

# BMJ Open

BMJ Open is committed to open peer review. As part of this commitment we make the peer review history of every article we publish publicly available.

When an article is published we post the peer reviewers' comments and the authors' responses online. We also post the versions of the paper that were used during peer review. These are the versions that the peer review comments apply to.

The versions of the paper that follow are the versions that were submitted during the peer review process. They are not the versions of record or the final published versions. They should not be cited or distributed as the published version of this manuscript.

BMJ Open is an open access journal and the full, final, typeset and author-corrected version of record of the manuscript is available on our site with no access controls, subscription charges or pay-per-view fees (<http://bmjopen.bmj.com>).

If you have any questions on BMJ Open's open peer review process please email [info.bmjopen@bmj.com](mailto:info.bmjopen@bmj.com)

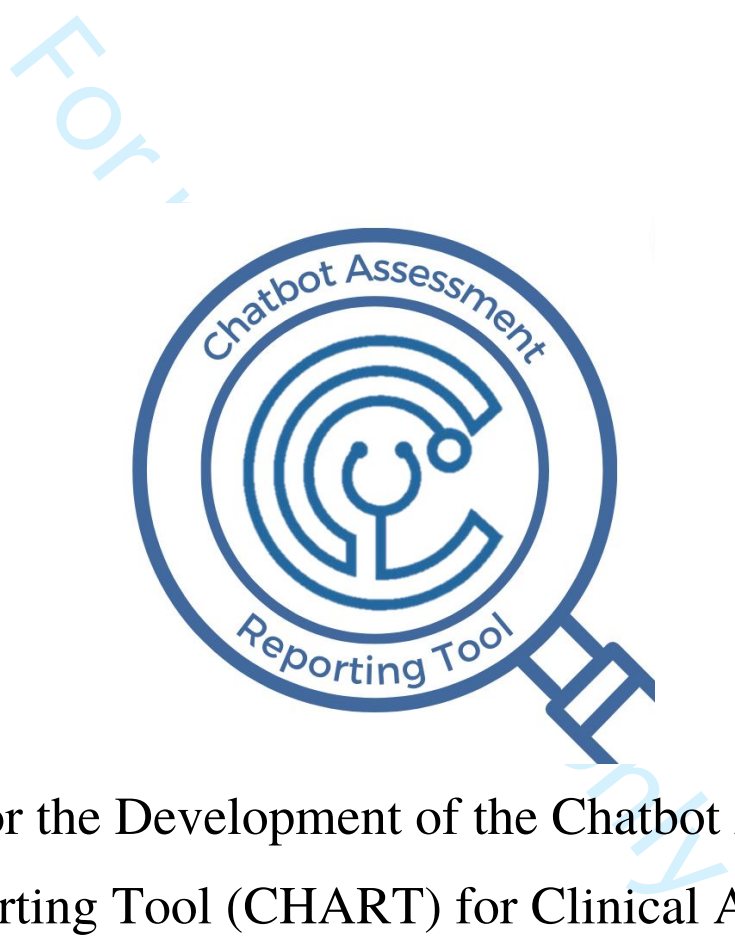
# BMJ Open

## Protocol for the Development of the Chatbot Assessment Reporting Tool (CHART) for Clinical Advice

Journal:	<i>BMJ Open</i>
Manuscript ID	bmjopen-2023-081155
Article Type:	Protocol
Date Submitted by the Author:	24-Oct-2023
Complete List of Authors:	CHART Collaborative, The; McMaster University Huo, Bright; Dalhousie Medical School,
Keywords:	MEDICAL ETHICS, STATISTICS & RESEARCH METHODS, Natural Language Processing

SCHOLARONE™  
Manuscripts

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60



# Protocol for the Development of the Chatbot Assessment Reporting Tool (CHART) for Clinical Advice

---

The CHART Collaborative\*

**Corresponding author:**

Bright Huo, on behalf of The CHART Collaborative\*

237 Barton St E, Hamilton, ON L8L 2X2

E: brighthuo@dal.ca

T: +1 902 448 6836

**Keywords:** Medical ethics, statistics & research methods, natural language processing

**ABSTRACT****Introduction:**

Large language model (LLM)-linked chatbots are being increasingly applied in healthcare due to their impressive functionality and public availability. Studies have assessed the ability of LLM-linked chatbots to provide accurate clinical advice. However, the methods applied in these Chatbot Assessment Studies are inconsistent due to the lack of reporting standards available, which obscures the interpretation of their study findings. This protocol outlines the development of the Chatbot Assessment Reporting Tool (CHART) reporting guideline.

**Methods and analysis:**

The development of the CHART reporting guideline will consist of three phases, led by the Steering Committee. During phase one, the team will identify relevant reporting guidelines with artificial

1  
2  
3 intelligence extensions that are published or in-development by searching preprint servers, protocol  
4  
5  
6 databases, and the Enhancing the Quality and Transparency of health research (EQUATOR) Network.  
7

8  
9 During phase two, we will conduct a scoping review to identify studies that have addressed the  
10  
11 performance of LLM-linked chatbots in summarizing evidence and providing clinical advice. The  
12  
13 Steering Committee will identify methodology used in previous Chatbot Assessment Studies. Finally, the  
14  
15 study team will use checklist items from prior reporting guidelines and findings from the scoping review  
16  
17 to develop a draft reporting checklist. We will then perform a Delphi consensus and host two synchronous  
18  
19 consensus meetings with an international, multidisciplinary group of stakeholders to refine reporting  
20  
21 checklist items and develop a flow diagram.  
22  
23  
24  
25  
26  
27  
28  
29

### 30 31 32 33 34 **Ethics and dissemination:**

35  
36  
37 We will publish the final CHART reporting guideline in peer-reviewed journals and will present findings  
38  
39 at peer-reviewed meetings. Ethical approval is not applicable for the development of the CHART  
40  
41 reporting guideline.  
42  
43  
44  
45  
46  
47  
48

### 49 **Registration:**

50  
51  
52 This study protocol is pre-registered with Open Science Framework:

53  
54  
55 <https://doi.org/10.17605/OSF.IO/59E2Q>.  
56  
57  
58

**Strengths and limitations of this study:**

- This initiative will address a lack of reporting standards for Chatbot Assessment Studies and will provide a framework to increase the transparent conduct of these studies.
- We will apply rigorous methodology of the highest standards to develop the CHART reporting guideline.
- A diverse group of international, multidisciplinary stakeholders will inform the development of the CHART reporting checklist and flow diagram, with key input from experts in LLMs.
- This reporting guideline will be developed swiftly while acknowledging the dynamically evolving technology of LLM-linked chatbots.
- The CHART reporting guideline will apply specifically for studies assessing the ability of LLM-linked chatbots to summarize evidence and provide clinical advice. It will not apply to their use in other settings.

**INTRODUCTION**

Novel chatbots such as ChatGPT have been integrating Large Language Models (LLMs), which are a popular technology in the field of natural language processing (NLP).<sup>1</sup> LLMs are large neural networks often comprised of hundreds of billions of parameters, which impact the model's input, size and shape,

1  
2  
3 and output.<sup>2</sup> LLMs are typically used to conditionally predict the next words in a sequence of text, given  
4  
5  
6 corresponding prompts (Table 1).<sup>3</sup> LLMs can be trained on a collection of massive amounts of raw data  
7  
8  
9 from online text sources including books, articles, websites, and more.<sup>1,4</sup> Coupled with reinforcement  
10  
11  
12 learning from human feedback,<sup>5</sup> LLMs exhibit striking text generation capabilities, producing outputs that  
13  
14  
15 are often indistinguishable from human language.<sup>6,7</sup> There has been a gold-rush movement of chatbots  
16  
17  
18 linked to LLMs, with recent releases including Bing Chat, Google Bard, Med-PaLM, and many more  
19  
20  
21 underway.<sup>8</sup>  
22  
23  
24  
25  
26  
27

28 Given their wide accessibility and ability to provide answers to lay prompts,<sup>8</sup> investigators have begun to  
29  
30  
31 assess LLM-linked chatbots as a potential source of health advice for both patients and clinicians.<sup>9-11</sup> We  
32  
33  
34 refer to these studies as Chatbot Assessment Studies, and they evaluate the performance of LLM-linked  
35  
36  
37 chatbots in summarizing health evidence and providing clinical advice. These studies represent a new  
38  
39  
40 genre of medical research, but the methodology and framing of results reported in these studies are highly  
41  
42  
43 variable. Inconsistent and incomplete reporting limits readers' ability to judge the methodology and  
44  
45  
46 results of these studies, complicating their interpretation.<sup>12</sup> A need exists to assess the rigour of their  
47  
48  
49 assessments,<sup>8</sup> but currently there are no standardized reporting tools for Chatbot Assessment Studies.  
50  
51  
52  
53  
54  
55

56 Instruments have been created to address issues of suboptimal reporting and raise the standard of research  
57  
58  
59  
60

quality, such as the Consolidated Standards of Reporting Trials (CONSORT) statement.<sup>13,14</sup> Such reporting guidelines provide a checklist and a flow diagram for a given study type. Since their development, extensions to reporting guidelines have been created to facilitate the integration of artificial intelligence.<sup>15-17</sup> However, LLM-linked chatbots and their accompanying applications have only recently emerged and are not captured by these reporting guidelines. This protocol outlines the development of a novel reporting checklist, the Chatbot Assessment Reporting Tool (CHART) to improve the reporting standards of Chatbot Assessment Studies.

### Key Terminology

Table 1 lists key terms included in this work.

Table 1. Glossary.

Term	Definition
Artificial Intelligence (AI)	The science of developing computer systems that can perform complex tasks approximating human cognitive performance.
Natural Language Processing (NLP)	A branch of information science that seeks to enable computers to interpret and manipulate human text.
Large Language Model (LLM)	A type of NLP model comprising large neural networks trained over large amounts of text, usually to produce an output of continuations of text from corresponding prompts, known as next word prediction. <sup>+</sup>
Next word prediction	The natural language processing task of predicting the next word in a sequence of text given context and model parameters.
Parameter	A <i>parameter</i> within an artificial intelligence algorithm is a variable that is tuned



	iteratively/automatically to optimize the intended outcome of the algorithm. Parameters may be at the model level to optimize tuning (hyperparameters) or "weights" within the model linking layer to layer (parameters)
LLM-Linked Chatbot	A program that permits users to interact with an algorithm (such as an LLM) designed to respond to user prompts.
Chatbot Assessment Study	Any research study assessing the performance of chatbots in summarizing health evidence and/or providing clinical advice.
Chat Instance	An interface in a computing device through which communication takes place between a chatbot and its user through text with only one prompt.
Chat Session	An interface in a computing device through which communication takes place between a chatbot and its user through text with more than one prompt.
Query	The act of communicating with a LLM by inputting a prompt into the chatbot which might be a question, comment, or phrase to elicit specific desired outputs from an LLM. For example, one might input a prompt asking the LLM to summarize the evidence supporting the use of a given intervention.
Check query	Following formal query completion and performance evaluation, the act of repeating the initial query to ensure that chatbot outputs are consistent in summarizing the same evidence and providing the same clinical advice.
Prompt	Text input by a user into the chatbot for the purpose of communicating with the LLM.
Prompt Engineering	An iterative testing phase where various pieces of text are inputted into a chatbot to achieve an output, informing the development of study prompts.
Delphi study	A structured research method applied to answer a research question through the establishment of consensus across respondents.

+Generally speaking, "next word" prediction is one basic "pre-training" objective, but LLMs similar to ChatGPT often undergo a subsequent round of "supervision" in which they are guided by human feedback.

-Chatbots are not necessarily built atop LLMs, but the modern tools that have captured public imagination (especially ChatGPT) are.

## METHODS & ANALYSIS

### Study Overview & Objectives

This study consists of three phases to address the following objectives:

1. To identify checklist items used in previous reporting guidelines and identify related reporting

standards for studies assessing the use of artificial intelligence in healthcare.

2. To perform a scoping review that will identify and characterize studies that have addressed the performance of LLMs in summarizing evidence and providing clinical advice. Specifically, the review will identify how authors evaluate chatbot performance in summarizing health evidence and providing clinical advice.
3. Informed by the scoping review and a review of prior checklists, to develop an evidence-informed, expert-derived reporting guideline comprised of a checklist and flow diagram for studies assessing chatbot performance in summarizing health evidence and providing clinical advice.

A Steering Committee will lead all key study initiatives. This group will include the following members: the project lead, the senior methodologist lead, an expert in chatbot assessment studies, a reporting checklist developer, and a journal editor. The group's responsibilities will be to guide the initiatives involved in the development of the CHART checklist. They will lead the review of relevant reporting checklists (phase one), the completion of the scoping review (phase two), and the development of the reporting guideline (phase three). Table 1 presents a glossary of key terms used in this work. Figure 1 demonstrates the timeline for the development of the CHART reporting guideline.

Figure 1. Development of the CHART Reporting Guideline.

## PHASE ONE

**Objective:** to identify checklist items used in previous reporting guidelines and identify related reporting standards for studies assessing the use of artificial intelligence in healthcare.

### *Identification of Existing Reporting Guidelines*

To identify relevant health research reporting guidelines to inform the development of our reporting guideline and checklist, the study team will search the EQUATOR network and identify reporting guidelines published prior to October 2023 that meet our inclusion criteria:

- Studies presenting primary data on the use of chatbots in any specialty in medicine.
- Studies applying chatbots to summarize evidence and provide clinical advice.
- Studies applying chatbots to answer one or more clinical question(s).
- Any studies applying chatbots as an intervention, with or without the use of a comparator.

We will review references from relevant reporting guidelines and related citations listed on PubMed for retrieved articles. To identify protocols of reporting guidelines, we will search Open Science Framework

1  
2  
3 as well as applicable results obtained from our scoping review. To identify ongoing or completed work  
4  
5  
6 not yet published in peer-reviewed sources, we will search Open Science Framework & MedRxiv.  
7  
8  
9

10  
11  
12 Reporting guidelines obtained from the search from phase one will inform the development of items for a  
13  
14  
15 preliminary draft version of the checklist.  
16  
17

## 18 19 20 21 22 **PHASE TWO**

23  
24  
25  
26  
27  
28 **Objective:** to perform a scoping review that will identify and characterize studies that have addressed the  
29  
30  
31 performance of LLMs in summarizing evidence and providing clinical advice. Specifically, the review  
32  
33  
34 will identify how authors evaluate chatbot performance in summarizing health evidence and providing  
35  
36  
37 clinical advice.  
38  
39  
40  
41  
42

43 For the scoping review, the project lead will recruit a team that will include two other members that have  
44  
45  
46 previous experience with performing systematic reviews and scoping reviews as well as the senior  
47  
48  
49 methodological lead. The scoping review team will identify articles assessing the performance of chatbots  
50  
51  
52 when applied in healthcare. A separate protocol presents our search strategy, inclusion criteria, exclusion  
53  
54  
55 criteria, and other details related to the scoping review in a separate protocol. Its development will be  
56  
57  
58  
59  
60

aligned with methodology guidance from the JBI Scoping Review Methodology Group.<sup>18</sup>

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.

Next, we will perform manual forward and backward citation searching. The team will complete two rounds of screening to identify articles of interest. Next, the team will perform data extraction to identify key items used in the reporting of these studies. We will report findings using descriptive statistics for quantitative data and present results graphically in diagrammatic form. A narrative summary will accompany the graphical results. The final report will adhere with reporting standards for the Preferred Reporting Items for Systematic Review and Meta-Analysis Extension for Scoping Reviews (PRISMA-ScR).<sup>19</sup>

### PHASE THREE

**Objective:** informed by the scoping review and a review of prior checklists, to develop an evidence-informed, expert-derived reporting guideline comprised of a checklist and flow diagram for studies assessing chatbot performance in summarizing health evidence and providing clinical advice.

## The CHART Reporting Guideline Research Protocol

2023/10/2

*Advisory Committee & Delphi*

An Advisory Committee will comprise epidemiologists, research methodologists, NLP researchers, journal editors, chatbot researchers, and patient partners. The Steering Committee will identify relevant stakeholders for inclusion in the Advisory Committee through the snowballing method. To achieve representation from all desired groups, they will prompt relevant stakeholders for member suggestions. Additionally, the Steering Committee will identify additional committee members by querying SCImago Journal Country Rank (SJR) portal ([www.scimagojr.com](http://www.scimagojr.com)) to obtain a list of the top 10 journals in each specialty in medicine. Using this list of journals, the Committee will query Web of Science to obtain a diverse list of researchers in medicine including general research methodologists and chatbot researchers. We will send an invitation email to our final list of contacts to invite them to join the Advisory Committee.

The Steering Committee will hold a synchronous virtual meeting open to all Advisory Committee members as an introduction to the project, as well as their role. Through a series of questionnaires shared through an online platform, the team will apply a Delphi consensus. The Steering Committee will develop a draft checklist informed by the scoping review and review of existing reporting guidelines. They will circulate the draft checklist to the Advisory Committee for a first round of voting. During this round, Advisory Committee members will select one of the following options for each checklist item: “include,

1  
2  
3 maybe include, uncertain, maybe exclude, exclude.” There will be an additional option for Advisory  
4  
5  
6 Committee members to once more add checklist items. The Steering Committee will then revise the  
7  
8  
9 checklist using comments from the first round. The team will re-circulate the updated draft checklist for a  
10  
11  
12 second round of voting, as above.  
13  
14  
15

16  
17  
18  
19 The Steering Committee will revise the checklist following the second round and present these items to  
20  
21  
22 the expert panel.  
23  
24  
25

### 26 27 28 *Expert Panel* 29

30  
31 We will create an international, multidisciplinary panel as per Moher and colleagues.<sup>12</sup> Participants will  
32  
33  
34 be purposefully selected to reflect a balanced representation of relevant stakeholders including  
35  
36  
37 statisticians, research methodologists, reporting checklist developers, NLP researchers, journal editors,  
38  
39  
40 chatbot researchers, and two patient partners. In preparation for the synchronous consensus meetings, the  
41  
42  
43 Steering Committee will share relevant materials with the panel such as the meeting agenda, participant  
44  
45  
46 list, and the completed scoping review highlighting the content and extent of reporting of the content area.  
47  
48  
49 The Committee will then circulate the draft checklist that emerged from the Delphi process to the Expert  
50  
51  
52 Panel through an electronic survey. We will group items with  $\geq 80\%$  consensus with the selection of  
53  
54  
55 “include” or “maybe include” together, posing to the panelists: “These items have been recommended for  
56  
57  
58  
59  
60

1  
2  
3 inclusion in our checklist. Do you agree or disagree?" Panelists will have the option of yes, no, unsure,  
4  
5  
6 and an additional option for comments.  
7  
8  
9

10  
11  
12 We will also group items with  $\geq 80\%$  consensus for items with the selection of "exclude" or "maybe  
13  
14  
15 exclude," posing to the panelists: "These items have been recommended for exclusion in our checklist.  
16  
17  
18 Do you agree or disagree?" Panelists will have the option of yes - include, no - exclude, and an additional  
19  
20  
21 option for comments. Items without 80% consensus will be gathered and panel members will indicate  
22  
23  
24 "include, maybe include, uncertain, maybe exclude, exclude." There will also be an additional option for  
25  
26  
27 each question to suggest additional checklist items. We will collate the results of this survey in  
28  
29  
30  
31 preparation for the Consensus Meetings.  
32  
33  
34  
35  
36  
37

### 38 Synchronous Consensus Meetings

39  
40 In advance of the Consensus Meetings, the Steering Committee will prompt panelists to share their  
41  
42  
43 conflicts of interest. Though we find it difficult to imagine circumstances that would lead to importance  
44  
45  
46 conflicts, we will stay alert to unanticipated conflicts. Should these arise, we will consider any panel  
47  
48  
49 member with significant conflicts as consultant who will not vote on the final checklist. Prior to the first  
50  
51  
52 of two Synchronous Consensus Meetings, the Steering Committee will share the candidate checklist items  
53  
54  
55  
56 with the Expert Panel which will have been revised following two Delphi rounds with the Advisory  
57  
58  
59  
60



1  
2  
3  
4 Committee, informed by findings from the scoping review.  
5  
6  
7  
8  
9

10 Additionally, the Steering Committee will construct a flow diagram prior to the Consensus Meetings  
11  
12 based on the candidate checklist items. The purpose of the flow diagram is to provide an overview to  
13  
14 guide authors in clearly reporting sequential stages of their study. The Steering Committee will also share  
15  
16 this flow diagram with the panel prior to the Consensus Meetings.  
17  
18  
19  
20  
21  
22  
23  
24

25 The project lead will organize two Synchronous Consensus Meetings that will be held over a video  
26  
27 conferencing platform. The Steering Committee will encourage panelists to attend both meetings, with the  
28  
29 expectation that panelists must attend one meeting, at minimum. The steering committee will circulate an  
30  
31 online scheduling survey in advance to control the number of participants in attendance, while also  
32  
33 selecting dates that optimize the attendance of panel members. As we will hold these meetings virtually,  
34  
35 no meeting will be longer than four hours in duration to mitigate burnout and encourage participation.  
36  
37  
38  
39  
40  
41  
42

43 The duration of both meetings will be eight hours in total.  
44  
45  
46  
47  
48  
49

50 During checklist item discussion, we will put forth any items rated as “no-exclude” to the panel for  
51  
52 exclusion from the checklist. We will then discuss any items without consensus or rated as “uncertain”  
53  
54 with  $\geq 80\%$  consensus after the second Delphi round. Finally, we will offer items rated as “yes-include” to  
55  
56  
57  
58  
59  
60

1  
2  
3 the panel for inclusion in the checklist. During the discussion for all checklist items, the meeting chair  
4  
5  
6 will present the following for each checklist item:  
7

- 8
- 9
- 10 • Previous use in a Chatbot Assessment Study
- 11
- 12
- 13 • Rationale for inclusion
- 14
- 15
- 16
- 17
- 18

19 All voting will take place virtually and anonymously over the video conferencing platform. A working  
20  
21 CHART checklist will emerge from the Synchronous Consensus Meetings. The panel will use this  
22  
23 working checklist to revise the draft CHART flow diagram during the Synchronous Consensus Meeting.  
24  
25  
26  
27  
28  
29  
30

31 Expert panel members who are unable to join will be able to review recordings of the meetings. The  
32  
33 project lead will record the meeting(s), and they will share both the recording and a summary of checklist  
34  
35 item decisions and rationale with absent panel members.  
36  
37  
38  
39  
40  
41  
42

43 Following the meetings, the Steering Committee will circulate the working CHART checklist and flow  
44  
45 diagram in the form of a survey reflecting checklist item decisions. This working checklist will outline a  
46  
47 final list of items for inclusion. Panellists will have the opportunity to provide any final comments, which  
48  
49 the Steering Committee will use to derive a preliminary CHART checklist.  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

## The CHART Reporting Guideline Research Protocol

2023/10/2

1  
2  
3  
4 The Steering Committee will pilot the preliminary CHART checklist and flow diagram with researchers  
5  
6 that have published Chatbot Assessment Studies and will identify authors by the included studies in the  
7  
8  
9  
10 scoping review. The Steering Committee will conduct pilot testing via an iterative process. Groups of five  
11  
12  
13 authors will provide feedback in each round until saturation is achieved, with a minimum of ten authors  
14  
15  
16 over two rounds of pilot testing. During synchronous sessions, we will ask authors to assess Chatbot  
17  
18  
19 Assessment Studies using the preliminary CHART checklist and flow diagram via think-aloud instrument  
20  
21  
22 testing. Authors will provide practical feedback regarding the development of these studies in the context  
23  
24  
25 of checklist items. They will also provide feedback regarding the practical application of the preliminary  
26  
27  
28 CHART checklist with respect to the length and content of the checklist.  
29  
30  
31  
32  
33

34 The Steering Committee will use the comments from Chatbot Assessment Study researchers to derive a  
35  
36  
37 final version of the CHART checklist and flow diagram.  
38  
39  
40  
41  
42

### 43 *Report Generation*

44  
45  
46 With the final CHART checklist and flow diagram, the Steering Committee will prepare a Statement  
47  
48  
49 document for submission for peer-reviewed conference presentation and publication. All panel members  
50  
51  
52  
53 will have the chance to review the draft manuscript, and all members of the research team satisfying the  
54  
55  
56 ICJME guidelines will join the group authorship.<sup>20</sup> The Statement article will consist of the checklist and  
57  
58  
59

1  
2  
3  
4 flow diagram. It will include the rationale for developing the CHART guideline and an overview of its  
5  
6  
7 development, including a brief description of the meeting and participants involved.  
8  
9  
10

11  
12 Separately, the Steering Committee will prepare a detailed explanation and elaboration paper (E&E). This  
13  
14  
15 paper will provide more detail for the inclusion of items in the final CHART checklist. For each checklist  
16  
17  
18 item, the E&E report will include three parts: 1) an explanation of the rationale supporting the checklist  
19  
20  
21 item, as well as reference to any supporting evidence for its inclusion 2) essential elements of the study  
22  
23  
24 that must be described to appropriately satisfy each checklist item 3) additional elements of the study  
25  
26  
27 which may be considered by authors depending on the context.  
28  
29  
30

31  
32  
33  
34 As per Moher and colleagues, we will simultaneously submit both the Statement and E&E articles for  
35  
36  
37 peer-reviewed publication.<sup>12</sup>  
38  
39  
40  
41  
42

#### 43 *Funding*

44  
45  
46 This protocol submission is funded by *the First Cut Research Competition* at McMaster University.  
47  
48

49  
50 Organizers of *the First Cut* had no involvement in planning the design of this study, the writing of this  
51  
52  
53 protocol manuscript, and will not be involved in the conduct of this study.  
54  
55  
56  
57  
58  
59  
60

*Updates & Monitoring*

The field of LLM-linked chatbot research is evolving, and it is paramount that the CHART Reporting Guidelines reflect the most modern advances in Chatbot Assessment Study research and LLM-linked technology. To address this need, the project lead and senior methodologist lead will actively survey news updates from both accessible and closed/proprietary chatbot models monthly. Beginning in 2025, the project lead will assess the need to initiate an updated scoping review annually if changes to the study aims, methodology, and/or quantity of published literature in this area is significant.

To inform the necessity of updates to the CHART reporting guidelines, both the project lead and senior methodologist lead will consider a combination of the updates in LLM-linked chatbot technology, as well as the study aims, methodology, and/or quantity of new Chatbot Assessment Studies.

For peer review only

## References

1. Thirunavukarasu AJ, Ting DSJ, Elangovan K, et al. Large language models in medicine. *Nat Med*. 2023;29:1930-1940. doi:10.1038/s41591-023-02448-8

2. Gholami S, Omar M. Do Generative Large Language Models need billions of parameters? *arXiv*. 2023;1-15. <http://arxiv.org/abs/2309.06589>
3. Krishna Vamsi G, Rasool A, Hajela G. Chatbot A Deep Neural Network Based Human to Machine Conversation Model, *IEEE*. 2023;1-7. <https://ieeexplore.ieee.org/document/9225395>
4. Cascella M, Montomoli J, Bellini V, et al. Evaluating the Feasibility of ChatGPT in Healthcare: An Analysis of Multiple Clinical and Research Scenarios. *J Med Syst*. 2023;47:33. doi:10.1007/s10916-023-01925-4
5. Ziegler DM, Stiennon N, Wu J, et al. Fine-Tuning Language Models from Human Preferences. *arXiv*. 2019;1-26. <http://arxiv.org/abs/1909.08593>
6. Bhirud N, Randive S, Tataale S, et al. A Literature Review On Chatbots In Healthcare Domain. *Int J Sci Technol Res*. 2019;8:225-232.
7. Sallam M. ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. *Healthcare*. 2023;11:1-20. doi:10.3390/healthcare11060887
8. Rudolph J, Tan S, Tan S. War of the chatbots: Bard, Bing Chat, ChatGPT, Ernie and beyond. The new AI gold rush and its impact on higher education. *JALT*. 2023;6:364-389. doi:10.37074/jalt.2023.6.1.23
9. Ayers JW, Poliak A, Dredze M, et al. Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. *JAMA Intern Med*. 2023;183:589-596. doi:10.1001/jamainternmed.2023.1838
10. Haver HL, Ambinder EB, Bahl M, et al. Appropriateness of Breast Cancer Prevention and Screening Recommendations Provided by ChatGPT. *Radiology*. 2023;307:e230424. doi:10.1148/radiol.230424
11. Rahsepar AA, Tavakoli N, Kim GHJ, et al. How AI Responds to Common Lung Cancer Questions: ChatGPT vs Google Bard. *Radiology*. 2023;307:e230922. doi:10.1148/radiol.230922
12. Moher D, Schulz KF, Simera I, et al. Guidance for developers of health research reporting guidelines. *PLoS Med*. 2010;7:e1000217. doi:10.1371/journal.pmed.1000217
13. Begg C, Cho M, Eastwood S, et al. Improving the Quality of Reporting of Randomized Controlled Trials. The CONSORT Statement. *JAMA*. 1996;276:637-9. doi:10.1001/jama.276.8.637
14. Moher D, Hopewell S, Schulz KF, et al. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *BMJ*. 2010;340:1-28. doi:10.1136/bmj.c869
15. Vasey B, Nagendran M, Campbell B, et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *Nat Med*. 2022;28:924-933. doi:10.1038/s41591-022-01772-9
16. Rivera SC, Liu X, Chan AW, et al. Guidelines for clinical trial protocols for interventions involving artificial intelligence: The SPIRIT-AI Extension. *The BMJ*. 2020;370:m3210. doi:10.1136/bmj.m3210
17. Liu X, Rivera SC, Moher D, et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: The CONSORT-AI Extension. *The BMJ*. 2020;370:m3164. doi:10.1136/bmj.m3164

- 1
  - 2
  - 3
  - 4
  - 5
  - 6
  - 7
  - 8
  - 9
  - 10
  - 11
  - 12
  - 13
  - 14
  - 15
  - 16
  - 17
  - 18
  - 19
  - 20
  - 21
  - 22
  - 23
  - 24
  - 25
  - 26
  - 27
  - 28
  - 29
  - 30
  - 31
  - 32
  - 33
  - 34
  - 35
  - 36
  - 37
  - 38
  - 39
  - 40
  - 41
  - 42
  - 43
  - 44
  - 45
  - 46
  - 47
  - 48
  - 49
  - 50
  - 51
  - 52
  - 53
  - 54
  - 55
  - 56
  - 57
  - 58
  - 59
  - 60
18. Peters MDJ, Marnie C, Tricco AC, et al. Updated methodological guidance for the conduct of scoping reviews. *JBI Evid Synth.* 2020;18:2119-2126. doi:10.11124/JBIES-20-00167
19. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *The BMJ.* 2021;372:n71. doi:10.1136/bmj.n71
20. Javed Ali M. ICMJE criteria for authorship: why the criticisms are not justified? *Graefes Arch Clin Exp Ophthalmol.* 2021;259:289-290. doi:10.1007/s00417-020-04825-2



**\*The CHART Collaborative.**

**Authors:** Bright Huo,<sup>1</sup> Tyler McKechnie,<sup>1,2</sup> David J Chartash,<sup>3,4</sup> Iain J Marshall,<sup>5</sup> David Moher,<sup>6</sup> Jeremy Ng,<sup>6,7</sup> Elizabeth Loder,<sup>8,9</sup> Timothy Feeney,<sup>8,10</sup> An-Wen Chan,<sup>11,12</sup> Michael Berkwits,<sup>13</sup> Annette Flanagan,<sup>13,14</sup> Stavros Antoniou,<sup>15</sup> Christine Laine,<sup>16, 17, 18</sup> Giovanni E Cacciamani,<sup>19, 20</sup> Gary S Collins,<sup>21</sup> Ashirbani Saha,<sup>22</sup> Piyush Mathur,<sup>23</sup> Alfonso Iorio,<sup>24,25</sup> Yung Lee,<sup>1,26</sup> Monica Ortenzi,<sup>27</sup> Julio Mayol,<sup>28</sup> Cynthia Lokker,<sup>24</sup> Thomas Agoritsas,<sup>24,29</sup> Per Olav Vandvik,<sup>30</sup> Farid Foroutan,<sup>24,31</sup> Hugo Campos,<sup>32</sup> Carolyn Canfield,<sup>33</sup> Karim Ramji,<sup>1</sup> Riaz Agha,<sup>34</sup> Hassaan Ahned,<sup>35</sup> Vanessa Boudreau,<sup>1</sup> Gordon Guyatt,<sup>24,36</sup>

**Affiliations:**

1. Division of General Surgery, Department of Surgery, McMaster University, Hamilton, Canada.
2. Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, Canada.
3. Khoury College of Computer Sciences, Northeastern University, Boston, USA. Section for Biomedical Informatics and Data Science, School of Medicine, Yale University, New Haven, Connecticut, 06511, United States of America
4. School of Medicine, University College Dublin - National University of Ireland, Dublin, County Dublin, Republic of Ireland.
5. School of Life Course and Population Sciences, King's College London, London, UK.
6. School of Epidemiology and Public Health, University of Ottawa, Ottawa, Canada.
7. Centre for Journalology, Ottawa Methods Centre, Ottawa Hospital Research Institute, Ottawa, Canada.
8. The BMJ, London, UK.
9. Division of Headache, Department of Neurology, Graham Headache Center, Brigham and

- 1  
2  
3  
4 Women's Hospital, Harvard Medical School, Boston, USA.
- 5 10. Gillings School of Global Public Health, The University of North Carolina, Chapel Hill, USA.
- 6 11. Phelan Senior Scientist, Women's College Research Institute and ICES, Toronto, Canada.
- 7 12. Department of Medicine, University of Toronto, Toronto, Canada.
- 8 13. JAMA and JAMA Network
- 9 14. Executive Managing Editor and Vice President
- 10 15. Department of General Surgery, Papageorgiou General Hospital, Thessaloniki, Greece.
- 11 16. Editor in Chief, Annals of Internal Medicine.
- 12 17. Senior VP, American College of Physicians.
- 13 18. Professor of Medicine, Sidney Kimmel Medical College, Thomas Jefferson University,
- 14 Philadelphia, USA.
- 15 19. USC Institute of Urology and Catherine and Joseph Aresty Department of Urology, Keck School
- 16 of Medicine, University of Southern California, Los Angeles, California.
- 17 20. AI Center at USC Urology, University of Southern California, Los Angeles, California.
- 18 21. Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology &
- 19 Musculoskeletal Sciences, University of Oxford, Oxford, UK.
- 20 22. Department of Oncology, McMaster University, Hamilton, Canada.
- 21 23. Department of General Anesthesiology, Anesthesiology Institute, Cleveland Clinic.
- 22 24. Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton,
- 23 Canada.
- 24 25. Michael Gent Chair in Healthcare Research
- 25 26. Harvard T.H. Chan School of Public Health, Harvard University, Boston, Massachusetts, USA.
- 26 27. Department of General Surgery, Università Politecnica delle Marche, Ancona, Italy.
- 27 28. Hospital Clinico San Carlos, IdISSC, Universidad Complutense de Madrid, Madrid, Spain.
- 28 29. Division of General Internal Medicine, Division of Clinical Epidemiology, University Hospitals
- 29 of Geneva, Geneva, Switzerland.
- 30 30. Department of Medicine, Lovisenberg Diaconal Hospital, Oslo, Norway.
- 31 31. Ted Rogers Computational Program (F.F., C.-P.S.F.), Peter Munk Cardiac Centre, University
- 32 Health Network, Toronto, ON, Canada.
- 33 32. University of California, Davis, USA.
- 34 33. Department of Family Practice, Faculty of Medicine, University of British Columbia, Vancouver,
- 35 Canada.
- 36 34. IJS Publishing Group, London, UK.
- 37 35. Phelix AI, Toronto, Canada.
- 38 36. Department of Medicine, McMaster University, Hamilton, Canada.
- 39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3  
4  
5 **Acknowledgements:** The study team would like to thank Byron Wallace for his expert input.  
6  
7

8 **Contributors:** Each author in the CHART collaborative contributed to the planning of the development of  
9 the CHART reporting guideline including the determination of its scope, study design, and the drafting of  
10 this protocol manuscript.  
11  
12  
13

14 **Funding:** This work was supported by *the First Cut* from the Department of Surgery at McMaster  
15 University (funding number not applicable). The organizers of *the First Cut* competition were not  
16 involved in this study at any stage including the conception, planning, or creation of this study protocol.  
17  
18  
19

20 **Competing interests:** None declared.  
21  
22

23 **Word Count:** 3,149  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

# BMJ Open

## Protocol for the Development of the Chatbot Assessment Reporting Tool (CHART) for Clinical Advice

Journal:	<i>BMJ Open</i>
Manuscript ID	bmjopen-2023-081155.R1
Article Type:	Protocol
Date Submitted by the Author:	16-Feb-2024
Complete List of Authors:	CHART Collaborative, The; McMaster University
<b>Primary Subject Heading</b>:	Research methods
Secondary Subject Heading:	Research methods, Ethics, Evidence based practice
Keywords:	MEDICAL ETHICS, Natural Language Processing, STATISTICS & RESEARCH METHODS

SCHOLARONE™  
Manuscripts

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60



# Protocol for the Development of the Chatbot Assessment Reporting Tool (CHART) for Clinical Advice

---

The CHART Collaborative\*

## The CHART Reporting Guideline Research Protocol

2024/02/15

1  
2  
3  
4 25

5 26

6  
7 278 28 **Corresponding author:**

9 29 Bright Huo, on behalf of The CHART Collaborative\*

10 30 237 Barton St E, Hamilton, ON L8L 2X2

11 31 E: brighthuo@dal.ca

12 32 T: +1 902 448 6836

13 33

14 34 **Keywords:** Medical ethics, statistics & research methods, natural language processing

15 35

16 36 **ABSTRACT**

17 37

18 38 **Introduction:**

19 39

20 40 Large language model (LLM)-linked chatbots are being increasingly applied in healthcare due to their

21 41

22 42 impressive functionality and public availability. Studies have assessed the ability of LLM-linked chatbots

23 43

24 44 to provide accurate clinical advice. However, the methods applied in these Chatbot Assessment Studies

25 45

26 46 are inconsistent due to the lack of reporting standards available, which obscures the interpretation of their

27 47

28 48 study findings. This protocol outlines the development of the Chatbot Assessment Reporting Tool

29 49

30 50 (CHART) reporting guideline.

31 51

32 52

33 53 **Methods and analysis:**

34 54

35 55 The development of the CHART reporting guideline will consist of three phases, led by the Steering

36 56

37 57 Committee. During phase one, the team will identify relevant reporting guidelines with artificial

38 58

39 59

40 60

1  
2  
3  
4 49 intelligence extensions that are published or in-development by searching preprint servers, protocol  
5  
6  
7 50 databases, and the Enhancing the Quality and Transparency of health research (EQUATOR) Network.  
8  
9  
10 51 During phase two, we will conduct a scoping review to identify studies that have addressed the  
11  
12  
13 52 performance of LLM-linked chatbots in summarizing evidence and providing clinical advice. The  
14  
15  
16 53 Steering Committee will identify methodology used in previous Chatbot Assessment Studies. Finally, the  
17  
18  
19 54 study team will use checklist items from prior reporting guidelines and findings from the scoping review  
20  
21  
22 55 to develop a draft reporting checklist. We will then perform a Delphi consensus and host two synchronous  
23  
24  
25 56 consensus meetings with an international, multidisciplinary group of stakeholders to refine reporting  
26  
27  
28 57 checklist items and develop a flow diagram.  
29  
30

31 58

**59 Ethics and dissemination:**

36  
37 60 We will publish the final CHART reporting guideline in peer-reviewed journals and will present findings  
38  
39  
40 61 at peer-reviewed meetings. Ethical approval is not applicable for the development of the CHART  
41  
42  
43 62 reporting guideline.  
44  
45

46 63

**64 Registration:**

51  
52  
53 65 This study protocol is pre-registered with Open Science Framework:  
54  
55

56 66 <https://doi.org/10.17605/OSF.IO/59E2Q>.  
57  
58  
59

67

**68 Strengths and limitations of this study:**

69 • This initiative will address a lack of reporting standards for Chatbot Assessment Studies and will  
70 provide a framework to increase the transparent conduct of these studies.

71 • We will apply rigorous methodology of the highest standards to develop the CHART reporting  
72 guideline.

73 • A diverse group of international, multidisciplinary stakeholders will inform the development of  
74 the CHART reporting checklist and flow diagram, with key input from experts in LLMs.

75 • This reporting guideline will be developed swiftly while acknowledging the dynamically  
76 evolving technology of LLM-linked chatbots.

77 • The CHART reporting guideline will apply specifically for studies assessing the ability of LLM-  
78 linked chatbots to summarize evidence and provide clinical advice. It will not apply to their use in  
79 other settings.

80

**81 INTRODUCTION**

82 Novel chatbots have been integrating Large Language Models (LLMs), which are a popular technology in  
83 the field of natural language processing (NLP) [1]. LLMs are large neural networks often comprised of  
84 hundreds of billions of parameters, which impact the model's input, size and shape, and output [2]. LLMs



1  
2  
3  
4 85 are typically used to conditionally predict the next words in a sequence of text, given corresponding  
5  
6 86 prompts (Table 1) [3]. LLMs can be trained on a collection of massive amounts of raw data from online  
7  
8  
9 87 text sources including books, articles, websites, and more [1,4]. Coupled with reinforcement learning  
10  
11  
12 88 from human feedback [5]. LLMs exhibit striking text generation capabilities, producing outputs that are  
13  
14  
15 89 often indistinguishable from human language [6,7]. There has been a gold-rush movement of chatbots  
16  
17  
18 90 linked to LLMs, with recent releases including ChatGPT, Bing Chat, Google Bard, Med-PaLM, and many  
19  
20  
21  
22 91 more underway [8].  
23  
24  
25 92  
26  
27  
28 93 Given their wide accessibility and ability to provide answers to lay prompts [8], investigators have begun  
29  
30  
31 94 to assess LLM-linked chatbots as a potential source of health advice for both patients and clinicians [9–  
32  
33  
34 95 11]. We refer to these studies as Chatbot Assessment Studies, and they evaluate the performance of LLM-  
35  
36  
37 96 linked chatbots in summarizing health evidence and providing clinical advice. These studies represent a  
38  
39  
40 97 new genre of medical research, but the methodology and framing of results reported in these studies are  
41  
42  
43 98 highly variable. Inconsistent and incomplete reporting limits readers' ability to judge the methodology  
44  
45  
46 99 and results of these studies, complicating their interpretation [12]. A need exists to assess the rigour of  
47  
48  
49 100 their assessments [8], but currently there are no standardized reporting tools for Chatbot Assessment  
50  
51  
52  
53 101 Studies.  
54  
55  
56 102

1  
2  
3  
4 103 Instruments have been created to address issues of suboptimal reporting and raise the standard of research  
5  
6  
7 104 quality, such as the Consolidated Standards of Reporting Trials (CONSORT) statement [13,14]. Such  
8  
9  
10 105 reporting guidelines provide a checklist and a flow diagram for a given study type. Since their  
11  
12  
13 106 development, extensions to reporting guidelines have been created to facilitate the integration of artificial  
14  
15  
16 107 intelligence [15–17]. However, LLM-linked chatbots and their accompanying applications have only  
17  
18  
19 108 recently emerged and are not captured by these reporting guidelines. This protocol outlines the  
20  
21  
22 109 development of a novel reporting checklist, the Chatbot Assessment Reporting Tool (CHART) to  
23  
24  
25 110 improve the reporting standards of Chatbot Assessment Studies.  
26  
27

111

## 112 **Key Terminology**

113 Table 1 lists key terms included in this work.  
114

115 Table 1. Glossary.

Term	Definition
Artificial Intelligence (AI)	The science of developing computer systems that can perform complex tasks approximating human cognitive performance.
Natural Language Processing (NLP)	A branch of information science that seeks to enable computers to interpret and manipulate human text.
Large Language Model (LLM)	A type of NLP model comprising large neural networks trained over large amounts of text, usually to produce an output of continuations of text from corresponding prompts, known as next word prediction. <sup>+</sup>
Multimodal LLM	LLMs with the capacity to integrate input from various data types, including text, speech,

	and/or visual sources.
Next word prediction	The natural language processing task of predicting the next word in a sequence of text given context and model parameters.
Parameter	A <i>parameter</i> within an artificial intelligence algorithm is a variable that is tuned iteratively/automatically to optimize the intended outcome of the algorithm. Parameters may be at the model level to optimize tuning (hyperparameters) or "weights" within the model linking layer to layer (parameters)
LLM-Linked Chatbot	A program that permits users to interact with an algorithm (such as an LLM) designed to respond to user prompts.
Chatbot Assessment Study	Any research study assessing the performance of chatbots in summarizing health evidence and/or providing clinical advice.
Chat Instance	An interface in a computing device through which communication takes place between a chatbot and its user through text with only one prompt.
Chat Session	An interface in a computing device through which communication takes place between a chatbot and its user through text with more than one prompt.
Query	The act of communicating with a LLM by inputting a prompt into the chatbot which might be a question, comment, or phrase to elicit specific desired outputs from an LLM. For example, one might input a prompt asking the LLM to summarize the evidence supporting the use of a given intervention.
Check query	Following formal query completion and performance evaluation, the act of repeating the initial query to ensure that chatbot outputs are consistent in summarizing the same evidence and providing the same clinical advice.
Prompt	Text input by a user into the chatbot for the purpose of communicating with the LLM.
Prompt Engineering	An iterative testing phase where various pieces of text are inputted into a chatbot to achieve an output, informing the development of study prompts.
Delphi study	A structured research method applied to answer a research question through the establishment of consensus across respondents.

116 +Generally speaking, "next word" prediction is one basic "pre-training" objective, but LLMs often undergo a subsequent round  
 117 of "supervision" in which they are guided by human feedback.

118 -Chatbots are not necessarily built atop LLMs, but the modern tools that have captured public imagination are.

119

## 120 **METHODS & ANALYSIS**

### 121 **Study Overview & Objectives**

1  
2  
3  
4 122 This study consists of three phases to address the following objectives:  
5

6 123 1. To identify checklist items used in previous reporting guidelines and identify related reporting  
7  
8  
9 124 standards for studies assessing the use of artificial intelligence in healthcare.  
10

11  
12 125 2. To perform a scoping review that will identify and characterize studies that have addressed the  
13  
14  
15 126 performance of LLMs in summarizing evidence and providing clinical advice. Specifically, the  
16  
17  
18 127 review will identify how authors evaluate chatbot performance in summarizing health evidence  
19  
20  
21  
22 128 and providing clinical advice.  
23

24  
25 129 3. Informed by the scoping review and a review of prior checklists, to develop an evidence-  
26  
27  
28 130 informed, expert-derived reporting guideline comprised of a checklist and flow diagram for  
29  
30  
31 131 studies assessing chatbot performance in summarizing health evidence and providing clinical  
32  
33  
34 132 advice.  
35

36  
37 133  
38  
39  
40 134 A Steering Committee will lead all key study initiatives. This group will include the following members:  
41

42  
43 135 the project lead, the senior methodologist lead, an expert in chatbot assessment studies, a reporting  
44

45  
46 136 checklist developer, and a journal editor. The group's responsibilities will be to guide the initiatives  
47

48  
49 137 involved in the development of the CHART checklist. They will lead the review of relevant reporting  
50

51  
52 138 checklists (phase one), the completion of the scoping review (phase two), and the development of the  
53

54  
55 139 reporting guideline (phase three). Table 1 presents a glossary of key terms used in this work. Figure 1  
56  
57  
58  
59  
60

140 demonstrates the timeline for the development of the CHART reporting guideline.

141

142 Figure 1. Timeline for the Development of the CHART Reporting Guideline.

143

144 This reporting guideline will emphasize transparent reporting standards for studies evaluating the

145 performance of LLMs when providing clinical advice to patients and clinicians. It will apply to LLM-

146 linked chatbots, but also LLMs more broadly. It will also apply to studies using both traditional and

147 multimodal LLMs.

148

## 149 PHASE ONE

150

151 **Objective:** to identify checklist items used in previous reporting guidelines and identify related reporting

152 standards for studies assessing the ability of LLMs to provide clinical advice.

153

### 154 *Identification of Existing Reporting Guidelines*

155 To identify relevant health research reporting guidelines to inform the development of our reporting

156 guideline and checklist, the study team will search the EQUATOR network and identify reporting

157 guidelines published prior to October 2023 that meet our inclusion criteria:

- 1  
2  
3  
4 158     • Studies presenting primary data on the use of chatbots in any specialty in medicine.  
5  
6  
7 159     • Studies applying chatbots to summarize evidence and provide clinical advice.  
8  
9  
10 160     • Studies applying chatbots to answer one or more clinical question(s).  
11  
12  
13 161     • Any studies applying chatbots as an intervention, with or without the use of a comparator.  
14  
15

16 162  
17  
18  
19 163 To achieve this, the study team will use the “search for reporting guidelines” feature and toggle through  
20  
21  
22 164 each study type. We will review all reporting guidelines in each study type for comprehensiveness. We  
23  
24  
25 165 will review references from relevant reporting guidelines and related citations listed on PubMed for  
26  
27  
28 166 retrieved articles. To identify protocols of reporting guidelines, we will search Open Science Framework  
29  
30  
31 167 as well as applicable results obtained from our scoping review. To identify ongoing or completed work  
32  
33  
34 168 not yet published in peer-reviewed sources, we will search Open Science Framework & MedRxiv.  
35  
36

37 169  
38  
39  
40 170 Reporting guidelines obtained from the search from phase one will inform the development of items for a  
41  
42  
43 171 preliminary draft version of the checklist.  
44  
45

46 172  
47  
48

49 173 **PHASE TWO**  
50

51  
52 174  
53  
54

55  
56 175 **Objective:** to perform a scoping review that will identify and characterize studies that have addressed the  
57  
58  
59  
60

1  
2  
3  
4 176 performance of LLMs in summarizing evidence and providing clinical advice. Specifically, the review  
5  
6  
7 177 will identify how authors evaluate chatbot performance in summarizing health evidence and providing  
8  
9  
10 178 clinical advice.  
11  
12  
13 179  
14  
15  
16 180 For the scoping review, the project lead will recruit a team that will include two other members that have  
17  
18  
19 181 previous experience with performing systematic reviews and scoping reviews as well as the senior  
20  
21  
22 182 methodological lead. The scoping review team will identify articles assessing the performance of chatbots  
23  
24  
25 183 when applied in healthcare. A separate protocol presents our search strategy, inclusion criteria, exclusion  
26  
27  
28 184 criteria, and other details related to the scoping review, which is under consideration for publication. Its  
29  
30  
31 185 development will be aligned with methodology guidance from the JBI Scoping Review Methodology  
32  
33  
34 186 Group [18].  
35  
36  
37 187  
38  
39  
40 188 In brief, the scoping review team will conduct a literature search using MEDLINE via Ovid, EMBASE  
41  
42  
43 189 via Elsevier, Scopus via Elsevier, and Web of Science to capture relevant studies published prior to  
44  
45  
46 190 October 2023. The team will identify studies that evaluate the performance of LLM-linked chatbots when  
47  
48  
49 191 providing clinical advice. We will only consider primary data. The team will complete two rounds of  
50  
51  
52 192 screening by title and abstract and full-text to identify articles of interest. Next, we will perform manual  
53  
54  
55 193 forward and backward citation searching. The team will then perform data extraction to identify key items  
56  
57  
58  
59  
60

1  
2  
3  
4 194 used in the reporting of these studies. The following variables will be extracted: clinical aims (health  
5  
6  
7 195 prevention, screening, differential diagnosis, diagnosis, treatment), prompt development (use of specific  
8  
9  
10 196 sources, engineering/testing phase, standardized prompts, prompt structure, prompt inclusion in-text)  
11  
12  
13 197 LLM, LLM model version, LLM characteristics (temperature, token length, fine-tuning availability,  
14  
15  
16 198 penalties, add-on availability, layers), date accessed/trained, language, location of query, use of chat  
17  
18  
19 199 windows/sessions, performance definition (objective use of literature such as guideline or systematic  
20  
21  
22 200 review versus subjective evaluation using experts), and whether a statement or discussion on ethics,  
23  
24  
25 201 regulation, or patient safety is included.  
26  
27  
28 202  
29  
30 203 We will report findings using descriptive statistics for quantitative data and present results graphically in  
31  
32  
33 204 diagrammatic form. A narrative summary will accompany the graphical results. The final report will  
34  
35  
36 205 adhere with reporting standards for the Preferred Reporting Items for Systematic Review and Meta-  
37  
38  
39 206 Analysis Extension for Scoping Reviews (PRISMA-ScR) [19].  
40  
41  
42 207  
43  
44

### 45 208 **PHASE THREE**

46  
47  
48 209  
49  
50  
51 210 **Objective:** informed by the scoping review and a review of prior checklists, to develop an evidence-  
52  
53  
54 211 informed, expert-derived reporting guideline comprised of a checklist and flow diagram for studies  
55  
56  
57  
58  
59  
60



1  
2  
3  
4 212 assessing chatbot performance in summarizing health evidence and providing clinical advice.  
5  
6  
7 213

8  
9  
10 214 *Advisory Committee & Delphi*

11  
12 215 An Advisory Committee will comprise epidemiologists, research methodologists, NLP researchers,  
13  
14  
15

16 216 journal editors, chatbot researchers, ethicists, regulatory experts, policy experts, and patient partners. The  
17  
18

19 217 Steering Committee will identify additional committee members by querying SCImago Journal Country  
20  
21

22 218 Rank (SJR) portal ([www.scimagojr.com](http://www.scimagojr.com)) to obtain a list of the top 10 journals in each specialty in  
23  
24

25 219 medicine. Using this list of journals, the Committee will query Web of Science to obtain a diverse list of  
26  
27

28 220 researchers in medicine including general research methodologists and chatbot researchers. Patient  
29  
30

31 221 partners will be identified through both public and internal calls through affiliate journals, as well as  
32  
33

34 222 through the snowballing method via our panel, including patient partner members. We will send an  
35  
36

37 223 invitation email to our final list of contacts to invite them to join the Advisory Committee.  
38  
39

40 224

41  
42  
43 225 The Steering Committee will hold a synchronous virtual meeting open to all Advisory Committee  
44  
45

46 226 members as an introduction to the project, as well as their role. Through a series of questionnaires shared  
47  
48

49 227 through an online platform, the team will apply a Delphi consensus. The Steering Committee will develop  
50  
51

52 228 a draft checklist informed by the scoping review and review of existing reporting guidelines. They will  
53  
54

55 229 circulate the draft checklist to the Advisory Committee for a first round of voting. During this round,  
56  
57  
58  
59  
60

## The CHART Reporting Guideline Research Protocol

2024/02/15

1  
2  
3  
4 230 Advisory Committee members will select one of the following options for each checklist item: “include,  
5  
6  
7 231 maybe include, uncertain, maybe exclude, exclude.” There will be an additional option for Advisory  
8  
9  
10 232 Committee members to once more add checklist items. The Steering Committee will then revise the  
11  
12  
13 233 checklist using comments from the first round. The team will re-circulate the updated draft checklist for a  
14  
15  
16 234 second round of voting, as above.

17  
18  
19 235  
20  
21  
22 236 The Steering Committee will revise the checklist following the second round and present these items to  
23  
24  
25 237 the expert panel. In preparation for the next phase, the steering committee will meet with an ethicist and  
26  
27  
28 238 regulatory expert to review draft checklist items from the Delphi process to revise or add key principles  
29  
30  
31 239 for ethics & safety for discussion during the consensus meeting.

32  
33  
34 24035  
36  
37 241 *Expert Panel*

38  
39  
40 242 We will create an international, multidisciplinary panel as per Moher and colleagues [12]. Participants  
41  
42  
43 243 will be purposefully selected to reflect a balanced representation of relevant stakeholders including  
44  
45  
46 244 statisticians, research methodologists, reporting checklist developers, NLP researchers, journal editors,  
47  
48  
49 245 chatbot researchers, ethicists, regulatory experts, and two patient partners. In advance of the Consensus  
50  
51  
52 246 Meetings, the Steering Committee will prompt panelists to share their conflicts of interest. Though we  
53  
54  
55 247 find it difficult to imagine circumstances that would lead to important conflicts, we will stay alert to

1  
2  
3  
4 248 unanticipated conflicts. Should these arise, we will consider any panel member with significant conflicts  
5  
6  
7 249 as consultant who will not vote on the final checklist. Prior to the first of two Synchronous Consensus  
8  
9  
10 250 Meetings, the Steering Committee will share the candidate checklist items with the Expert Panel which  
11  
12  
13 251 will have been revised following two Delphi rounds with the Advisory Committee, informed by findings  
14  
15  
16 252 from the scoping review.  
17  
18  
19 253  
20  
21  
22 254 Additionally, the Steering Committee will construct a flow diagram prior to the Consensus Meetings  
23  
24  
25 255 based on the candidate checklist items. The purpose of the flow diagram is to provide an overview to  
26  
27  
28 256 guide authors in clearly reporting sequential stages of their study. The Steering Committee will also share  
29  
30  
31 257 this flow diagram with the panel prior to the Consensus Meetings.  
32  
33  
34 258  
35  
36  
37 259 In preparation for the synchronous consensus meetings, the Steering Committee will share relevant  
38  
39  
40 260 materials with the panel such as the meeting agenda, participant list, and the completed scoping review  
41  
42  
43 261 highlighting the content and extent of reporting of the content area. The Committee will also circulate the  
44  
45  
46 262 draft checklist that emerged from the Delphi process to the Expert Panel through an electronic survey in  
47  
48  
49 263 advance of the meeting. The steering group has pre-specified an 80% threshold for inclusion to  
50  
51  
52 264 demonstrate majority consensus based on prior work [17]. We will group items with  $\geq 80\%$  consensus  
53  
54  
55 265 with the selection of “include” or “maybe include” together, posing to the panelists: “These items have  
56  
57  
58  
59  
60

1  
2  
3  
4 266 been recommended for inclusion in our checklist. Do you agree or disagree?" Panelists will have the  
5  
6  
7 267 option of yes - include, no - exclude, unsure, and an additional option for comments.  
8

9  
10 268  
11  
12 269 We will also group items with  $\geq 80\%$  consensus for items with the selection of "exclude" or "maybe  
13  
14  
15  
16 270 exclude," posing to the panelists: "These items have been recommended for exclusion in our checklist.  
17  
18  
19 271 Do you agree or disagree?" Panelists will have the option of yes - exclude, no - include, and an additional  
20  
21  
22 272 option for comments. Items without 80% consensus will be gathered and panel members will indicate  
23  
24  
25 273 "include, maybe include, uncertain, maybe exclude, exclude." There will also be an additional option for  
26  
27  
28 274 each question to suggest additional checklist items. We will collate the results of this survey in  
29  
30  
31 275 preparation for the Consensus Meetings.  
32

33  
34 276  
35  
36  
37 277  
38  
39  
40 278 *Synchronous Consensus Meetings*  
41  
42  
43 279 The project lead will organize two Synchronous Consensus Meetings that will be held over a video  
44  
45  
46 280 conferencing platform. The Steering Committee will encourage panelists to attend both meetings, with the  
47  
48  
49 281 expectation that panelists must attend one meeting, at minimum. The steering committee will circulate an  
50  
51  
52 282 online scheduling survey in advance to control the number of participants in attendance, while also  
53  
54  
55 283 selecting dates that optimize the attendance of panel members. As we will hold these meetings virtually,  
56  
57  
58  
59  
60

284 no meeting will be longer than four hours in duration to mitigate burnout and encourage participation.

285 The duration of both meetings will be eight hours in total. A contingency plan is set to pre-emptively

286 arrange and hold a third meeting of two to four hours should additional time be needed following the

287 eight hours of consensus meetings.

288

289 During checklist item discussion, we will put forth any items rated as “no-exclude” by panelists during

290 the pre-consensus meeting survey for exclusion from the checklist. We will then discuss any items

291 without consensus or rated as “uncertain” with  $\geq 80\%$  consensus after the second Delphi round. Finally,

292 we will offer items rated as “yes-include” to the panel for inclusion in the checklist. During the discussion

293 for all checklist items, the meeting chair will present the following for each checklist item:

294 • Previous use in a Chatbot Assessment Study

295 • Rationale for inclusion

296

297 All voting will take place virtually and anonymously over the video conferencing platform. A working

298 CHART checklist will emerge from the Synchronous Consensus Meetings. The panel will use this

299 working checklist to revise the draft CHART flow diagram during the Synchronous Consensus Meeting.

300

301 Expert panel members who are unable to join will be able to review recordings of the meetings. The

## The CHART Reporting Guideline Research Protocol

2024/02/15

1  
2  
3  
4 302 project lead will record the meeting(s), and they will share both the recording and a summary of checklist  
5  
6  
7 303 item decisions and rationale with absent panel members.  
8

9  
10 304  
11  
12 305 Following the meetings, the Steering Committee will circulate the working CHART checklist and flow  
13  
14  
15 306 diagram in the form of a survey reflecting checklist item decisions. This working checklist will outline a  
16  
17  
18 307 final list of items for inclusion. Panellists will have the opportunity to provide any final comments, which  
19  
20  
21  
22 308 the Steering Committee will use to derive a preliminary CHART checklist. The preliminary checklist will  
23  
24  
25 309 also be shared with the public for open comment on the EQUATOR website, while links to the checklist  
26  
27  
28 310 will be shared on the website of affiliate journals of editors involved in the development of the CHART  
29  
30  
31 311 reporting guideline.  
32

33  
34 312  
35  
36  
37 313 Prior to pilot testing, the study team will share the preliminary checklist following the consensus meetings  
38  
39  
40 314 with patient partners identified a priori through snowballing and journal contacts to ensure that themes of  
41  
42  
43 315 patient access and safety are sufficiently addressed.  
44

45  
46 316  
47  
48  
49 317 *Pilot Testing*  
50  
51  
52 318 The Steering Committee will pilot the preliminary CHART checklist and flow diagram with researchers  
53  
54  
55 319 that have published Chatbot Assessment Studies and will identify authors by the included studies in the  
56  
57  
58

1  
2  
3  
4 320 scoping review. The Steering Committee will conduct pilot testing via an iterative process. Groups of five  
5  
6  
7 321 authors will provide feedback in each round until saturation is achieved, with a minimum of ten authors  
8  
9  
10 322 over two rounds of pilot testing. Authors will not evaluate their own studies but will use the checklist to  
11  
12  
13 323 assess Chatbot Assessment Studies published by other authors. During synchronous sessions, we will ask  
14  
15  
16 324 authors to assess Chatbot Assessment Studies using the preliminary CHART checklist and flow diagram  
17  
18  
19 325 via think-aloud instrument testing. Authors will provide practical feedback regarding the development of  
20  
21  
22 326 these studies in the context of checklist items. They will also provide feedback regarding the practical  
23  
24  
25 327 application of the preliminary CHART checklist with respect to the length and content of the checklist.  
26  
27

28 328  
29  
30  
31 329 The Steering Committee will use the comments from Chatbot Assessment Study researchers to derive a  
32  
33  
34 330 final version of the CHART checklist and flow diagram.  
35  
36

37 331  
38  
39  
40 332 *Report Generation*  
41

42  
43 333 With the final CHART checklist and flow diagram, the Steering Committee will prepare a Statement  
44  
45  
46 334 document for submission for peer-reviewed conference presentation and publication. All panel members  
47  
48  
49 335 will have the chance to review the draft manuscript, and all members of the research team satisfying the  
50  
51  
52 336 International Committee of Medical Journal Editors (ICJME) criteria will join the group authorship.[20]  
53  
54  
55 337 The Statement article will consist of the checklist and flow diagram. It will include the rationale for  
56  
57  
58  
59  
60

1  
2  
3  
4 338 developing the CHART guideline and an overview of its development, including a brief description of the  
5  
6  
7 339 meeting and participants involved.  
8

9  
10 340  
11  
12  
13 341 Separately, the Steering Committee will prepare a detailed explanation and elaboration paper (E&E). This  
14  
15  
16 342 paper will provide more detail for the inclusion of items in the final CHART checklist. For each checklist  
17  
18  
19 343 item, the E&E report will include three parts: 1) an explanation of the rationale supporting the checklist  
20  
21  
22 344 item, as well as reference to any supporting evidence for its inclusion 2) essential elements of the study  
23  
24  
25 345 that must be described to appropriately satisfy each checklist item 3) additional elements of the study  
26  
27  
28 346 which may be considered by authors depending on the context. Both the Statement and E&E articles will  
29  
30  
31 347 be written in collaboration with the multidisciplinary panel.  
32

33  
34 348  
35  
36  
37 349 As per Moher and colleagues, we will simultaneously submit both the Statement and E&E articles for  
38  
39  
40 350 peer-reviewed publication [12].  
41

42  
43 351  
44  
45  
46 352 *Patient & Public Involvement*  
47  
48  
49 353 Patients will be involved in the development of the CHART reporting guideline through participation in  
50  
51  
52 354 the Delphi process, as outlined above. Two patients will also be involved in the revision of the reporting  
53  
54  
55 355 guideline including the checklist, flow diagram, and resulting reports as panel members.  
56  
57  
58



1  
2  
3  
4 3565  
6  
7 357 *Funding*8  
9  
10 358 This protocol submission is funded by *the First Cut Research Competition* at McMaster University.11  
12  
13 359 Organizers of *the First Cut* had no involvement in planning the design of this study, the writing of this14  
15  
16 360 protocol manuscript, and will not be involved in the conduct of this study.17  
18  
19 36120  
21  
22 362 *Updates & Monitoring*23  
24  
25 363 The field of LLM-linked chatbot research is evolving, and it is paramount that the CHART Reporting26  
27  
28 364 Guidelines reflect the most modern advances in Chatbot Assessment Study research and LLM-linked29  
30  
31 365 technology. To address this need, the project lead and senior methodologist lead will actively survey news32  
33  
34 366 updates from both accessible and closed/proprietary chatbot models monthly. Beginning in 2025, the35  
36  
37 367 project lead will assess the need to initiate an updated scoping review annually if changes to the study38  
39  
40 368 aims, methodology, and/or quantity of published literature in this area is significant.41  
42  
43 36944  
45  
46 370 To inform the necessity of updates to the CHART reporting guidelines, both the project lead and senior47  
48  
49 371 methodologist lead will consider a combination of the updates in LLM-linked chatbot technology, as well50  
51  
52 372 as the study aims, methodology, and/or quantity of new Chatbot Assessment Studies.53  
54  
55 373

374 *Ethics*

375 This study was submitted to the Hamilton Integrated Research Ethics Board (HiREB). It was deemed that  
376 HiREB review and approval was not required. This study will adhere to key principles. All work will  
377 adhere to the World Medical Association Declaration of Helsinki Ethical Principles for Medical Research  
378 Involving Human Subjects [21]. Furthermore, all checklist items for future studies involving the use of  
379 LLMs for clinical advice will be reviewed in the context of these ethical principles [21]. The involvement  
380 of ethicists and regulatory experts in health technology will aid the steering committee and panel in  
381 considering these key principles, including accessibility and patient safety.

382  
383 Limitations

384 This study has limitations. The reporting checklist will be applicable for the most current, conventional  
385 LLMs at the time of publication due to the dynamic pace at which this field is evolving. To address this,  
386 the steering committee will assess the need to update the checklist on an annual basis, driven by the junior  
387 primary investigator.

388  
389  
390 **Acknowledgements:** The study team would like to thank Byron Wallace for his expert input.

391  
392 **Contributors:** Each author in the CHART collaborative contributed to the planning of the development of  
393 the CHART reporting guideline including the determination of its scope, study design, and the drafting of  
394 this protocol manuscript.

395

1  
2  
3  
4 396 **Funding:** This work was supported by *the First Cut* from the Department of Surgery at McMaster  
5 397 University (funding number not applicable). The organizers of *the First Cut* competition were not  
6 398 involved in this study at any stage including the conception, planning, or creation of this study protocol.  
7 399

8 399  
9 400 **Competing interests:** None declared.  
10 401

11 401  
12 402 **Word Count:** 3,752  
13 403

14 403  
15 404  
16 404  
17 405  
18 405  
19 406  
20 407  
21 407  
22 408  
23 408  
24 409  
25 410  
26 410  
27 411  
28 411  
29 412  
30 412  
31 413  
32 413  
33 414  
34 414  
35 415  
36 415  
37 416  
38 416  
39 417  
40 417  
41 418  
42 418  
43 419  
44 419  
45 420  
46 420  
47 421  
48 421  
49 422  
50 422  
51 423  
52 423  
53 424  
54 424  
55 425  
56 425  
57 426  
58 426  
59 427  
60 427  
428  
429  
430

431 **References**

- 432 1. Thirunavukarasu AJ, Ting DSJ, Elangovan K, et al. Large language models in medicine. *Nat*  
433 *Med*. 2023;29:1930-1940. doi:10.1038/s41591-023-02448-8
- 434 2. Gholami S, Omar M. Do Generative Large Language Models need billions of parameters?  
435 *arXiv*. 2023:1-15. <http://arxiv.org/abs/2309.06589>
- 436 3. Krishna Vamsi G, Rasool A, Hajela G. Chatbot A Deep Neural Network Based Human to  
437 Machine Conversation Model, *IEEE*. 2023;1-7.  
438 <https://ieeexplore.ieee.org/document/9225395>
- 439 4. Cascella M, Montomoli J, Bellini V, et al. Evaluating the Feasibility of ChatGPT in  
440 Healthcare: An Analysis of Multiple Clinical and Research Scenarios. *J Med Syst*.  
441 2023;47:33. doi:10.1007/s10916-023-01925-4
- 442 5. Ziegler DM, Stiennon N, Wu J, et al. Fine-Tuning Language Models from Human  
443 Preferences. *arXiv*. 2019;1-26. <http://arxiv.org/abs/1909.08593>
- 444 6. Bhirud N, Randive S, Tataale S, et al. A Literature Review On Chatbots In Healthcare  
445 Domain. *Int J Sci Technol Res*. 2019;8:225-232.
- 446 7. Sallam M. ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic  
447 Review on the Promising Perspectives and Valid Concerns. *Healthcare*. 2023;11:1-20.  
448 doi:10.3390/healthcare11060887
- 449 8. Rudolph J, Tan S, Tan S. War of the chatbots: Bard, Bing Chat, ChatGPT, Ernie and  
450 beyond. The new AI gold rush and its impact on higher education. *JALT*. 2023;6:364-389.  
451 doi:10.37074/jalt.2023.6.1.23
- 452 9. Ayers JW, Poliak A, Dredze M, et al. Comparing Physician and Artificial Intelligence  
453 Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. *JAMA*  
454 *Intern Med*. 2023;183:589-596. doi:10.1001/jamainternmed.2023.1838
- 455 10. Haver HL, Ambinder EB, Bahl M, et al. Appropriateness of Breast Cancer Prevention and  
456 Screening Recommendations Provided by ChatGPT. *Radiology*. 2023;307:e230424.  
457 doi:10.1148/radiol.230424
- 458 11. Rahsepar AA, Tavakoli N, Kim GHJ, et al. How AI Responds to Common Lung Cancer  
459 Questions: ChatGPT vs Google Bard. *Radiology*. 2023;307:e230922.  
460 doi:10.1148/radiol.230922
- 461 12. Moher D, Schulz KF, Simera I, et al. Guidance for developers of health research reporting  
462 guidelines. *PLoS Med*. 2010;7:e1000217. doi:10.1371/journal.pmed.1000217
- 463 13. Begg C, Cho M, Eastwood S, et al. Improving the Quality of Reporting of Randomized  
464 Controlled Trials. The CONSORT Statement. *JAMA*. 1996;276:637-9.  
465 doi:10.1001/jama.276.8.637
- 466 14. Moher D, Hopewell S, Schulz KF, et al. CONSORT 2010 explanation and elaboration:  
467 updated guidelines for reporting parallel group randomised trials. *BMJ*. 2010;340:1-28.  
468 doi:10.1136/bmj.c869
- 469 15. Vasey B, Nagendran M, Campbell B, et al. Reporting guideline for the early-stage clinical  
470 evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *Nat*  
471 *Med*. 2022;28:924-933. doi:10.1038/s41591-022-01772-9

- 1  
2  
3 472 16. Rivera SC, Liu X, Chan AW, et al. Guidelines for clinical trial protocols for interventions  
4 473 involving artificial intelligence: The SPIRIT-AI Extension. *The BMJ*. 2020;370:m3210.  
5 474 doi:10.1136/bmj.m3210  
6  
7 475 17. Liu X, Rivera SC, Moher D, et al. Reporting guidelines for clinical trial reports for  
8 476 interventions involving artificial intelligence: The CONSORT-AI Extension. *The BMJ*.  
9 477 2020;370:m3164. doi:10.1136/bmj.m3164  
10  
11 478 18. Peters MDJ, Marnie C, Tricco AC, et al. Updated methodological guidance for the conduct  
12 479 of scoping reviews. *JBI Evid Synth*. 2020;18:2119-2126. doi:10.11124/JBIES-20-00167  
13 480 19. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: An updated  
14 481 guideline for reporting systematic reviews. *The BMJ*. 2021;372:n71. doi:10.1136/bmj.n71  
15 482 20. Javed Ali M. ICMJE criteria for authorship: why the criticisms are not justified? *Graefes*  
16 483 *Arch Clin Exp Ophthalmol*. 2021;259:289-290. doi:10.1007/s00417-020-04825-2  
17 484 21. World Medical Association declaration of Helsinki: Ethical principles for medical  
18 485 research involving human subjects. *JAMA*. 2013;310(20):2191-2194.  
19 486 doi:10.1001/jama.2013.281053  
20  
21  
22 487  
23 488  
24  
25 489  
26 490  
27  
28 491  
29 492  
30  
31 493  
32 494  
33  
34 495  
35 496  
36  
37 497  
38 498  
39  
40 499  
41 500  
42  
43 501  
44  
45 502  
46 503  
47  
48 504  
49 505  
50  
51 506  
52 507  
53  
54 508  
55 509  
56  
57  
58  
59  
60

1  
2  
3  
4 510  
5 511  
6 512  
7  
8 513  
9  
10 514  
11 515  
12  
13 516  
14 517  
15  
16 518  
17 519  
18  
19 520  
20 521  
21  
22 522  
23  
24 523

25 524 **\*The CHART Collaborative.**

26 525  
27  
28 526 **Authors:** Bright Huo,<sup>1</sup> Tyler McKechnie,<sup>1,2</sup> David Chartash,<sup>3,4</sup> Iain J Marshall,<sup>5</sup> David Moher,<sup>6,7</sup> Jeremy Y  
29 527 Ng,<sup>7</sup> Elizabeth Loder,<sup>8,9</sup> Timothy Feeney,<sup>8,10</sup> An-Wen Chan,<sup>11, 12</sup> Michael Berkwits,<sup>13</sup> Annette  
30 528 Flanagan,<sup>13,14</sup> Stavros A Antoniou,<sup>15</sup> Christine Laine,<sup>16, 17, 18</sup> Giovanni E Cacciamani,<sup>19, 20</sup> Gary S Collins,<sup>21</sup>  
31 529 Ashirbani Saha,<sup>22</sup> Piyush Mathur,<sup>23</sup> Alfonso Iorio,<sup>2,24</sup> Yung Lee,<sup>1,25</sup> Diana Samuel,<sup>26</sup> Helen Frankish,<sup>27</sup>  
32 530 Monica Ortenzi,<sup>28</sup> Julio Mayol,<sup>29</sup> Cynthia Lokker,<sup>2</sup> Thomas Agoritsas,<sup>2,30</sup> Per Olav Vandvik,<sup>31</sup> Farid  
33 531 Foroutan,<sup>2,32</sup> Joerg J. Meerpohl,<sup>33,34</sup> Hugo Campos,<sup>35</sup> Carolyn Canfield,<sup>36</sup> Xufei Luo,<sup>37</sup> Yaolong Chen,<sup>37</sup>  
34 532 Hugh Harvey,<sup>38</sup> Stacy Loeb,<sup>39</sup> Riaz Agha,<sup>40</sup> Karim Ramji,<sup>1,41</sup> Hassaan Ahned,<sup>41</sup> Vanessa Boudreau,<sup>1</sup>  
35 533 Gordon Guyatt.<sup>2,42</sup>

40 534  
41  
42 535  
43  
44 536

**Affiliations:**

- 45 537 1. Division of General Surgery, Department of Surgery, McMaster University, Hamilton, Canada.  
46 538 2. Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton,  
47 539 Canada.  
48 540 3. Section of Biomedical Informatics and Data Science, Yale University School of Medicine, New  
49 541 Haven, USA.  
50 542 4. School of Medicine, University College Dublin - National University of Ireland, Dublin,  
51 543 Republic of Ireland.  
52 544 5. School of Life Course and Population Sciences, King's College London, London, UK.

- 1
- 2
- 3
- 4 545 6. School of Epidemiology and Public Health, University of Ottawa, Ottawa, Canada.
- 5 546 7. Centre for Journalology, Ottawa Methods Centre, Ottawa Hospital Research Institute, Ottawa,
- 6 547 Canada.
- 7
- 8 548 8. The BMJ, London, UK.
- 9
- 10 549 9. Department of Neurology, Harvard Medical School, Boston, USA.
- 11 550 10. Gillings School of Global Public Health, The University of North Carolina, Chapel Hill, USA.
- 12
- 13 551 11. Phelan Senior Scientist, Women's College Research Institute and ICES, Toronto, Canada.
- 14 552 12. Department of Medicine, University of Toronto, Toronto, Canada.
- 15
- 16 553 13. JAMA and JAMA Network, Chicago, USA.
- 17 554 14. Executive Managing Editor and Vice President, JAMA and the JAMA Network, Chicago, USA.
- 18
- 19 555 15. Department of Surgery, Papageorgiou General Hospital, Thessaloniki, Greece.
- 20 556 16. Editor in Chief, Annals of Internal Medicine, Philadelphia, USA.
- 21
- 22 557 17. Senior VP, American College of Physicians, Philadelphia, USA.
- 23
- 24 558 18. Professor of Medicine, Sidney Kimmel Medical College, Thomas Jefferson University,
- 25 559 Philadelphia, USA.
- 26 560 19. USC Institute of Urology and Catherine and Joseph Aresty Department of Urology, Keck School
- 27 561 of Medicine, University of Southern California, Los Angeles, USA.
- 28
- 29 562 20. AI Center at USC Urology, University of Southern California, Los Angeles, USA.
- 30
- 31 563 21. Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology &
- 32 564 Musculoskeletal Sciences, University of Oxford, Oxford, UK.
- 33
- 34 565 22. Department of Oncology, McMaster University, Hamilton, Canada.
- 35
- 36 566 23. Department of General Anesthesiology, Anesthesiology Institute, Cleveland Clinic, Cleveland,
- 37 567 USA.
- 38
- 39 568 24. Michael Gent Chair in Healthcare Research, McMaster University, Hamilton, Canada.
- 40 569 25. Harvard T.H. Chan School of Public Health, Harvard University, Boston, USA.
- 41
- 42 570 26. Lancet Digital Health, London, UK.
- 43
- 44 571 27. The Lancet, London, UK.
- 45 572 28. Department of General Surgery, Università Politecnica delle Marche, Ancona, Italy.
- 46
- 47 573 29. Hospital Clinico San Carlos, IdISSC, Universidad Complutense de Madrid, Madrid, Spain.
- 48 574 30. Division of General Internal Medicine, Department of Medicine, University Hospitals of Geneva,
- 49 575 Geneva, Switzerland.
- 50
- 51 576 31. Department of Medicine, Lovisenberg Diaconal Hospital, Oslo, Norway.
- 52
- 53 577 32. Ted Rogers Computational Program (F.F., C.-P.S.F.), Peter Munk Cardiac Centre, University
- 54 578 Health Network, Toronto, ON, Canada.
- 55
- 56 579 33. Institute for Evidence in Medicine, Medical Center – University of Freiburg, Faculty of Medicine,
- 57
- 58
- 59
- 60

## The CHART Reporting Guideline Research Protocol

2024/02/15

- 1  
2  
3  
4 580 University of Freiburg, Freiburg, Germany.  
5 581 34. Cochrane Germany, Cochrane Germany Foundation, Freiburg, Germany.  
6  
7 582 35. University of California, Davis, USA.  
8 583 36. Department of Family Practice, Faculty of Medicine, University of British Columbia, Vancouver,  
9  
10 584 Canada.  
11 585 37. Evidence-Based Medicine Center, School of Basic Medical Sciences, Lanzhou University,  
12  
13 586 Lanzhou, China.  
14 587 38. Hardien Health, Haywards Heath, UK.  
15  
16 588 39. Department of Urology and Population Health, New York University, New York, USA.  
17 589 40. IJS Publishing Group, London, UK.  
18  
19 590 41. Phelix AI, Toronto, Canada.  
20 591 42. Department of Medicine, McMaster University, Hamilton, Canada.  
21  
22 592  
23 593  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60



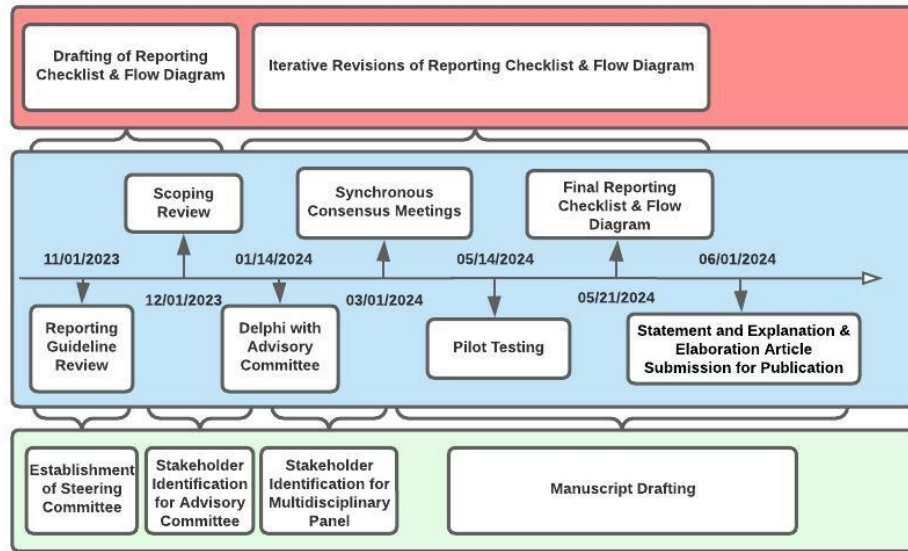


Figure 1. Timeline for the Development of the CHART Reporting Guideline.

136x87mm (160 x 160 DPI)

# BMJ Open

## Protocol for the Development of the Chatbot Assessment Reporting Tool (CHART) for Clinical Advice

Journal:	<i>BMJ Open</i>
Manuscript ID	bmjopen-2023-081155.R2
Article Type:	Protocol
Date Submitted by the Author:	11-Mar-2024
Complete List of Authors:	CHART Collaborative, The; McMaster University Huo, Bright; Dalhousie Medical School,
<b>Primary Subject Heading</b>:	Research methods
Secondary Subject Heading:	Research methods, Ethics, Evidence based practice
Keywords:	MEDICAL ETHICS, Natural Language Processing, STATISTICS & RESEARCH METHODS

SCHOLARONE™  
Manuscripts

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60



# Protocol for the Development of the Chatbot Assessment Reporting Tool (CHART) for Clinical Advice

---

The CHART Collaborative\*

## The CHART Reporting Guideline Research Protocol

2024/02/15

25

26

27

**Corresponding author:**

Bright Huo, on behalf of The CHART Collaborative\*

237 Barton St E, Hamilton, ON L8L 2X2

E: brighthuo@dal.ca

T: +1 902 448 6836

33

**Keywords:** Medical ethics, statistics & research methods, natural language processing

35

**ABSTRACT**

37

**Introduction:**

Large language model (LLM)-linked chatbots are being increasingly applied in healthcare due to their

impressive functionality and public availability. Studies have assessed the ability of LLM-linked chatbots

to provide accurate clinical advice. However, the methods applied in these Chatbot Assessment Studies

are inconsistent due to the lack of reporting standards available, which obscures the interpretation of their

study findings. This protocol outlines the development of the Chatbot Assessment Reporting Tool

(CHART) reporting guideline.

45

**Methods and analysis:**

The development of the CHART reporting guideline will consist of three phases, led by the Steering

Committee. During phase one, the team will identify relevant reporting guidelines with artificial

1  
2  
3  
4 49 intelligence extensions that are published or in-development by searching preprint servers, protocol  
5  
6  
7 50 databases, and the Enhancing the Quality and Transparency of health research (EQUATOR) Network.  
8  
9  
10 51 During phase two, we will conduct a scoping review to identify studies that have addressed the  
11  
12  
13 52 performance of LLM-linked chatbots in summarizing evidence and providing clinical advice. The  
14  
15  
16 53 Steering Committee will identify methodology used in previous Chatbot Assessment Studies. Finally, the  
17  
18  
19 54 study team will use checklist items from prior reporting guidelines and findings from the scoping review  
20  
21  
22 55 to develop a draft reporting checklist. We will then perform a Delphi consensus and host two synchronous  
23  
24  
25 56 consensus meetings with an international, multidisciplinary group of stakeholders to refine reporting  
26  
27  
28 57 checklist items and develop a flow diagram.  
29  
30

31 58

**59 Ethics and dissemination:**

36  
37 60 We will publish the final CHART reporting guideline in peer-reviewed journals and will present findings  
38  
39  
40 61 at peer-reviewed meetings. Ethical approval is not applicable for the development of the CHART  
41  
42  
43 62 reporting guideline.  
44  
45

46 63

**64 Registration:**

51  
52  
53 65 This study protocol is pre-registered with Open Science Framework:

54  
55  
56 66 <https://doi.org/10.17605/OSF.IO/59E2Q>.  
57  
58

67

**68 Strengths and limitations of this study:**

69 • This initiative will address a lack of reporting standards for Chatbot Assessment Studies and will  
70 provide a framework to increase the transparent conduct of these studies.

71 • We will apply rigorous methodology of the highest standards to develop the CHART reporting  
72 guideline. A diverse group of international, multidisciplinary stakeholders will inform the  
73 development of the CHART reporting checklist and flow diagram, with key input from experts in  
74 LLMs.

75 • This reporting guideline will be developed swiftly while acknowledging the dynamically  
76 evolving technology of LLM-linked chatbots.

77 • The CHART reporting guideline will apply specifically for studies assessing the ability of LLM-  
78 linked chatbots to summarize evidence and provide clinical advice. It will not apply to their use in  
79 other settings.

80 • To avoid the limitation that this reporting checklist may become outdated sooner than  
81 conventional reporting tools, the steering committee will assess the need to update the checklist  
82 on an annual basis, driven by the junior primary investigator.

**83**  
**84 INTRODUCTION**

1  
2  
3  
4 85 Novel chatbots have been integrating Large Language Models (LLMs), which are a popular technology in  
5  
6 86 the field of natural language processing (NLP) [1]. LLMs are large neural networks often comprised of  
7  
8  
9 87 hundreds of billions of parameters, which impact the model's input, size and shape, and output [2]. LLMs  
10  
11  
12 88 are typically used to conditionally predict the next words in a sequence of text, given corresponding  
13  
14  
15 89 prompts (Table 1) [3]. LLMs can be trained on a collection of massive amounts of raw data from online  
16  
17  
18 90 text sources including books, articles, websites, and more [1,4]. Coupled with reinforcement learning  
19  
20  
21 91 from human feedback [5]. LLMs exhibit striking text generation capabilities, producing outputs that are  
22  
23  
24 92 often indistinguishable from human language [6,7]. There has been a gold-rush movement of chatbots  
25  
26  
27 93 linked to LLMs, with recent releases including ChatGPT, Bing Chat, Google Bard, Med-PaLM, and many  
28  
29  
30 94 more underway [8].  
31  
32  
33  
34 95  
35  
36  
37 96 Given their wide accessibility and ability to provide answers to lay prompts [8], investigators have begun  
38  
39  
40 97 to assess LLM-linked chatbots as a potential source of health advice for both patients and clinicians [9–  
41  
42  
43 98 11]. We refer to these studies as Chatbot Assessment Studies, and they evaluate the performance of LLM-  
44  
45  
46 99 linked chatbots in summarizing health evidence and providing clinical advice. These studies represent a  
47  
48  
49 100 new genre of medical research, but the methodology and framing of results reported in these studies are  
50  
51  
52 101 highly variable. Inconsistent and incomplete reporting limits readers' ability to judge the methodology  
53  
54  
55 102 and results of these studies, complicating their interpretation [12]. A need exists to assess the rigour of  
56  
57  
58  
59  
60

1  
2  
3  
4 103 their assessments [8], but currently there are no standardized reporting tools for Chatbot Assessment  
5  
6  
7 104 Studies.  
8  
9  
10 105  
11  
12  
13 106 Instruments have been created to address issues of suboptimal reporting and raise the standard of research  
14  
15  
16 107 quality, such as the Consolidated Standards of Reporting Trials (CONSORT) statement [13,14]. Such  
17  
18  
19 108 reporting guidelines provide a checklist and a flow diagram for a given study type. Since their  
20  
21  
22 109 development, extensions to reporting guidelines have been created to facilitate the integration of artificial  
23  
24  
25 110 intelligence [15–17]. However, LLM-linked chatbots and their accompanying applications have only  
26  
27  
28 111 recently emerged and are not captured by these reporting guidelines. This protocol outlines the  
29  
30  
31 112 development of a novel reporting checklist, the Chatbot Assessment Reporting Tool (CHART) to  
32  
33  
34 113 improve the reporting standards of Chatbot Assessment Studies.  
35  
36  
37  
38  
39  
40  
41  
42

#### 40 115 **Key Terminology**

43 116 Table 1 lists key terms included in this work.  
44  
45  
46  
47  
48  
49  
50 118 Table 1. Glossary.

Term	Definition
Artificial Intelligence (AI)	The science of developing computer systems that can perform complex tasks approximating human cognitive performance.
Natural Language	A branch of information science that seeks to enable computers to interpret and manipulate



Processing (NLP)	human text.
Large Language Model (LLM)	A type of NLP model comprising large neural networks trained over large amounts of text, usually to produce an output of continuations of text from corresponding prompts, known as next word prediction. <sup>+</sup>
Multimodal LLM	LLMs with the capacity to integrate input from various data types, including text, speech, and/or visual sources.
Next word prediction	The natural language processing task of predicting the next word in a sequence of text given context and model parameters.
Parameter	A <i>parameter</i> within an artificial intelligence algorithm is a variable that is tuned iteratively/automatically to optimize the intended outcome of the algorithm. Parameters may be at the model level to optimize tuning (hyperparameters) or "weights" within the model linking layer to layer (parameters)
LLM-Linked Chatbot	A program that permits users to interact with an algorithm (such as an LLM) designed to respond to user prompts. <sup>-</sup>
Chatbot Assessment Study	Any research study assessing the performance of chatbots in summarizing health evidence and/or providing clinical advice.
Chat Instance	An interface in a computing device through which communication takes place between a chatbot and its user through text with only one prompt.
Chat Session	An interface in a computing device through which communication takes place between a chatbot and its user through text with more than one prompt.
Query	The act of communicating with a LLM by inputting a prompt into the chatbot which might be a question, comment, or phrase to elicit specific desired outputs from an LLM. For example, one might input a prompt asking the LLM to summarize the evidence supporting the use of a given intervention.
Check query	Following formal query completion and performance evaluation, the act of repeating the initial query to ensure that chatbot outputs are consistent in summarizing the same evidence and providing the same clinical advice.
Prompt	Text input by a user into the chatbot for the purpose of communicating with the LLM.
Prompt Engineering	An iterative testing phase where various pieces of text are inputted into a chatbot to achieve an output, informing the development of study prompts.
Delphi study	A structured research method applied to answer a research question through the establishment of consensus across respondents.

119 +Generally speaking, "next word" prediction is one basic "pre-training" objective, but LLMs often undergo a subsequent round  
 120 of "supervision" in which they are guided by human feedback.

121 -Chatbots are not necessarily built atop LLMs, but the modern tools that have captured public imagination are.

1  
2  
3  
4 1225 123 **METHODS & ANALYSIS**6  
7  
8 124 **Study Overview & Objectives**9  
10  
11 125 This study consists of three phases to address the following objectives:12  
13  
14 126 1. To identify checklist items used in previous reporting guidelines and identify related reporting15  
16  
17 127 standards for studies assessing the use of artificial intelligence in healthcare.18  
19  
20 128 2. To perform a scoping review that will identify and characterize studies that have addressed the21  
22  
23 129 performance of LLMs in summarizing evidence and providing clinical advice. Specifically, the24  
25  
26 130 review will identify how authors evaluate chatbot performance in summarizing health evidence27  
28  
29 131 and providing clinical advice.30  
31  
32 132 3. Informed by the scoping review and a review of prior checklists, to develop an evidence-33  
34  
35 133 informed, expert-derived reporting guideline comprised of a checklist and flow diagram for36  
37  
38 134 studies assessing chatbot performance in summarizing health evidence and providing clinical39  
40  
41 135 advice.42  
43  
44  
45 13646  
47  
48 137 A Steering Committee will lead all key study initiatives. This group will include the following members:49  
50  
51 138 the project lead, the senior methodologist lead, an expert in chatbot assessment studies, a reporting52  
53  
54 139 checklist developer, and a journal editor. The group's responsibilities will be to guide the initiatives

1  
2  
3  
4 140 involved in the development of the CHART checklist. They will lead the review of relevant reporting  
5  
6  
7 141 checklists (phase one), the completion of the scoping review (phase two), and the development of the  
8  
9  
10 142 reporting guideline (phase three). Table 1 presents a glossary of key terms used in this work. Figure 1  
11  
12  
13 143 demonstrates the timeline for the development of the CHART reporting guideline, which began in  
14  
15  
16 144 November 2023 and will terminate in June 2024.

17  
18  
19 145  
20  
21  
22 146 Figure 1. Timeline for the Development of the CHART Reporting Guideline.

23  
24  
25 147  
26  
27  
28 148 This reporting guideline will emphasize transparent reporting standards for studies evaluating the  
29  
30  
31 149 performance of LLMs when providing clinical advice to patients and clinicians. It will apply to LLM-  
32  
33  
34 150 linked chatbots, but also LLMs more broadly. It will also apply to studies using both traditional and  
35  
36  
37 151 multimodal LLMs.

38  
39  
40 152  
41  
42  
43 153 **PHASE ONE**

44  
45  
46 154  
47  
48  
49 155 **Objective:** to identify checklist items used in previous reporting guidelines and identify related reporting  
50  
51  
52 156 standards for studies assessing the ability of LLMs to provide clinical advice.

53  
54  
55 157  
56  
57  
58  
59  
60

1  
2  
3  
4 158 *Identification of Existing Reporting Guidelines*

5  
6  
7 159 To identify relevant health research reporting guidelines to inform the development of our reporting  
8  
9  
10 160 guideline and checklist, the study team will search the EQUATOR network and identify reporting  
11  
12  
13 161 guidelines published prior to October 2023 that meet our inclusion criteria:

- 14  
15  
16 162 • Studies presenting primary data on the use of chatbots in any specialty in medicine.  
17  
18  
19 163 • Studies applying chatbots to summarize evidence and provide clinical advice.  
20  
21  
22 164 • Studies applying chatbots to answer one or more clinical question(s).  
23  
24  
25 165 • Any studies applying chatbots as an intervention, with or without the use of a comparator.  
26  
27

28 166  
29  
30  
31 167 To achieve this, the study team will use the “search for reporting guidelines” feature and toggle through  
32  
33  
34 168 each study type. We will review all reporting guidelines in each study type for comprehensiveness. We  
35  
36  
37 169 will review references from relevant reporting guidelines and related citations listed on PubMed for  
38  
39  
40 170 retrieved articles. To identify protocols of reporting guidelines, we will search Open Science Framework  
41  
42  
43 171 as well as applicable results obtained from our scoping review. To identify ongoing or completed work  
44  
45  
46 172 not yet published in peer-reviewed sources, we will search Open Science Framework & MedRxiv.  
47  
48  
49 173  
50  
51  
52 174 Reporting guidelines obtained from the search from phase one will inform the development of items for a  
53  
54  
55 175 preliminary draft version of the checklist.  
56  
57  
58  
59  
60

176

177 **PHASE TWO**

178

179 **Objective:** to perform a scoping review that will identify and characterize studies that have addressed the  
180 performance of LLMs in summarizing evidence and providing clinical advice. Specifically, the review  
181 will identify how authors evaluate chatbot performance in summarizing health evidence and providing  
182 clinical advice.

183

184 For the scoping review, the project lead will recruit a team that will include two other members that have  
185 previous experience with performing systematic reviews and scoping reviews as well as the senior  
186 methodological lead. The scoping review team will identify articles assessing the performance of chatbots  
187 when applied in healthcare. A separate protocol presents our search strategy, inclusion criteria, exclusion  
188 criteria, and other details related to the scoping review, which is under consideration for publication. Its  
189 development will be aligned with methodology guidance from the JBI Scoping Review Methodology  
190 Group [18].

191

192 In brief, the scoping review team will conduct a literature search using MEDLINE via Ovid, EMBASE  
193 via Elsevier, Scopus via Elsevier, and Web of Science to capture relevant studies published prior to

## The CHART Reporting Guideline Research Protocol

2024/02/15

1  
2  
3  
4 194 October 2023. The team will identify studies that evaluate the performance of LLM-linked chatbots when  
5  
6  
7 195 providing clinical advice. We will only consider primary data. The team will complete two rounds of  
8  
9  
10 196 screening by title and abstract and full-text to identify articles of interest. Next, we will perform manual  
11  
12  
13 197 forward and backward citation searching. The team will then perform data extraction to identify key items  
14  
15  
16 198 used in the reporting of these studies. The following variables will be extracted: clinical aims (health  
17  
18  
19 199 prevention, screening, differential diagnosis, diagnosis, treatment), prompt development (use of specific  
20  
21  
22 200 sources, engineering/testing phase, standardized prompts, prompt structure, prompt inclusion in-text)  
23  
24  
25 201 LLM, LLM model version, LLM characteristics (temperature, token length, fine-tuning availability,  
26  
27  
28 202 penalties, add-on availability, layers), date accessed/trained, language, location of query, use of chat  
29  
30  
31 203 windows/sessions, performance definition (objective use of literature such as guideline or systematic  
32  
33  
34 204 review versus subjective evaluation using experts), and whether a statement or discussion on ethics,  
35  
36  
37 205 regulation, or patient safety is included.  
38  
39  
40 206  
41  
42 207 We will report findings using descriptive statistics for quantitative data and present results graphically in  
43  
44  
45 208 diagrammatic form. A narrative summary will accompany the graphical results. The final report will  
46  
47  
48 209 adhere with reporting standards for the Preferred Reporting Items for Systematic Review and Meta-  
49  
50  
51 210 Analysis Extension for Scoping Reviews (PRISMA-ScR) [19].  
52  
53

211

1  
2  
3  
4 212 **PHASE THREE**

5  
6 213

7  
8  
9 214 **Objective:** informed by the scoping review and a review of prior checklists, to develop an evidence-

10  
11  
12 215 informed, expert-derived reporting guideline comprised of a checklist and flow diagram for studies

13  
14  
15 216 assessing chatbot performance in summarizing health evidence and providing clinical advice.

16  
17  
18  
19 217

20  
21  
22 218 *Advisory Committee & Delphi*

23  
24  
25 219 An Advisory Committee will comprise epidemiologists, research methodologists, NLP researchers,

26  
27  
28 220 journal editors, chatbot researchers, ethicists, regulatory experts, policy experts, and patient partners. The

29  
30  
31 221 Steering Committee will identify additional committee members by querying SCImago Journal Country

32  
33  
34 222 Rank (SJR) portal ([www.scimagojr.com](http://www.scimagojr.com)) to obtain a list of the top 10 journals in each specialty in

35  
36  
37 223 medicine. Using this list of journals, the Committee will query Web of Science to obtain a diverse list of

38  
39  
40 224 researchers in medicine including general research methodologists and chatbot researchers. Patient

41  
42  
43 225 partners will be identified through both public and internal calls through affiliate journals, as well as

44  
45  
46 226 through the snowballing method via our panel, including patient partner members. We will send an

47  
48  
49 227 invitation email to our final list of contacts to invite them to join the Advisory Committee.

50  
51  
52 228

53  
54  
55 229 The Steering Committee will hold a synchronous virtual meeting open to all Advisory Committee

## The CHART Reporting Guideline Research Protocol

2024/02/15

1  
2  
3  
4 230 members as an introduction to the project, as well as their role. Through a series of questionnaires shared  
5  
6  
7 231 through an online platform, the team will apply a Delphi consensus. The Steering Committee will develop  
8  
9  
10 232 a draft checklist informed by the scoping review and review of existing reporting guidelines. They will  
11  
12  
13 233 circulate the draft checklist to the Advisory Committee for a first round of voting. During this round,  
14  
15  
16 234 Advisory Committee members will select one of the following options for each checklist item: “include,  
17  
18  
19 235 maybe include, uncertain, maybe exclude, exclude.” There will be an additional option for Advisory  
20  
21  
22 236 Committee members to once more add checklist items. The Steering Committee will then revise the  
23  
24  
25 237 checklist using comments from the first round. The team will re-circulate the updated draft checklist for a  
26  
27  
28 238 second round of voting, as above.

29  
30  
31 239  
32  
33  
34 240 The Steering Committee will revise the checklist following the second round and present these items to  
35  
36  
37 241 the expert panel. In preparation for the next phase, the steering committee will meet with an ethicist and  
38  
39  
40 242 regulatory expert to review draft checklist items from the Delphi process to revise or add key principles  
41  
42  
43 243 for ethics & safety for discussion during the consensus meeting.

244

245 *Expert Panel*

51  
52  
53 246 We will create an international, multidisciplinary panel as per Moher and colleagues [12]. Participants  
54  
55  
56 247 will be purposefully selected to reflect a balanced representation of relevant stakeholders including  
57  
58  
59



1  
2  
3  
4 248 statisticians, research methodologists, reporting checklist developers, NLP researchers, journal editors,  
5  
6  
7 249 chatbot researchers, ethicists, regulatory experts, and two patient partners. In advance of the Consensus  
8  
9  
10 250 Meetings, the Steering Committee will prompt panelists to share their conflicts of interest. Though we  
11  
12  
13 251 find it difficult to imagine circumstances that would lead to important conflicts, we will stay alert to  
14  
15  
16 252 unanticipated conflicts. Should these arise, we will consider any panel member with significant conflicts  
17  
18  
19 253 as consultant who will not vote on the final checklist. Prior to the first of two Synchronous Consensus  
20  
21  
22 254 Meetings, the Steering Committee will share the candidate checklist items with the Expert Panel which  
23  
24  
25 255 will have been revised following two Delphi rounds with the Advisory Committee, informed by findings  
26  
27  
28 256 from the scoping review.  
29  
30  
31 257  
32  
33  
34 258 Additionally, the Steering Committee will construct a flow diagram prior to the Consensus Meetings  
35  
36  
37 259 based on the candidate checklist items. The purpose of the flow diagram is to provide an overview to  
38  
39  
40 260 guide authors in clearly reporting sequential stages of their study. The Steering Committee will also share  
41  
42  
43 261 this flow diagram with the panel prior to the Consensus Meetings.  
44  
45  
46 262  
47  
48  
49 263 In preparation for the synchronous consensus meetings, the Steering Committee will share relevant  
50  
51  
52 264 materials with the panel such as the meeting agenda, participant list, and the completed scoping review  
53  
54  
55 265 highlighting the content and extent of reporting of the content area. The Committee will also circulate the  
56  
57  
58  
59  
60

1  
2  
3  
4 266 draft checklist that emerged from the Delphi process to the Expert Panel through an electronic survey in  
5  
6  
7 267 advance of the meeting. The steering group has pre-specified an 80% threshold for inclusion to  
8  
9  
10 268 demonstrate majority consensus based on prior work [17]. We will group items with  $\geq 80\%$  consensus  
11  
12  
13 269 with the selection of “include” or “maybe include” together, posing to the panelists: “These items have  
14  
15  
16 270 been recommended for inclusion in our checklist. Do you agree or disagree?” Panelists will have the  
17  
18  
19 271 option of yes - include, no - exclude, unsure, and an additional option for comments.  
20  
21  
22 272  
23  
24  
25 273 We will also group items with  $\geq 80\%$  consensus for items with the selection of “exclude” or “maybe  
26  
27  
28 274 exclude,” posing to the panelists: “These items have been recommended for exclusion in our checklist.  
29  
30  
31 275 Do you agree or disagree?” Panelists will have the option of yes - exclude, no - include, and an additional  
32  
33  
34 276 option for comments. Items without 80% consensus will be gathered and panel members will indicate  
35  
36  
37 277 “include, maybe include, uncertain, maybe exclude, exclude.” There will also be an additional option for  
38  
39  
40 278 each question to suggest additional checklist items. We will collate the results of this survey in  
41  
42  
43 279 preparation for the Consensus Meetings.

44  
45  
46 28047  
48  
49 28150  
51  
52 282 *Synchronous Consensus Meetings*53  
54  
55 283 The project lead will organize two Synchronous Consensus Meetings that will be held over a video  
56  
57  
58  
59  
60

1  
2  
3  
4 284 conferencing platform. The Steering Committee will encourage panelists to attend both meetings, with the  
5  
6  
7 285 expectation that panelists must attend one meeting, at minimum. The steering committee will circulate an  
8  
9  
10 286 online scheduling survey in advance to control the number of participants in attendance, while also  
11  
12  
13 287 selecting dates that optimize the attendance of panel members. As we will hold these meetings virtually,  
14  
15  
16 288 no meeting will be longer than four hours in duration to mitigate burnout and encourage participation.  
17  
18  
19 289 The duration of both meetings will be eight hours in total. A contingency plan is set to pre-emptively  
20  
21  
22 290 arrange and hold a third meeting of two to four hours should additional time be needed following the  
23  
24  
25 291 eight hours of consensus meetings.

26  
27  
28 292  
29  
30  
31 293 During checklist item discussion, we will put forth any items rated as “no-exclude” by panelists during  
32  
33  
34 294 the pre-consensus meeting survey for exclusion from the checklist. We will then discuss any items  
35  
36  
37 295 without consensus or rated as “uncertain” with  $\geq 80\%$  consensus after the second Delphi round. Finally,  
38  
39  
40 296 we will offer items rated as “yes-include” to the panel for inclusion in the checklist. During the discussion  
41  
42  
43 297 for all checklist items, the meeting chair will present the following for each checklist item:

- 44  
45  
46 298
- Previous use in a Chatbot Assessment Study
- 47  
48  
49 299
- Rationale for inclusion
- 50

51  
52  
53 300  
54  
55  
56 301 All voting will take place virtually and anonymously over the video conferencing platform. A working  
57  
58  
59  
60

## The CHART Reporting Guideline Research Protocol

2024/02/15

1  
2  
3  
4 302 CHART checklist will emerge from the Synchronous Consensus Meetings. The panel will use this  
5  
6  
7 303 working checklist to revise the draft CHART flow diagram during the Synchronous Consensus Meeting.  
8  
9  
10 304  
11  
12  
13 305 Expert panel members who are unable to join will be able to review recordings of the meetings. The  
14  
15  
16 306 project lead will record the meeting(s), and they will share both the recording and a summary of checklist  
17  
18  
19 307 item decisions and rationale with absent panel members.  
20  
21  
22 308  
23  
24  
25 309 Following the meetings, the Steering Committee will circulate the working CHART checklist and flow  
26  
27  
28 310 diagram in the form of a survey reflecting checklist item decisions. This working checklist will outline a  
29  
30  
31 311 final list of items for inclusion. Panellists will have the opportunity to provide any final comments, which  
32  
33  
34 312 the Steering Committee will use to derive a preliminary CHART checklist. The preliminary checklist will  
35  
36  
37 313 also be shared with the public for open comment on the EQUATOR website, while links to the checklist  
38  
39  
40 314 will be shared on the website of affiliate journals of editors involved in the development of the CHART  
41  
42  
43 315 reporting guideline.  
44  
45  
46 316  
47  
48  
49 317 Prior to pilot testing, the study team will share the preliminary checklist following the consensus meetings  
50  
51  
52 318 with patient partners identified a priori through snowballing and journal contacts to ensure that themes of  
53  
54  
55 319 patient access and safety are sufficiently addressed.  
56  
57  
58  
59  
60

1  
2  
3  
4 320  
5  
67 321 *Pilot Testing*  
8  
910 322 The Steering Committee will pilot the preliminary CHART checklist and flow diagram with researchers  
11  
1213 323 that have published Chatbot Assessment Studies and will identify authors by the included studies in the  
14  
1516 324 scoping review. The Steering Committee will conduct pilot testing via an iterative process. Groups of five  
17  
1819 325 authors will provide feedback in each round until saturation is achieved, with a minimum of ten authors  
20  
2122 326 over two rounds of pilot testing. Authors will not evaluate their own studies but will use the checklist to  
23  
2425 327 assess Chatbot Assessment Studies published by other authors. During synchronous sessions, we will ask  
26  
2728 328 authors to assess Chatbot Assessment Studies using the preliminary CHART checklist and flow diagram  
29  
3031 329 via think-aloud instrument testing. Authors will provide practical feedback regarding the development of  
32  
3334 330 these studies in the context of checklist items. They will also provide feedback regarding the practical  
35  
3637 331 application of the preliminary CHART checklist with respect to the length and content of the checklist.  
38  
3940 332  
41  
4243 333 The Steering Committee will use the comments from Chatbot Assessment Study researchers to derive a  
44  
4546 334 final version of the CHART checklist and flow diagram.  
47  
4849 335  
50  
5152 336 *Report Generation*  
53  
5455 337 With the final CHART checklist and flow diagram, the Steering Committee will prepare a Statement  
56  
57  
58  
59  
60

## The CHART Reporting Guideline Research Protocol

2024/02/15

1  
2  
3  
4 338 document for submission for peer-reviewed conference presentation and publication. All panel members  
5  
6  
7 339 will have the chance to review the draft manuscript, and all members of the research team satisfying the  
8  
9  
10 340 International Committee of Medical Journal Editors (ICJME) criteria will join the group authorship.[20]  
11  
12  
13 341 The Statement article will consist of the checklist and flow diagram. It will include the rationale for  
14  
15  
16 342 developing the CHART guideline and an overview of its development, including a brief description of the  
17  
18  
19 343 meeting and participants involved.  
20  
21  
22 344  
23  
24  
25 345 Separately, the Steering Committee will prepare a detailed explanation and elaboration paper (E&E). This  
26  
27  
28 346 paper will provide more detail for the inclusion of items in the final CHART checklist. For each checklist  
29  
30  
31 347 item, the E&E report will include three parts: 1) an explanation of the rationale supporting the checklist  
32  
33  
34 348 item, as well as reference to any supporting evidence for its inclusion 2) essential elements of the study  
35  
36  
37 349 that must be described to appropriately satisfy each checklist item 3) additional elements of the study  
38  
39  
40 350 which may be considered by authors depending on the context. Both the Statement and E&E articles will  
41  
42  
43 351 be written in collaboration with the multidisciplinary panel.  
44  
45  
46 352  
47  
48  
49 353 As per Moher and colleagues, we will simultaneously submit both the Statement and E&E articles for  
50  
51  
52 354 peer-reviewed publication [12].  
53  
54  
55  
56 355

1  
2  
3  
4 356 *Patient & Public Involvement*

5  
6 357 Patients will be involved in the development of the CHART reporting guideline through participation in  
7  
8  
9 358 the Delphi process, as outlined above. Two patients will also be involved in the revision of the reporting  
10  
11  
12 359 guideline including the checklist, flow diagram, and resulting reports as panel members.  
13  
14  
15

16 360

17  
18  
19 361 *Funding*

20  
21  
22 362 This protocol submission is funded by *the First Cut Research Competition* at McMaster University.  
23  
24

25 363 Organizers of *the First Cut* had no involvement in planning the design of this study, the writing of this  
26  
27  
28 364 protocol manuscript, and will not be involved in the conduct of this study.  
29  
30

31 365

32  
33  
34 366 *Updates & Monitoring*

35  
36  
37 367 The field of LLM-linked chatbot research is evolving, and it is paramount that the CHART Reporting  
38  
39

40 368 Guidelines reflect the most modern advances in Chatbot Assessment Study research and LLM-linked  
41  
42

43 369 technology. To address this need, the project lead and senior methodologist lead will actively survey news  
44  
45

46 370 updates from both accessible and closed/proprietary chatbot models monthly. Beginning in 2025, the  
47  
48

49 371 project lead will assess the need to initiate an updated scoping review annually if changes to the study  
50  
51

52 372 aims, methodology, and/or quantity of published literature in this area is significant.  
53  
54

55  
56 373  
57  
58  
59  
60

1  
2  
3  
4 374 To inform the necessity of updates to the CHART reporting guidelines, both the project lead and senior  
5  
6  
7 375 methodologist lead will consider a combination of the updates in LLM-linked chatbot technology, as well  
8  
9  
10 376 as the study aims, methodology, and/or quantity of new Chatbot Assessment Studies.

11  
12  
13 37714 378 *Ethics*

15  
16  
17 379 This study was submitted to the Hamilton Integrated Research Ethics Board (HiREB). It was deemed that  
18  
19  
20 380 HiREB review and approval was not required. This study will adhere to key principles. All work will  
21  
22  
23 381 adhere to the World Medical Association Declaration of Helsinki Ethical Principles for Medical Research  
24  
25  
26 382 Involving Human Subjects [21]. Furthermore, all checklist items for future studies involving the use of  
27  
28  
29 383 LLMs for clinical advice will be reviewed in the context of these ethical principles [21]. The involvement  
30  
31  
32 384 of ethicists and regulatory experts in health technology will aid the steering committee and panel in  
33  
34  
35 385 considering these key principles, including accessibility and patient safety.

36  
37  
38  
39 38640 387 *Limitations*

41  
42  
43 388 This study has limitations. The reporting checklist will be applicable for the most current, conventional  
44  
45  
46 389 LLMs at the time of publication due to the dynamic pace at which this field is evolving. To address this,  
47  
48  
49 390 the steering committee will assess the need to update the checklist on an annual basis, driven by the junior  
50  
51  
52 391 primary investigator.

53  
54  
55  
56 392



1  
2  
3  
4 3935 394 **Acknowledgements:** The study team would like to thank Byron Wallace for his expert input.

6 395

7 396  
8 397 **Contributors:** Each author in the CHART collaborative contributed to the planning of the development of  
9 398 the CHART reporting guideline including the determination of its scope, study design, and the drafting of  
10 399 this protocol manuscript.

11 400

12 401 **Funding:** This work was supported by *the First Cut* from the Department of Surgery at McMaster  
13 402 University (funding number not applicable). The organizers of *the First Cut* competition were not  
14 403 involved in this study at any stage including the conception, planning, or creation of this study protocol.

15 404

16 405 **Competing interests:** None declared.

17 406

18 407 **Word Count:** 3,752

19 408

20 409

21 410

22 411

23 412

24 413

25 414

26 415

27 416

28 417

29 418

30 419

31 420

32 421

33 422

34 423

35 424

36 425

37 426

38 427

39 428

40 429

41 430

42 431

43 432

1  
2  
3  
4 428  
5 429  
6  
7 430  
8 431  
9  
10 432  
11 433  
12  
13 434  
14 435  
15  
16 436  
17 437  
18 438  
19 439  
20 440  
21 441  
22 442  
23 443  
24 444  
25 444  
26 445  
27 446  
28 447  
29 448  
30 448  
31 449  
32 450  
33 451  
34 451  
35 452  
36 453  
37 454  
38 454  
39 455  
40 456  
41 457  
42 458  
43 459  
44 460  
45 460  
46 461  
47 462  
48 463  
49 463  
50 464  
51 465  
52 466  
53 467  
54 467  
55 468  
56 469  
57  
58  
59  
60

## References

1. Thirunavukarasu AJ, Ting DSJ, Elangovan K, et al. Large language models in medicine. *Nat Med*. 2023;29:1930-1940. doi:10.1038/s41591-023-02448-8
2. Gholami S, Omar M. Do Generative Large Language Models need billions of parameters? *arXiv*. 2023:1-15. <http://arxiv.org/abs/2309.06589>
3. Krishna Vamsi G, Rasool A, Hajela G. Chatbot A Deep Neural Network Based Human to Machine Conversation Model, *IEEE*. 2023;1-7. <https://ieeexplore.ieee.org/document/9225395>
4. Cascella M, Montomoli J, Bellini V, et al. Evaluating the Feasibility of ChatGPT in Healthcare: An Analysis of Multiple Clinical and Research Scenarios. *J Med Syst*. 2023;47:33. doi:10.1007/s10916-023-01925-4
5. Ziegler DM, Stiennon N, Wu J, et al. Fine-Tuning Language Models from Human Preferences. *arXiv*. 2019;1-26. <http://arxiv.org/abs/1909.08593>
6. Bhirud N, Randive S, Tataale S, et al. A Literature Review On Chatbots In Healthcare Domain. *Int J Sci Technol Res*. 2019;8:225-232.
7. Sallam M. ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. *Healthcare*. 2023;11:1-20. doi:10.3390/healthcare11060887
8. Rudolph J, Tan S, Tan S. War of the chatbots: Bard, Bing Chat, ChatGPT, Ernie and beyond. The new AI gold rush and its impact on higher education. *JALT*. 2023;6:364-389. doi:10.37074/jalt.2023.6.1.23
9. Ayers JW, Poliak A, Dredze M, et al. Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. *JAMA Intern Med*. 2023;183:589-596. doi:10.1001/jamainternmed.2023.1838
10. Haver HL, Ambinder EB, Bahl M, et al. Appropriateness of Breast Cancer Prevention and Screening Recommendations Provided by ChatGPT. *Radiology*. 2023;307:e230424. doi:10.1148/radiol.230424
11. Rahsepar AA, Tavakoli N, Kim GHJ, et al. How AI Responds to Common Lung Cancer Questions: ChatGPT vs Google Bard. *Radiology*. 2023;307:e230922. doi:10.1148/radiol.230922
12. Moher D, Schulz KF, Simera I, et al. Guidance for developers of health research reporting guidelines. *PLoS Med*. 2010;7:e1000217. doi:10.1371/journal.pmed.1000217
13. Begg C, Cho M, Eastwood S, et al. Improving the Quality of Reporting of Randomized Controlled Trials. The CONSORT Statement. *JAMA*. 1996;276:637-9. doi:10.1001/jama.276.8.637

- 1  
2  
3 470 14. Moher D, Hopewell S, Schulz KF, et al. CONSORT 2010 explanation and elaboration:  
4 471 updated guidelines for reporting parallel group randomised trials. *BMJ*. 2010;340:1-28.  
5 472 doi:10.1136/bmj.c869  
6  
7 473 15. Vasey B, Nagendran M, Campbell B, et al. Reporting guideline for the early-stage clinical  
8 474 evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *Nat*  
9 475 *Med*. 2022;28:924-933. doi:10.1038/s41591-022-01772-9  
10  
11 476 16. Rivera SC, Liu X, Chan AW, et al. Guidelines for clinical trial protocols for interventions  
12 477 involving artificial intelligence: The SPIRIT-AI Extension. *The BMJ*. 2020;370:m3210.  
13 478 doi:10.1136/bmj.m3210  
14 479 17. Liu X, Rivera SC, Moher D, et al. Reporting guidelines for clinical trial reports for  
15 480 interventions involving artificial intelligence: The CONSORT-AI Extension. *The BMJ*.  
16 481 2020;370:m3164. doi:10.1136/bmj.m3164  
17  
18 482 18. Peters MDJ, Marnie C, Tricco AC, et al. Updated methodological guidance for the conduct  
19 483 of scoping reviews. *JBI Evid Synth*. 2020;18:2119-2126. doi:10.11124/JBIES-20-00167  
20 484 19. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: An updated  
21 485 guideline for reporting systematic reviews. *The BMJ*. 2021;372:n71. doi:10.1136/bmj.n71  
22 486 20. Javed Ali M. ICMJE criteria for authorship: why the criticisms are not justified? *Graefes*  
23 487 *Arch Clin Exp Ophthalmol*. 2021;259:289-290. doi:10.1007/s00417-020-04825-2  
24 488 21. World Medical Association declaration of Helsinki: Ethical principles for medical  
25 489 research involving human subjects. *JAMA*. 2013;310(20):2191-2194.  
26 490 doi:10.1001/jama.2013.281053  
27  
28 491  
29  
30 492  
31  
32 493  
33  
34 494  
35 495  
36 496  
37  
38 497  
39  
40 498  
41 499  
42  
43 500  
44 501  
45  
46 502  
47  
48 503  
49 504  
50  
51 505  
52 506  
53  
54 507  
55 508  
56  
57  
58  
59  
60

1  
2  
3  
4 509  
5 510  
6  
7 511  
8 512  
9  
10 513  
11 514  
12  
13 515  
14 516  
15  
16 517  
17 518  
18  
19 519  
20  
21 520  
22 521  
23  
24 522  
25 523  
26  
27 524  
28 525  
29  
30 526  
31 527  
32

33 **\*The CHART Collaborative.**

34 529  
35  
36 530  
37 531  
38 532  
39 533  
40 534  
41 535  
42 536  
43 537  
44 538  
45 539  
46 540  
47 541  
48 542  
49 543  
50 544  
51 545  
52 546  
53 547  
54 548  
55 549  
56 550  
57 551  
58 552  
59 553  
60 554

**Authors:** Bright Huo,<sup>1</sup> Tyler McKechnie,<sup>1,2</sup> David Chartash,<sup>3,4</sup> Iain J Marshall,<sup>5</sup> David Moher,<sup>6,7</sup> Jeremy Y Ng,<sup>7</sup> Elizabeth Loder,<sup>8,9</sup> Timothy Feeney,<sup>8,10</sup> An-Wen Chan,<sup>11, 12</sup> Michael Berkwits,<sup>13</sup> Annette Flanagan,<sup>13,14</sup> Stavros A Antoniou,<sup>15</sup> Christine Laine,<sup>16, 17, 18</sup> Giovanni E Cacciamani,<sup>19, 20</sup> Gary S Collins,<sup>21</sup> Ashirbani Saha,<sup>22</sup> Piyush Mathur,<sup>23</sup> Alfonso Iorio,<sup>2,24</sup> Yung Lee,<sup>1,25</sup> Diana Samuel,<sup>26</sup> Helen Frankish,<sup>27</sup> Monica Ortenzi,<sup>28</sup> Julio Mayol,<sup>29</sup> Cynthia Lokker,<sup>2</sup> Thomas Agoritsas,<sup>2,30</sup> Per Olav Vandvik,<sup>31</sup> Farid Foroutan,<sup>2,32</sup> Joerg J. Meerpohl,<sup>33, 34</sup> Hugo Campos,<sup>35</sup> Carolyn Canfield,<sup>36</sup> Xufei Luo,<sup>37</sup> Yaolong Chen,<sup>37</sup> Hugh Harvey,<sup>38</sup> Stacy Loeb,<sup>39</sup> Riaz Agha,<sup>40</sup> Karim Ramji,<sup>1,41</sup> Hassaan Ahned,<sup>41</sup> Vanessa Boudreau,<sup>1</sup> Gordon Guyatt.<sup>2,42</sup>

**Affiliations:**

1. Division of General Surgery, Department of Surgery, McMaster University, Hamilton, Canada.
2. Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, Canada.

3. Section of Biomedical Informatics and Data Science, Yale University School of Medicine, New Haven, USA.
4. School of Medicine, University College Dublin - National University of Ireland, Dublin, Republic of Ireland.
5. School of Life Course and Population Sciences, King's College London, London, UK.
6. School of Epidemiology and Public Health, University of Ottawa, Ottawa, Canada.
7. Centre for Journalology, Ottawa Methods Centre, Ottawa Hospital Research Institute, Ottawa, Canada.
8. The BMJ, London, UK.
9. Department of Neurology, Harvard Medical School, Boston, USA.
10. Gillings School of Global Public Health, The University of North Carolina, Chapel Hill, USA.
11. Phelan Senior Scientist, Women's College Research Institute and ICES, Toronto, Canada.
12. Department of Medicine, University of Toronto, Toronto, Canada.
13. JAMA and JAMA Network, Chicago, USA.
14. Executive Managing Editor and Vice President, JAMA and the JAMA Network, Chicago, USA.
15. Department of Surgery, Papageorgiou General Hospital, Thessaloniki, Greece.
16. Editor in Chief, Annals of Internal Medicine, Philadelphia, USA.
17. Senior VP, American College of Physicians, Philadelphia, USA.
18. Professor of Medicine, Sidney Kimmel Medical College, Thomas Jefferson University, Philadelphia, USA.
19. USC Institute of Urology and Catherine and Joseph Aresty Department of Urology, Keck School of Medicine, University of Southern California, Los Angeles, USA.
20. AI Center at USC Urology, University of Southern California, Los Angeles, USA.
21. Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology & Musculoskeletal Sciences, University of Oxford, Oxford, UK.
22. Department of Oncology, McMaster University, Hamilton, Canada.
23. Department of General Anesthesiology, Anesthesiology Institute, Cleveland Clinic, Cleveland, USA.
24. Michael Gent Chair in Healthcare Research, McMaster University, Hamilton, Canada.
25. Harvard T.H. Chan School of Public Health, Harvard University, Boston, USA.
26. Lancet Digital Health, London, UK.
27. The Lancet, London, UK.
28. Department of General Surgery, Università Politecnica delle Marche, Ancona, Italy.
29. Hospital Clinico San Carlos, IdISSC, Universidad Complutense de Madrid, Madrid, Spain.
30. Division of General Internal Medicine, Department of Medicine, University Hospitals of Geneva,

- 1  
2  
3  
4 579 Geneva, Switzerland.
- 5 580 31. Department of Medicine, Lovisenberg Diaconal Hospital, Oslo, Norway.
- 6  
7 581 32. Ted Rogers Computational Program (F.F., C.-P.S.F.), Peter Munk Cardiac Centre, University  
8 582 Health Network, Toronto, ON, Canada.
- 9  
10 583 33. Institute for Evidence in Medicine, Medical Center – University of Freiburg, Faculty of Medicine,  
11 584 University of Freiburg, Freiburg, Germany.
- 12  
13 585 34. Cochrane Germany, Cochrane Germany Foundation, Freiburg, Germany.
- 14 586 35. University of California, Davis, USA.
- 15  
16 587 36. Department of Family Practice, Faculty of Medicine, University of British Columbia, Vancouver,  
17 588 Canada.
- 18  
19 589 37. Evidence-Based Medicine Center, School of Basic Medical Sciences, Lanzhou University,  
20 590 Lanzhou, China.
- 21  
22 591 38. Hardien Health, Haywards Heath, UK.
- 23  
24 592 39. Department of Urology and Population Health, New York University, New York, USA.
- 25 593 40. IJS Publishing Group, London, UK.
- 26  
27 594 41. Phelix AI, Toronto, Canada.
- 28 595 42. Department of Medicine, McMaster University, Hamilton, Canada.
- 29  
30 596  
31 597

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

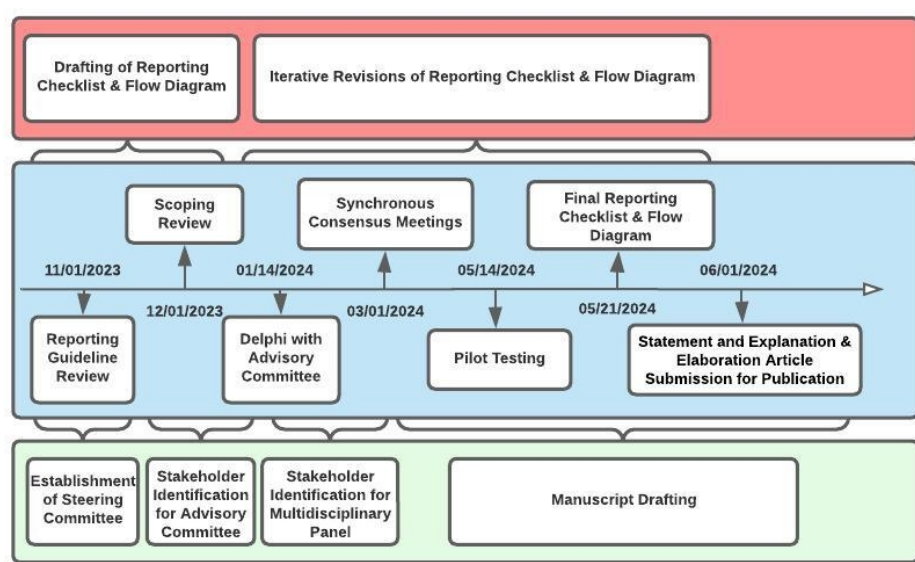


Figure 1. Timeline for the Development of the CHART Reporting Guideline.

136x87mm (160 x 160 DPI)