BMJ Open

Protocol for the development of the Chatbot Assessment Reporting Tool (CHART) for clinical advice

The CHART Collaborative

To cite: The CHART Collaborative. Protocol for the development of the Chatbot Assessment Reporting Tool (CHART) for clinical advice. *BMJ Open* 2024;**14**:e081155. doi:10.1136/bmjopen-2023-081155

► Prepublication history for this paper is available online. To view these files, please visit the journal online (https://doi.org/10.1136/bmjopen-2023-081155).

Received 24 October 2023
Accepted 26 March 2024

Check for updates

Correspondence to
Bright Huo; brighthuo@dal.ca

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 intelligence extensions that are published or in development by searching preprint servers, protocol databases, and the Enhancing the Quality and Transparency of health research Network. During phase two, we will conduct a scoping review to identify studies that have addressed the performance of LLM-linked chatbots in summarising evidence and providing clinical advice. The Steering Committee will identify methodology used in previous Chatbot Assessment Studies. Finally, the study team will use checklist items from prior reporting guidelines and findings from the scoping review to develop a draft reporting checklist. We will then perform a Delphi consensus and host two synchronous consensus meetings with an international, multidisciplinary group of stakeholders to refine reporting checklist items and develop a flow diagram.

Ethics and dissemination We will publish the final CHART reporting guideline in peer-reviewed journals and will present findings at peer-reviewed meetings. Ethical approval was submitted to the Hamilton Integrated Research Ethics Board and deemed "not required" in accordance with the Tri-Council Policy Statement (TCPS2) for the development of the CHART reporting guideline (#17025).

Registration This study protocol is preregistered with Open Science Framework: https://doi.org/10.17605/OSF.IO/59E2Q.

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 Chatbot Assessment Reporting Tool (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 large language models (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 to studies assessing the ability of LLM-linked chatbots to summarise evidence and provide clinical advice. It will not apply to their use in other settings.

⇒ To avoid the limitation that this reporting checklist may become outdated sooner than conventional reporting tools, the Steering Committee will assess the need to update the checklist on an annual basis, driven by the junior primary investigator.

INTRODUCTION

Novel chatbots have been integrating large language models (LLMs), which are a popular technology in the field of natural language processing (NLP).[1] LLMs are large neural networks often comprised of hundreds of billions of parameters, which impact the model's input, size and shape, and output.[2] LLMs are typically used to conditionally predict the next words in a sequence of text, given corresponding prompts (table 1).[3] LLMs can be trained on a collection of massive amounts of raw data from online text sources including books, articles, websites and more,[1,4] coupled with reinforcement learning from human feedback.[5] LLMs exhibit striking text generation capabilities, producing outputs that are often indistinguishable from human language.[6,7] There has been a goldrush movement of chatbots linked to LLMs, with recent releases including ChatGPT, Bing Chat, Google Bard, Med-PaLM and many more underway.[8]

Given their wide accessibility and ability to provide answers to lay prompts,[8] investigators

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.*
Multimodal LLM	LLMs with the capacity to integrate input from various data types, including text, speech and/or visual sources.
Next word prediction	The NLP task of predicting the next word in a sequence of text given context and model parameters.
Parameter	A *parameter* within an AI algorithm is a variable that is tuned iteratively/automatically to optimise the intended outcome of the algorithm. Parameters may be at the model level to optimise 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 summarising 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 an 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 summarise 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 summarising 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 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 are.

have begun to assess LLM-linked chatbots as a potential source of health advice for both patients and clinicians.[9–11] We refer to these studies as Chatbot Assessment Studies, and they evaluate the performance of LLM-linked chatbots in summarising health evidence and providing clinical advice. These studies represent a new genre of medical research, but the methodology and framing of results reported in these studies are highly variable. Inconsistent and incomplete reporting limits readers' ability to judge the methodology and results of these studies, complicating their interpretation.[12] A need exists to assess the rigour of their assessments,[8] but currently, there are no standardised reporting tools for Chatbot Assessment Studies.

Instruments have been created to address issues of suboptimal reporting and raise the standard of research quality, such as the Consolidated Standards of Reporting Trials statement.[13 14] 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.[15–17] 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.

METHODS AND ANALYSIS

Study overview and objectives

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

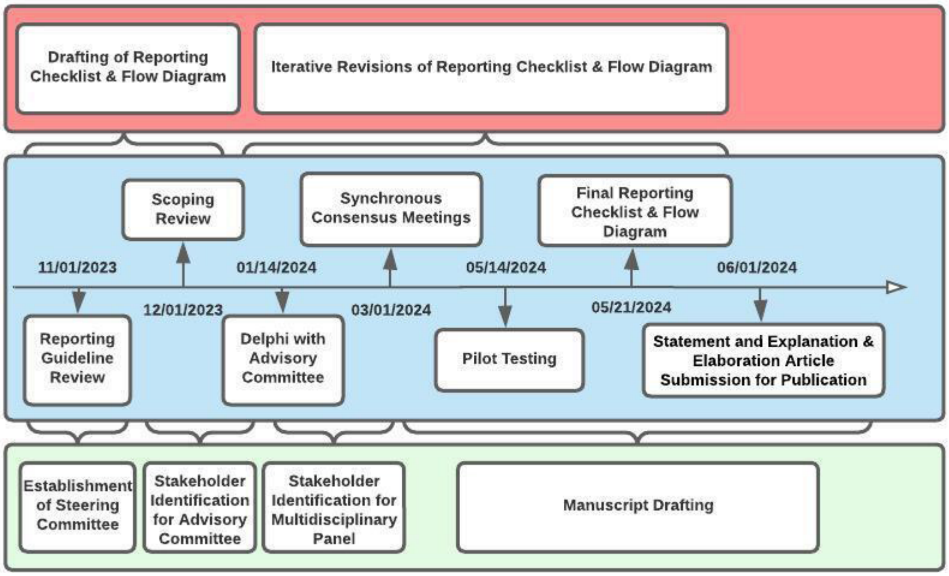
Drafting of Reporting Checklist & Flow Diagram
Iterative Revisions of Reporting Checklist & Flow Diagram
Scoping Review
Synchronous Consensus Meetings
Final Reporting Checklist & Flow Diagram
11/01/2023
01/14/2024
05/14/2024
06/01/2024
12/01/2023
03/01/2024
05/21/2024
Reporting Guideline Review
Delphi with Advisory Committee
Pilot Testing
Statement and Explanation & Elaboration Article Submission for Publication
Establishment of Steering Committee
Stakeholder Identification for Advisory Committee
Stakeholder Identification for Multidisciplinary Panel
Manuscript Drafting

Figure 1 Timeline for the development of the Chatbot Assessment Reporting Tool reporting guideline.

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 characterise studies that have addressed the performance of LLMs in summarising evidence and providing clinical advice. Specifically, the review will identify how authors evaluate chatbot performance in summarising 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 summarising 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, which began in November 2023 and will terminate in June 2024.

This reporting guideline will emphasise transparent reporting standards for studies evaluating the performance of LLMs when providing clinical advice to patients and clinicians. It will apply to LLM-linked chatbots, but also LLMs more broadly. It will also apply to studies using both traditional and multimodal LLMs.

PHASE ONE

Objective

To identify checklist items used in previous reporting guidelines and identify related reporting standards for studies assessing the ability of LLMs to provide clinical advice.

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 Enhancing the Quality and Transparency of health research (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 summarise 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.

To achieve this, the study team will use the 'search for reporting guidelines' feature and toggle through each study type. We will review all reporting guidelines in each study type for comprehensiveness. 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 as well as applicable results obtained from our scoping review. To identify ongoing or completed work not yet published in peer-reviewed sources, we will search Open Science Framework and MedRxiv.

Reporting guidelines obtained from the search from phase one will inform the development of items for a preliminary draft version of the checklist.

PHASE TWO

Objective

To perform a scoping review that will identify and characterise studies that have addressed the performance of LLMs in summarising evidence and providing clinical advice. Specifically, the review will identify how authors evaluate chatbot performance in summarising health evidence and providing clinical advice.

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

In brief, the scoping review team will conduct a literature search using MEDLINE via Ovid, EMBASE via Elsevier, Scopus via Elsevier and Web of Science to capture relevant studies published prior to October 2023. The team will identify studies that evaluate the performance of LLM-linked chatbots when providing clinical advice. We will only consider primary data. The team will complete two rounds of screening by title and abstract and full text to identify articles of interest. Next, we will perform manual forward and backward citation searching. The team will then perform data extraction to identify key items used in the reporting of these studies. The following variables will be extracted: clinical aims (health prevention, screening, differential diagnosis, diagnosis, treatment), prompt development (use of specific sources, engineering/testing phase, standardised prompts, prompt structure, prompt inclusion in-text) of LLM, LLM model version, LLM characteristics (temperature, token length, fine-tuning availability, penalties, add-on availability, layers), date accessed/trained, language, location of query, use of chat windows/sessions, performance definition (objective use of literature such as guideline or systematic review vs subjective evaluation using experts), and whether a statement or discussion on ethics, regulation, or patient safety is included.

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 to reporting standards for the Preferred Reporting Items for Systematic Review and Meta-Analysis Extension for Scoping Reviews.[19]

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 summarising health evidence and providing clinical advice.

Advisory Committee and Delphi

An Advisory Committee will comprise epidemiologists, research methodologists, NLP researchers, journal editors, chatbot researchers, ethicists, regulatory experts, policy experts and patient partners. The Steering Committee will identify additional committee members by querying SCImago Journal Country Rank portal (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. Patient partners will be identified through both public and internal calls through affiliate journals, as well as through the snowballing method via our panel, including patient partner members. 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, maybe include, uncertain, maybe exclude, exclude'. There will be an additional option for Advisory Committee members to once more add checklist items. The Steering Committee will then revise the checklist using comments from the first round. The team will recirculate the updated draft checklist for a second round of voting, as above.

The Steering Committee will revise the checklist following the second round and present these items to the expert panel. In preparation for the next phase, the Steering Committee will meet with an ethicist and regulatory expert to review draft checklist items from the Delphi process to revise or add key principles for ethics and safety for discussion during the consensus meeting.

Expert panel

We will create an international, multidisciplinary panel as per Moher and colleagues.[12] Participants will be purposefully selected to reflect a balanced representation of relevant stakeholders including statisticians, research methodologists, reporting checklist developers, NLP researchers, journal editors, chatbot researchers, ethicists, regulatory experts and two patient partners. In advance of the consensus meetings, the Steering Committee will prompt panellists to share their conflicts of interest. Though we find it difficult to imagine circumstances that would lead to important conflicts, we will stay alert to unanticipated conflicts. Should these arise, we will consider any panel member with significant conflicts as consultant who will not vote on the final checklist. Prior to the first of two synchronous consensus meetings, the Steering Committee will share the candidate checklist items with the expert panel which will have been revised following two Delphi rounds with the Advisory Committee, informed by findings from the scoping review.

Additionally, the Steering Committee will construct a flow diagram prior to the consensus meetings based on the candidate checklist items. The purpose of the flow diagram is to provide an overview to guide authors in clearly reporting sequential stages of their study. The Steering Committee will also share this flow diagram with the panel prior to the consensus meetings.

In preparation for the synchronous consensus meetings, the Steering Committee will share relevant materials with the panel such as the meeting agenda, participant list and the completed scoping review, highlighting the content and extent of reporting of the content area. The committee will also circulate the draft checklist that emerged from the Delphi process to the expert panel through an electronic survey in advance of the meeting. The steering group has prespecified an 80% threshold for inclusion to demonstrate majority consensus based on prior work.[17] We will group items with ≥80% consensus with the selection of 'include' or 'maybe include' together, posing to the panellists: 'These items have been recommended for inclusion in our checklist. Do you agree or disagree?' Panellists will have the option of yes-include, no-exclude, unsure and an additional option for comments.

We will also group items with ≥80% consensus for items with the selection of 'exclude' or 'maybe exclude,' posing to the panellists: 'These items have been recommended for exclusion in our checklist. Do you agree or disagree?' Panellists will have the option of yes-exclude, no-include and an additional option for comments. Items without 80% consensus will be gathered and panel members will indicate 'include, maybe include, uncertain, maybe exclude, exclude'. There will also be an additional option for each question to suggest additional checklist items. We will collate the results of this survey in preparation for the consensus meetings.

Synchronous consensus meetings

The project lead will organise two synchronous consensus meetings that will be held over a video-conferencing platform. The Steering Committee will encourage panellists to attend both meetings, with the expectation that panellists must attend one meeting, at minimum. The Steering Committee will circulate an online scheduling survey in advance to control the number of participants in attendance, while also selecting dates that optimise the attendance of panel members. As we will hold these meetings virtually, no meeting will be longer than 4 hours in duration to mitigate burnout and encourage participation. The duration of both meetings will be 8 hours in total. A contingency plan is set to pre-emptively arrange and hold a third meeting of 2–4 hours should additional time be needed following the 8 hours of consensus meetings.

During checklist item discussion, we will put forth any items rated as 'no-exclude' by panellists during the pre-consensus meeting survey for exclusion from the checklist. We will then discuss any items without consensus or rated as 'uncertain' with ≥80% consensus after the second Delphi round. Finally, we will offer items rated as 'yes-include' to the panel for inclusion in the checklist. During the discussion for all checklist items, the meeting chair will present the following for each checklist item:

- Previous use in a Chatbot Assessment Study.
- Rationale for inclusion.

All voting will take place virtually and anonymously over the video-conferencing platform. A working CHART checklist will emerge from the synchronous consensus meetings. The panel will use this working checklist to revise the draft CHART flow diagram during the synchronous consensus meeting.

Expert panel members who are unable to join will be able to review recordings of the meetings. The project lead will record the meeting(s), and they will share both the recording and a summary of checklist item decisions and rationale with absent panel members.

Following the meetings, the Steering Committee will circulate the working CHART checklist and flow diagram in the form of a survey reflecting checklist item decisions. This working checklist will outline a final list of items for inclusion. Panellists will have the opportunity to provide any final comments, which the Steering Committee will use to derive a preliminary CHART checklist. The preliminary checklist will also be shared with the public for open comment on the EQUATOR website, while links to the checklist will be shared on the website of affiliate journals of editors involved in the development of the CHART reporting guideline.

Prior to pilot testing, the study team will share the preliminary checklist following the consensus meetings with patient partners identified a priori through snowballing and journal contacts to ensure that themes of patient access and safety are sufficiently addressed.

Pilot testing

The Steering Committee will pilot the preliminary CHART checklist and flow diagram with researchers who have published Chatbot Assessment Studies and will identify authors by the included studies in the scoping review. The Steering Committee will conduct pilot testing via an iterative process. Groups of five authors will provide feedback in each round until saturation is achieved, with a minimum of 10 authors over two rounds of pilot testing. Authors will not evaluate their own studies but will use the checklist to assess Chatbot Assessment Studies published by other authors. During synchronous sessions, we will ask authors to assess Chatbot Assessment Studies using the preliminary CHART checklist and flow diagram via think-aloud instrument testing. Authors will provide practical feedback regarding the development of these studies in the context of checklist items. They will also provide feedback regarding the practical application of the preliminary CHART checklist with respect to the length and content of the checklist.

The Steering Committee will use the comments from Chatbot Assessment Study researchers to derive a final version of the CHART checklist and flow diagram.

Report generation

With the final CHART checklist and flow diagram, the Steering Committee will prepare a statement document for submission for peer-reviewed conference presentation and publication. All panel members will have the chance to review the draft manuscript, and all members of the research team satisfying the International Committee of Medical Journal Editors criteria will join the group authorship.[20] The statement article will consist of the checklist and flow diagram. It will include the rationale for developing the CHART guideline and an overview of its development, including a brief description of the meeting and participants involved.

Separately, the Steering Committee will prepare a detailed explanation and elaboration paper (E&E). This paper will provide more detail for the inclusion of items in the final CHART checklist. For each checklist item, the E&E report will include three parts: (1) an explanation of the rationale supporting the checklist item, as well as reference to any supporting evidence for its inclusion; (2) essential elements of the study that must be described to appropriately satisfy each checklist item; (3) additional elements of the study which may be considered by authors depending on the context. Both the statement and E&E articles will be written in collaboration with the multidisciplinary panel.

As per Moher and colleagues, we will simultaneously submit both the statement and E&E articles for peer-reviewed publication.[12]

Patient and public involvement

Patients will be involved in the development of the CHART reporting guideline through participation in the Delphi process, as outlined above. Two patients will also be involved in the revision of the reporting guideline including the checklist, flow diagram and resulting reports as panel members.

Funding

This protocol submission is funded by the *First Cut Research Competition* at McMaster University. Organisers of the *First Cut* had no involvement in planning the design of this study, the writing of this protocol manuscript and will not be involved in the conduct of this study.

Updates and 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 are 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.

Ethics

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

Limitations

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

Acknowledgements The study team would like to thank Byron Wallace for his expert input.

Collaborators Bright Huo(Division of General Surgery, Department of Surgery, McMaster University, Hamilton, Canada) Tyler McKechnie(Division of General Surgery, Department of Surgery, McMaster University, Hamilton, Canada;Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, Canada) David Chartash(Section of Biomedical Informatics and Data Science, Yale University School of Medicine, New Haven, USA;School of Medicine, University College Dublin - National University of Ireland, Dublin, Republic of Ireland)Iain J Marshall(School of Life Course and Population Sciences, King's College London, London, UK) David Moher(School of Epidemiology and Public Health, University of Ottawa, Ottawa, Canada;Centre for Journalology, Ottawa Methods Centre, Ottawa Hospital Research Institute, Ottawa, Canada)Jeremy Y Ng(Centre for Journalology, Ottawa Methods Centre, Ottawa Hospital Research Institute, Ottawa, Canada)Elizabeth Loder(The BMJ, London, UK;Department

of Neurology, Harvard Medical School, Boston, USA) Timothy Feeney(The BMJ, London, UK;GIlings School of Global Public Health, The University of North Carolina, Chapel Hill, USA)An-Wen Chan(Phelan Senior Scientist, Women's College Research Institute and ICES, Toronto, Canada;Department of Medicine, University of Toronto, Toronto, Canada)Michael Berkwits(JAMA and JAMA Network, Chicago, USA) Annette Flanagin(JAMA and JAMA Network, Chicago, USA;Executive Managing Editor and Vice President, JAMA and the JAMA Network, Chicago, USA) Stavros A Antoniou(Department of Surgery, Papageorgiou General Hospital, Thessaloniki, Greece)Christine Laine(Editor in Chief, Annals of Internal Medicine, Philadelphia, USA;Senior VP, American College of Physicians, Philadelphia, USA;professor of Medicine,Sidney Kimmel Medical College, Thomas Jefferson University, Philadelphia, USA)Giovanni E Cacciamani(USC Institute of Urology and Catherine and Joseph Aresty Department of Urology, Keck School of Medicine, University of Southern California, Los Angeles, USA;AI Center at USC Urology, University of Southern California, Los Angeles, USA)Gary S Collins(Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology & Musculoskeletal Sciences, University of Oxford, Oxford, UK)ShirbaniSaha(Department of Oncology, McMaster University, Hamilton, Canada)Piyush Mathur(Department of General Anesthesiology, Anesthesiology Institute, Cleveland Clinic, Cleveland, USA)Alfonso Iorio(Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, Canada;Michael Gent Chair in Healthcare Research, McMaster University, Hamilton, Canada)Yung Lee(Division of General Surgery, Department of Surgery, McMaster University, Hamilton, Canada;Harvard T.H. Chan School of Public Health, Harvard University, Boston, USA)Diana Samuel(Lancet Digital Health, London, UK)Helen Frankish(The Lancet, London, UK) Monica Ortenzi(Department of General Surgery, Università Politecnica delle Marche, Ancona, Italy) Julio Mayol(Hospital Clinico San Carlos, IdISSC, Universidad Complutense de Madrid, Madrid, Spain)Cynthia Lokker(Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, Canada)Thomas Agoritsas(Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, Canada;Division of General Internal Medicine, Department of Medicine, University Hospitals of Geneva, Geneva, Switzerland) Per Olav Vandvik(Department of Medicine, Lovisenberg Diaconal Hospital, Oslo, Norway)Farid Foroutan(Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, Canada;Ted Rogers Computational Program (F.F., C.-P.S.F.), Peter Munk Cardiac Centre, University Health Network, Toronto, ON, Canada)Joerg J. Meerpohl(Institute for Evidence in Medicine, Medical Center – University of Freiburg, Faculty of Medicine, University of Freiburg, Freiburg, Germany;Cochrane Germany, Cochrane Germany Foundation, Freiburg, Germany)Hugo Campos(University of California, Davis, USA)Carolyn Canfield(Department of Family Practice, Faculty of Medicine, University of British Columbia, Vancouver, Canada)Xufei Luo(Evidence-Based Medicine Center, School of Basic Medical Sciences, Lanzhou University, Lanzhou, China) Yaolong Chen(Evidence-Based Medicine Center, School of Basic Medical Sciences, Lanzhou University, Lanzhou, China) Hugh Harvey(Hardien Health, Haywards Heath, UK)Stacy Loeb(Department of Urology and Population Health, New York University, New York, USA)Riaz Agha(IJS Publishing Group, London, UK)Karim Ramji(Division of General Surgery, Department of Surgery, McMaster University, Hamilton, Canada; Phelix AI, Toronto, Canada)Hassaan Ahmed(Phelix AI, Toronto, Canada)Vanessa Boudreau(Division of General Surgery, Department of Surgery, McMaster University, Hamilton, Canada)Gordon Guyatt(Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, Canada;Department of Medicine, McMaster University, Hamilton, Canada)

Contributors Each author in the CHART Collaborative contributed to the planning of the development of the CHART reporting guideline including the determination of its scope, study design and the drafting of this protocol manuscript.

Funding This work was supported by the First Cut from the Department of Surgery at McMaster University (funding number not applicable).

Disclaimer The organisers of the First Cut Competition were not involved in this study at any stage including the conception, planning or creation of this study protocol.

Competing interests None declared.

Patient and public involvement Patients and/or the public were involved in the design, or conduct, or reporting, or dissemination plans of this research. Refer to the Methods section for further details.

Patient consent for publication Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1 Thirunavukarasu AJ, Ting DSJ, Elangovan K, *et al*. Large language models in medicine. *Nat Med* 2023;29:1930–40.
2 Gholami S, Omar M. Do Generative large language models need billions of parameters? *arXiv* 2023;1–15. Available: 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. Available: 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.
5 Ziegler DM, Stiennon N, Wu J, *et al*. Fine-tuning language models from human preferences. *arXiv* 2019;1–26. Available: 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;8:225–32.
7 Sallam M. Chatgpt utility in Healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. *Healthcare* 2023;11:887.
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;6:364–89.
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;183:589–96.
10 Haver HL, Ambinder EB, Bahl M, *et al*. Appropriateness of breast cancer prevention and screening recommendations provided by Chatgpt. *Radiology* 2023;307:e230424.
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.
12 Moher D, Schulz KF, Simera I, *et al*. Guidance for developers of health research reporting guidelines. *PLoS Med* 2010;7:e1000217.
13 Begg C, Cho M, Eastwood S, *et al*. Improving the quality of reporting of randomized controlled trials. *JAMA* 1996;276:637–9.
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: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;28:924–33.
16 Rivera SC, Liu X, Chan A-W, *et al*. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *BMJ* 2020;370: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. *BMJ* 2020;370:m3164.
18 Peters MDJ, Marnie C, Tricco AC, *et al*. Updated methodological guidance for the conduct of Scoping reviews. *JBI Evid Synth* 2020;18:2119–26.
19 Page MJ, McKenzie JE, Bossuyt PM, *et al*. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* 2021;372:71.
20 Ali MJ. ICMJE criteria for authorship: why the criticisms are not justified *Graefes Arch Clin Exp Ophthalmol* 2021;259:289–90.
21 World Medical Association. World Medical Association declaration of Helsinki: ethical principles for medical research involving human subjects. *JAMA* 2013;310:2191–4.