. the second phase? Here should be made clear that the chatbot study authors do evaluate their own study based on the checklist provided by the authors of the protocol. Would this be step 4 "evaluation or validation" then?

- We appreciate Reviewer #1's comment. We have added a subheading "Pilot Testing."

- This is listed under “Pilot Testing” rather than the scoping review section to avoid confusion (readers would not understand the purpose of identifying authors if stated in the scoping review section).
- To avoid bias, study authors of chatbot assessment studies will use the checklist to evaluate studies written by other authors and provide feedback on their experience.
- We have clarified this under “Pilot Testing” as follows:
    a. “Authors will not evaluate their own studies but will use the checklist to assess Chatbot Assessment Studies published by other authors.”

Reviewer: 2

Some issues might be considered:

1. Following editorial guidelines, proposed dates for the various phases of the study might be provided.

- We are grateful for this feedback.
- The revised protocol now outlines the study timeline in Figure 1, in “Study Overview & Objectives.”

2. For Phase One, the use of only the EQUATOR network for initial identification of reporting guidelines might be justified. Moreover, the subsequent literature search does not appear to be as extensively described/defined as for Phase Two. This might be addressed.

- We thank reviewer #2 for this feedback.
- We agree that the use of only the EQUATOR network can be justified, as we have done in response to Reviewer #1 point 2a as follows:
    a. “As we were interested in checklists that meet rigorous standards, the search through the EQUATOR network was added to explicitly outline a systematic, comprehensive approach to identifying high-quality reporting guidelines.”
- Reviewer #2 acknowledged in point #3 (below) that we have a separate protocol for the scoping review. This is being submitted for publication, however, we have also added further detail regarding the literature search, as follows:
    a. “In brief, the scoping review team will conduct a literature search using MEDLINE via Ovid, EMBASE via Elsevier, Scopus, and Web of Science to capture relevant studies published prior to October 2023. The team will identify studies that evaluate the performance of LLM-linked chatbots when providing clinical advice. We will only consider primary data. The team will complete two rounds of screening by title and abstract and full-text to identify articles of interest. Next, we will perform manual forward and backward citation searching.”

3. For Phase Two, it is stated that a separate protocol presents the search strategy and other details of the scoping review (the sentence also contains a repeated "in a separate protocol" at the end). The nature of this separate protocol might be described further - is it still under development, in which case its development and due date might be provided, or has it been completed/published, in which case it might be cited.

- We thank reviewer #2 for this feedback.
- This scoping review protocol was just completed and will be submitted for publication – this has been stated in the protocol now under “PHASE TWO.”

4. For the Advisory Committee makeup in Phase Three, it is stated that a snowballing method will be used to identify members. This snowballing method might be described further, in particular justifying its ability to select a sufficiently-comprehensive committee.

- We thank reviewer #2 for the opportunity to re-evaluate our approach.
- We have removed the snowballing method for the recruitment of general committee members, as we will attract a sufficiently comprehensive group of stakeholders (we anticipate several hundred) by identifying authors of prior studies using SCImago and Web of Science, as described in the protocol.

5. In the Expert Panel subsection, a threshold of 80% consensus is stated for grouping purposes. The determination of this threshold (e.g. ad-hoc, by convention, etc.) might be briefly explained.

- This 80% threshold of consensus has been employed in many reporting guidelines/ consensus studies: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7598943/.
- We specify in the protocol that this decision was made by the Steering Committee.

6. It is stated that the duration of both Synchronous Consensus Meetings will be at most four hours each. It might be clarified as to whether eight hours total is expected to be sufficient to reach a conclusion, and also what contingencies are in place if not.

- Based on collective experience of study authors including prior reporting guideline developers and research methodologists, they anticipate that eight hours will be sufficient.
- Still, we have added a contingency plan to be proactive as follows: “A contingency plan is set to pre-emptively arrange and hold a third meeting of two to four hours should additional time be needed following the eight hours of consensus meetings.”

We thank you for your time and consideration.

VERSION 2 – REVIEW

REVIEWER	Lim, Gilbert National University of Singapore, School of Computing
REVIEW RETURNED	16-Feb-2024

GENERAL COMMENTS	We thank the authors for addressing our previous comments.

VERSION 2 – AUTHOR RESPONSE