OXFORD
Collaborative Research

BJS, 2025, znaf142
https://doi.org/10.1093/bjs/znaf142
Guideline

Reporting guideline for chatbot health advice studies: the Chatbot Assessment Reporting Tool (CHART) statement

The CHART Collaborative

*Correspondence to: Bright Huo, Division of General Surgery, Department of Surgery, McMaster University, 50 Charlton Ave. E., Hamilton, Ontario, Canada L8N 1Y3 (e-mail: brighthuo@dal.ca)

Members of The CHART Collaborative are co-authors of this study and are listed under the heading Collaborators.

This article is being published jointly in *Artificial Intelligence in Medicine*, *Annals of Family Medicine*, *BJS*, *BMC Medicine*, *BMJ Medicine*, and *JAMA Network Open*. The article is identical except for minor stylistic and spelling differences in keeping with each journal's style. Citations from any of the journals can be used when citing this article.

Abstract

The Chatbot Assessment Reporting Tool (CHART) is a reporting guideline developed to provide reporting recommendations for studies evaluating the performance of generative artificial intelligence (AI)-driven chatbots when summarizing clinical evidence and providing health advice, referred to as chatbot health advice studies. CHART was developed in several phases after performing a comprehensive systematic review to identify variation in the conduct, reporting, and method in chatbot health advice studies. Findings from the review were used to develop a draft checklist that was revised through an international, multidisciplinary, modified, asynchronous Delphi consensus process of 531 stakeholders, three synchronous panel consensus meetings of 48 stakeholders, and subsequent pilot testing of the checklist. CHART includes 12 items and 39 subitems to promote transparent and comprehensive reporting of chatbot health advice studies. These include title (subitem 1a), abstract/summary (subitem 1b), background (subitems 2a,b), model identifiers (subitems 3a,b), model details (subitems 4a-c), prompt engineering (subitems 5a,b), query strategy (subitems 6a-d), performance evaluation (subitems 7a,b), sample size (subitem 8), data analysis subitem 9a), results (subitems 10a-c), discussion (subitems 11a-c), disclosures (subitem 12a), funding (subitem 12b), ethics (subitem 12c), protocol (subitem 12d), and data availability (subitem 12e). The CHART checklist and corresponding diagram of the method were designed to support key stakeholders including clinicians, researchers, editors, peer reviewers, and readers in reporting, understanding, and interpreting the findings of chatbot health advice studies.

Key messages

- CHART was developed by performing a systematic review, Delphi consensus of 531 international stakeholders, and several consensus meetings among an expert panel comprised of 48 members.
- The CHART statement outlines 12 key reporting items for chatbot health advice studies in the form of a checklist and methodology diagram.
- All stakeholders including clinicians, researchers, and journal editors should encourage the transparent reporting of chatbot health advice studies.

Introduction

Artificial intelligence (AI) has made great strides toward clinical applications in healthcare, with deep learning algorithms performing comparably to current gold standards in several areas in patient care[1,2]. With the introduction of large language models (LLMs) into mainstream use, there has been a considerable rise in the number of studies evaluating the performance of generative AI-driven chatbots in summarizing evidence and providing health advice[3], termed chatbot health advice (CHA) studies. Investigators typically develop prompts to query generative AI models through a chat-based interface for the purpose of summarizing clinical evidence or obtaining health advice including, but not limited to, health promotion, prevention, screening, diagnosis, treatment, and/or general health information. For example, physicians may query generative AI-driven chatbots to identify whether their patient should receive colorectal cancer screening[4]. Similarly, a patient may ask questions about their upcoming surgery for gastro-oesophageal reflux disease[5]. The intense interest in using generative AI-driven chatbots for health advice has generated numerous CHA studies in a short timeframe[6]. Investigators may include clinicians, scientists, or patients, bringing different technical expertise and personal perspectives to study methodology including prompt engineering and model response evaluation.

These studies represent a growing genre of medical AI research[7]. At least 137 CHA studies were published less than a year after the release of ChatGPT in November 2022, but the completeness of reporting among these studies has been highly variable[6]. For instance, few articles elaborate on the development of their prompts, while fewer than 40% of articles report key elements of

Received: June 15, 2025. **Accepted:** June 18, 2025

Table 1 Glossary

Term	Definition
Artificial intelligence (AI)	The science of developing computer systems that can perform complex tasks approximating human cognitive performance.
Base model	A pre-existing generative AI model.
Chat session	An interface in a computing device through which communication takes place between a chatbot and its user through text-based prompts.
Chatbot health advice study (CHAS)	Any research study evaluating the performance of chatbots when summarizing health evidence and/or providing clinical advice.
Fine-tuned model	A base model that has been manipulated through various methods of algorithmic tuning to alter its performance including, but not limited to, reinforcement learning or retrieval-augmented generation (RAG).
Generative AI-driven chatbot	A program that permits users to interact with an algorithm (such as an LLM) designed to respond to user prompts.
Ground truth	The reference standard, or criteria, on which the model is evaluated to define successful performance.
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. LLMs are a subset of generative AI models.
Multimodal LLM	LLMs with the capacity to integrate input from various data types including text speech and/or visual sources.
Natural language processing (NLP)	A branch of information science that seeks to enable computers to interpret and manipulate human text.
Next word prediction	The natural language processing task of predicting the next word in a sequence of text given context and model parameters.
Novel model	A novel base model.
Parameter	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).
Post-implementation/deployment	Refers to alteration of the generative AI model following its release for user accessibility.
Pre-implementation/deployment	Refers to alteration of the generative AI model prior to its release for user accessibility.
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.
Query	The act of communicating with a generative AI-driven chatbot by inputting a prompt into the chatbot, which might be a question comment, or phrase, to elicit specific desired outputs from the generative AI model.
Response	The output of the generative AI-driven chatbot.
Tuned model	A base model that has been altered to provide focused responses by means other than fine-tuning.
Zero shot	A machine learning paradigm in which the task (such as classification) is performed without explicit training of data (or classes).

their query strategy including the date of their search, the number of chat sessions used, or the number of prompts[6]. Raw prompts and model output are infrequently reported, and most articles present an insufficient amount of information to identify the model and chatbot under evaluation[6]. This problem is important because inadequate reporting impairs the ability of readers to interpret the validity and reliability of study findings[8]. Flaws in the design, data collection, or conduct of a study may lead to erroneous conclusions or raise the risk of patient harm, particularly if generative AI-driven models are used for health purposes[9]. Complete and standardized reporting facilitates critical appraisal, and may help to identify applications with genuine potential to improve healthcare, building trust in the use of generative AI-models in clinical practice among clinicians, patients, and the general public[9].

In response to the growing need for reporting standards for evaluating CHA studies for clinical purposes[10], we developed the Chatbot Assessment Reporting Tool (CHART). This reporting standard is an international, multidisciplinary initiative registered with the Enhancing the QUAlity and Transparency Of health Research (EQUATOR) Network[11], and was announced in December 2023[3]. This article describes the methods used to identify, evaluate, and gain consensus on the checklist items and diagram that comprise CHART. We aimed to develop robust guidance to promote high methodological rigour and transparent reporting of CHA studies evaluating the performance of generative-AI driven chatbots when summarizing clinical evidence and providing health advice. The terminology used in this reporting guideline is listed in *Table 1*.

Methods

We formed a steering group responsible for overseeing the development of CHART. We developed CHART in alignment with the EQUATOR Network's framework according to the highest methodological standards for reporting guideline development[8], and published the protocol in May 2024[7].

To inform the development of CHART, we conducted a comprehensive systematic review to identify information reported in CHA studies. The review protocol was prospectively registered on the Open Sciences Framework: https://osf.io/cxsk3. The systematic review was devised according to methodological guidance from the Joanna Briggs Institute[12]. A systematic literature search was performed with the support of a health sciences librarian using Medline via Ovid, Embase via Elsevier, and Web of Science on 27 October 2023. A full search syntax from all database searches is provided in the *supplementary*

section of our systematic review[6]. We screened 7752 articles to identify 137 eligible articles of interest. Considerable variation in methodology and reporting was observed, and we identified 120 candidate checklist items for CHART (Appendix S1). Full details on this process can be found in our protocol[7]. To evaluate these candidate checklist items for inclusion in the CHART checklist, we invited an advisory committee to perform a modified Delphi consensus process and formed an expert panel to conduct synchronous consensus meetings. Full details on this recruitment process can be found in the protocol[7]. We considered 'experts' as individuals who have made important contributions academically to their discipline, with an emphasis on individuals that have participated in reporting guideline development previously.

Modified Delphi consensus survey

The steering group invited 1043 members globally to form an advisory committee to participate in a Delphi survey, comprising clinicians, epidemiologists, research methodologists, generative AI researchers, journal editors, chatbot researchers, ethicists, regulatory experts, policy experts, and patient partners. We identified potential committee members using a multipronged approach through co-authors published in the top medical journals, public and internal calls through affiliate journals, as well as through snowballing via all members of our expert panel. To identify the top ten journals across all specialties, we used the journal ranking feature in Scimago. Full details are listed in our protocol[6]. Via convenience sampling, we included four editors from the top journals identified. We invited members by e-mail and provided project details, as well as our correspondence article and study protocol[3,7]. Members voluntarily registered to participate in our Delphi consensus survey by providing basic demographic information, as well as details surrounding their prior research experience and content expertise. We presented candidate checklist items to the advisory committee using the online Delphi consensus platform Welphi, *Decision Eyes* (www.welphi.com). Members rated candidate checklist items as one of the following: 'include', 'maybe include', 'uncertain', 'maybe exclude', or 'exclude'. They also suggested additional checklist items. After the first round of voting, advisory committee members engaged in a second round of voting via a modified Delphi consensus survey. Members were able to view the results from the first round and review comments supporting voting considerations. During the second Delphi round, members voted on the same checklist items, as well as any additional checklist items from the first round. Advisory committee members were also able to suggest additional checklist items during the second round, generating a total of 28 additional candidate checklist items across both Delphi rounds. A total of 531 of 1043 (50.9%) members participated in both Delphi Consensus rounds, rating a total of 140 candidate checklist items for review by the expert panel (Appendix S1).

Expert panel consensus

The steering group assembled an international, multidisciplinary panel comprising a balanced representation of 48 relevant stakeholders including clinicians, statisticians, research methodologists, reporting guideline developers, generative AI researchers, journal editors, chatbot researchers, ethicists, regulatory experts, policy experts, and four patient partners. The distribution of stakeholders among the panel is presented in the supplementary material. The steering group used a prespecified threshold of 80% agreement for inclusion to show majority consensus based on prior work[7,13]. We identified items with at least 80% consensus with the selection of either 'include' and 'maybe include' together, or 'exclude' and 'maybe exclude' and posed to the panel whether to include or exclude suggested items. Items not meeting 80% consensus were posed to the panel for further discussion. We also presented raw scores including absolute and relative and frequencies to the expert panel to support their interpretation and decision-making. We held synchronous discussions over three separate panel consensus meetings on Zoom spanning over 12 collective hours on 30 June, 5 August, and 2 September 2024. Items on which the expert panel disagreed with the advisory committee, as well as items rated as 'uncertain' by the advisory committee, were discussed among panel members until consensus was reached. Panel members were able to suggest changes to the phrasing of checklist items, as well as suggest additional checklist items. After extensive discussion, the expert panel reached consensus on 12 checklist items (Appendix S2) and nine abstract checklist items (Appendix S3). A fillable methodological diagram can be found in Appendix S4. A list of panel members can be found in Appendix S5. No items or subitems required voting, as contentious items were discussed thoroughly until consensus was achieved.

Pilot testing

Following the panel consensus meetings, draft checklist items were presented to authors of separate, prior CHA studies via an iterative process for pilot testing. Groups of five authors used the draft CHART checklist to evaluate ten published CHA studies and provide feedback in each round until saturation was reached with respect to no new comments or areas for improvement. Pilot testers were provided with feedback from each round of testing to inform their evaluations. Authors were physicians or CHA study researchers and were not affiliated with the articles under evaluation. We instructed pilot testers to flag any item or subitem that they perceived as unclear or inappropriate for further assessment by the steering group and re-evaluation by the panel if needed. However, we received positive feedback regarding the length, content, and user experience with the checklist. No items or subitems were flagged as inappropriate. Minor changes were made to the checklist including the phrasing of items, the order of items, and the formatting of the fillable document to optimize user experience with the checklist. No additional items or subitems were suggested. Saturation was reached after two rounds of pilot testing. Full details regarding our methodology can be found in our research protocol[7].

Deviations from the protocol

Based on feedback from the multidisciplinary expert panel, we broadened the scope beyond LLMs to include any applications using generative AI due to the dynamically evolving nature of AI research in medicine. Moreover, two expert subgroups were assembled after the panel reviewed the candidate checklist items after the first consensus meeting. First, an expert generative AI subgroup met to evaluate and revise the terminology and checklist items used in this reporting guideline. Second, an expert data analysis subgroup reviewed checklist items related to statistical analysis. The results of both subgroups were presented to the expert panel and were reviewed for approval and discussed at subsequent panel consensus meetings. Finally, due to the complex nature of the conduct and reporting of CHA studies, we developed the checklist items and accompanying diagram for CHART over three separate synchronous, 4-h panel consensus

meetings, rather than two, as initially planned in our protocol[7]. Further guidance and points of emphasis are detailed in the CHART Explanation & Elaboration article[14].

Results

The CHART methodological diagram can be seen in *Fig. 1*. The CHART checklist consists of 12 items comprising 39 subitems for the complete and transparent reporting of CHA studies. Items relate to title & abstract (item 1), introduction (item 2), methods (items 3–9), results (item 10), discussion (item 11), and open science (item 12).

The Delphi advisory committee and the expert panel both emphasized the importance of several checklist items. Specific examples are highlighted here, but the thorough reporting of all items listed in *Table 2* is recommended. Delphi and panel members both voiced that authors must adequately identify the generative AI model and chatbot that they evaluated (items 3 and 4). This includes model identifiers, whether it is an open source or proprietary model, and whether the model was novel or a base model (*Table 2*). Our expert stakeholders further stressed that authors must report the details involved during prompt engineering as well as the query strategy applied by investigators (items 5 and 6). This information must include the process used to develop prompts, the members of the study team involved, and the dates and locations of queries (*Table 2*). Our panelists also underscored the necessity of explicitly defining a reference standard and describing the performance

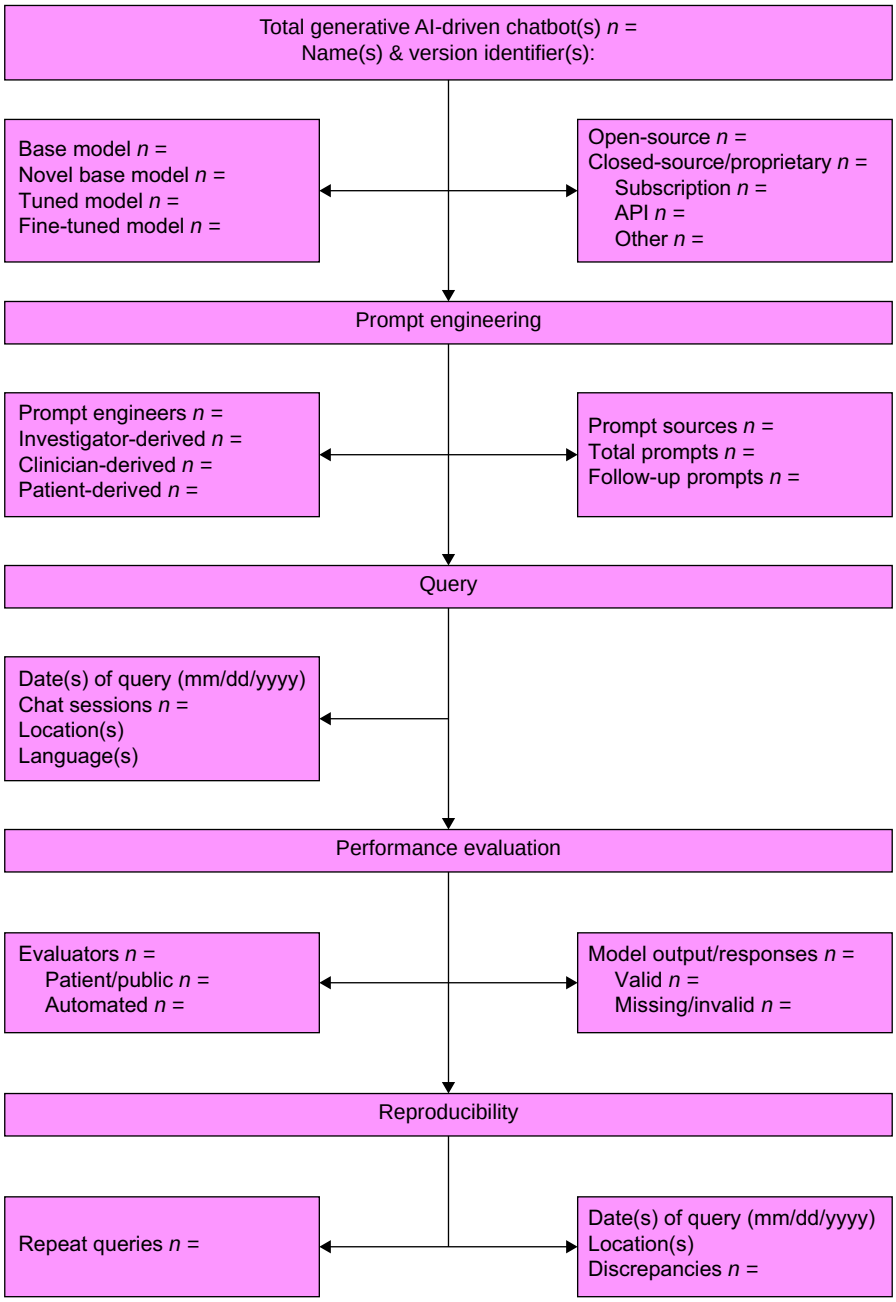

Fig. 1 CHART methodological diagram

AI, artificial intelligence; API, application programming interfaces.

Table 2 CHART checklist

Heading	No.	CHART checklist item	Page no.
Title and abstract			
Title	1a	State that the study is assessing one or more generative AI-driven chatbots for clinical evidence or health advice.	
Abstract/summary	1b	Apply a structured format, if applicable.	
Introduction			
Background	2a	State the scientific background, rationale, and healthcare context for evaluating the generative AI-driven chatbot(s), referencing relevant literature when applicable.	
	2b	State the aims and research questions including the target audience, intervention, comparator(s), and outcome(s).	
Methods			
Model identifiers	3a	State the name and version identifier(s) of the generative AI model(s) and chatbot(s) under evaluation, as well as their date of release or last update.	
	3b	State whether the generative AI model(s) and chatbot(s) are open-source or closed-source/proprietary.	
Model details	4a	State whether the generative AI model was a base model or a novel base model, tuned model, or fine-tuned model.	
	4b	If a base model is used, cite its development in sufficient detail to identify the model.	
	4c	If a novel base model, tuned model, or fine-tuned model is used, describe the pre- and/or post-implementation/deployment data and parameters.	
Prompt engineering	5a	Describe the evolution of study prompt development.	
	5ai	Describe the sources of prompts.	
	5aii	State the number and characteristics of the individual(s) involved in prompt engineering.	
	5aiii	Provide details of any patient and public involvement during prompt engineering.	
	5b	Provide study prompts.	
Query strategy	6a	State route of access to generative AI model.	
	6b	State the date(s) and location(s) of queries for the generative AI-driven chatbot(s) including the day, month, and year, as well as city and country.	
	6c	Describe whether prompts were input into separate chat session(s).	
	6d	Provide all generative AI-driven chatbot output/responses.	
Performance evaluation	7a	Define the ground truth or reference standard used to define successful generative AI-driven chatbot performance.	
	7b	Describe the process undertaken for generative AI-driven chatbot performance evaluation.	
	7bi	State the number and characteristics of team members involved in performance evaluation.	
	7bii	Provide details of any patients and public involvement during the evaluation process.	
	7biii	State whether evaluators were blinded to the identity of the generative AI-driven chatbot(s) under assessment.	
Sample size	8	Report how the sample size was determined.	
Data analysis	9a	Describe statistical analysis methods including any evaluation of reproducibility of generative AI-driven chatbot responses.	
	9ai	Report the measures used for performance evaluation.	
Results			
	10a	Report the performance evaluation undertaken including the alignment between generative AI-driven chatbot output and ground truth or reference standard using quantitative or mixed methods approaches as applicable.	
	10b	For responses deviating from the ground truth or reference standard, state the nature of the difference(s).	
	10c	Report the evaluation for potentially harmful, biased, or misleading responses.	
Discussion			
	11a	Interpret study findings in the context of relevant evidence.	
	11b	Describe the strengths and limitations of the study.	
	11c	Describe the potential implications for practice, education, policy, regulation, and research.	
Open science			
Disclosures	12a	Report any relevant conflicts of interest for all authors.	
Funding	12b	Report sources of funding and their role in the conduct and reporting of the study.	
Ethics	12c	Describe the process undertaken for ethical approval.	
	12ci	Describe the measures taken to safeguard data privacy of patient health information, as applicable.	
	12cii	State whether permission/licensing was obtained for the use of original, copyrighted data.	
Protocol	12d	Provide a study protocol.	
Data availability	12e	State where study data, code repository, and model parameters can be accessed.	

AI, artificial intelligence.

evaluation process (item 7). Stakeholders emphasized the importance of providing a sample size, which includes the number of independent responses from one or more generative AI-driven chatbots. Panelists also identified that the sample size of training data points may also be relevant if authors evaluate a novel or tuned model. Additionally, panelists stressed the importance of reporting the training data used, the ethical approval process undertaken, measures to safeguard the privacy of patient data, the permission or licensing obtained for the use of training data, and whether the training data can be accessed (item 12) (*Table 3*).

Discussion

CHART was developed in accordance with the highest methodological standards through a comprehensive systematic

Table 3 CHART abstract checklist

Heading	CHART checklist no.	Item	Page no.
Background	2a	State the scientific background, rationale, and healthcare context for evaluating the generative AI-driven chatbot(s), referencing relevant literature when applicable.	
	2b	State the aims and research questions including the target audience, intervention, comparator(s), and outcome(s).	
Methods			
Model identifiers	3a	State the name and version identifier(s) of the generative AI model(s) and chatbot(s) under evaluation, as well as their date of release or last update.	
	3b	State whether generative AI model(s) and chatbot(s) are open-source *versus* closed-source/proprietary.	
Model details	4a	State whether the generative AI model was a base model or a novel base model, tuned model, or fine-tuned model.	
Prompt engineering	5a	Describe the evolution of study prompt development.	
	5ai	Describe the sources of prompts.	
	5aii	State the number and characteristics of the individual(s) involved in prompt engineering.	
	5aiii	Provide details of any patient and public involvement during prompt engineering.	
Query strategy	6a	State route of access to generative AI model.	
	6b	State the date(s) and location(s) of queries for the generative AI-driven chatbot(s) including the day, month, and year, as well as city and country.	
Performance evaluation	7a	Define the ground truth or reference standard used to define successful generative AI-driven chatbot performance.	
	7b	Describe the process undertaken for the performance evaluation of the generative AI-driven chatbot(s).	
Sample size	8	Report how the sample size was determined.	
Data analysis	9a	Describe statistical analysis methods including any evaluation of reproducibility of generative AI-driven chatbot responses.	
Results			
	10a	Report the alignment between generative AI-driven chatbot output and ground truth or reference standard using quantitative or mixed methods approaches as applicable.	

AI, artificial intelligence.

review of CHA studies, a modified asynchronous Delphi process conducted by an international, multidisciplinary advisory committee, and three synchronous international, multidisciplinary expert panel consensus meetings[7]. Detailed rationales for each subitem are described in our explanation & elaboration article[14]. The CHART checklist outlines essential items for the reporting of CHA studies, which typically evaluate the performance of generative AI-driven chatbots when summarizing clinical evidence or providing health advice. At the time of writing, substantial advancements are being made in other forms of generative AI such as large multimodal models (LMMs), to which our reporting checklist—developed in the context of studies evaluating LLM performance—may not fully apply[15]. Thus, due to the rapidly evolving nature of these studies, a dynamic process must be in place for the monitoring and updating of this reporting guideline[16].

Applicability and scope

The CHART checklist applies to CHA studies where generative AI-driven chatbots are queried and their responses are reported and evaluated. The CHART checklist does not apply to CHA studies applying randomization techniques (randomized controlled trials), nor to studies that follow patients over time (prospective cohort studies). Future CHART extensions of relevant checklists for various study designs are planned, but in the interim authors are encouraged to apply both the CHART checklist and relevant reporting guidelines according to the appropriate study design such as CONSORT or STROBE[17,18]. Authors using applications in the field of AI more broadly (but not generative AI) are encouraged to use more generic reporting guidelines[13,19,20]. Authors using generative AI models for medical writing are encouraged to apply the CANGARU reporting guidelines, which are in development[21]. CHART applies to the current landscape of CHA studies, and will evolve as a living reporting guideline.

How to use CHART

We suggest that authors use the CHART checklist early in the writing of CHA studies to ensure all items in the checklist have been reported somewhere in their manuscript. Many of the recommendations in the CHART checklist have a natural order and sequence in a CHA study, but some may not. We do not prescribe a specific format or dictate where each individual reporting recommendation should appear in a CHA study, because this order might also depend on journal formatting policies. A downloadable and editable checklist can be found in the *supplementary material*. Authors are recommended to complete the checklist indicating the page number where each subitem has been reported. The completed checklist can then be submitted alongside the CHA study manuscript. A detailed explanation and elaboration paper accompanies the CHART checklist and explains why the reporting of each item is recommended[14].

Copyright protections and fair use doctrine

The accuracy of LLMs is significantly influenced by the nature of the data on which they were trained[10,22]. This principle is the first of four according to the fair use doctrine, which are addressed throughout the CHART checklist as they relate to CHA studies. The first principle refers to the purpose and character of use of the model[11]. The second principle is the nature of the original training data[10,23]. While many LLMs will be trained on non-medical data, it is essential that factual, evidence-based information must be prioritized in the healthcare setting[10]. The third principle pertains to the amount and substantiality of original material used to train the

generative AI model[10], and clarity regarding the origin of training data and permission or license to use content or data protected by copyright is recommended. Finally, the fourth principle relates to the impact on original work, where generative AI models may be trained with copyrighted data[10]. We address these principles in the CHART checklist by encouraging authors to state the purpose of the study, and whether they are evaluating a pre-existing base model, rather than one that is a novel base model, a tuned model, or a fine-tuned model (items 3 and 4). The CHART checklist promotes open science practices and calls for authors to share their code and training data sets to optimize transparency and mitigate uncertainty over data provenance (item 12e). The CHART checklist further uses an evidence-based approach by encouraging authors to state the source of their prompts, their definition of successful model/bot performance, and the process behind performance evaluation (items 5 and 7). The CHART checklist recommends that authors state whether permission or license was obtained by investigators for use of the original work (item 12cii). Readers may also identify the presence of copyrighted data as authors share their coding and training data (item 12e).

Bias and patient safety

In the setting of model development, the outputs of generative AI models such as LLMs are further impacted by the presence of bias in their training data sets[10]. This introduces the risk of LLMs producing misleading or harmful information when applied for the purposes of patient care. These biases may pertain to many factors including, but not limited to, race or ethnicity, sex or gender, language, and culture[24,25]. This risk further highlights the importance of the open science checklist item (item 12) in CHART because the risk of bias from data used to develop LLM-driven chatbots may be identified and/or mitigated by open coding and training data sharing[25]. Furthermore, data used to train generative AI models may pose a threat to data security and patient privacy. The use of identifiable patient data during model training is of particular concern, as sensitive information may be inadvertently disclosed in the absence of appropriate data security measures[10,26]. The risk of data breaches must be met accordingly with robust cybersecurity measures[10]. This concept underscores the importance of the CHART checklist item related to steps taken to ensure safeguarding of patient health information (item 12ci). The push for clinically integrating generative AI models necessitates human oversight of the ethical and safe inclusion of patients and their health information to provide guidance for the safe conduct of CHA studies[27,28]. Although we recognize the importance of making advancements by including patients in CHA studies to develop more patient-centered studies (items 5biii and 7bii), we encourage authors to report whether ethics approval was obtained in these instances for the responsible conduct of their study (item 12c).

Monitoring and updates

This reporting guideline will follow and adapt the traditional methodology for a living clinical practice guideline[16]. The update interval for this reporting guideline will apply to individual checklist items, rather than the entire guideline[16]. Core members of the steering group will perform a systematic search of the literature to continuously survey the literature per living guideline best practices[16], and will meet to discuss any relevant developments in the generative AI field every 6 months for the first 2 years (until 2026). If important changes occur sooner, the group will meet ad hoc as needed. The timing for monitoring and updating the guideline will be reviewed and revised at the time of the next reporting guideline update or by the end of 2026, whichever occurs sooner.

Furthermore, a living expert panel consisting of 14 expert panel members was selected following the third expert panel consensus meeting in accordance with living guideline best practices[16], and comprised of panel members committed to making themselves available to meet virtually at very short notice[16]. Living expert panel members represent backgrounds stemming from medicine, epidemiology, data science, health research methodology, reporting guideline methodology, and statistics. If no changes to the reporting guideline are warranted within a given year, the living expert panel will be updated with the activities of the core steering group and will be alerted to any relevant literature or topics within generative AI to monitor and be aware of. This update will occur at a minimum of once per year at a meeting between the core members of the steering group and the living expert panel. Finally, living peer reviewers will be selected following the peer review process for the CHART statement and elaboration & explanation articles[16]. They will similarly be provided with an annual update, but will only be contacted if checklist items must be updated. If new candidate checklist items or revisions to existing items are identified by the core members of the steering group, the living expert panel will be convened at its earliest convenience to review the relevant literature. In alignment with living guideline best practices[16], the minimum threshold will be set at 90% agreement among living expert panel members for changing checklist items to mitigate the risk of false positives inherent to frequent updates, while avoiding an excessively high threshold[16]. If applicable, the updated manuscript will be co-published in relevant journals with interest.

Target users and implications for stakeholders

CHART applies to individuals performing and reviewing CHA studies such as study investigators, peer reviewers, and journal editors for academic purposes, as well as the wider readership of CHA studies including clinicians, statisticians, generative AI researchers, regulatory experts, ethicists, research methodologists, policy makers, hospital managers, funders, patients, and the wider public. To promote the transparent reporting of CHA studies, we call for clinical journals to adopt CHART: a comprehensive reporting standard developed with high methodological rigour. The main barrier that we anticipate to CHART uptake is the failure to reach the appropriate audience. Therefore, this reporting guideline will be listed on the EQUATOR Network website, and we will disseminate the publication of this reporting guideline widely. CHART will also be presented at peer-reviewed meetings across various medical specialties to optimize the dissemination and reach of the checklist and accompanying diagram. Finally, we will develop a website to house fillable versions of the abstract checklist, the full checklist, and the methodological diagram, which can be found in *Appendices S2–S4* of this publication, to facilitate the application of CHART by CHA researchers.

Following the publication of previous reporting guidelines, the reporting quality of applicable studies improves[29,30]. As investigators and journals apply CHART and the completeness of reporting of CHA studies improves, higher quality studies may be produced. Researchers, ethicists, clinicians, and regulators in the clinical generative AI community must then turn toward the validation of generative AI-driven chatbots for

the purposes of providing health advice[10]. This step may include the prioritization of standardized quality validation metrics, clarifying the role of human involvement in validation studies, validation methodology[31], and the reporting of validation results using CHART. Regulators must further look toward data sensitivity and privacy, ensuring that data security measures are put in place by generative AI developers according to risk category[10]. Funders must invest in the development of high-quality benchmarking and validation studies, as well as highly rigorous CHA studies in the context of the healthcare setting of interest. Funders may also encourage applicants to include a research plan in alignment with the CHART checklist. With studies exhibiting greater transparency and improved methodological rigour, clinicians, patients, and the public will develop progressively increased trust in the clinical integration of generative AI-driven chatbots.

Finally, quality appraisal tools do not exist for CHA studies and remain a future area of study. CHART is a reporting guideline, rather than a critical appraisal tool. Still, we hope that attention to CHART's core checklist items will indirectly improve the methodological rigour of studies in this field[32]. As high-quality evidence builds, the path forward for integrating generative AI into the clinical practice environment will become clearer for both hospital managers and policy makers.

Collaborators

Bright Huo (Division of General Surgery, Department of Surgery, McMaster University, Hamilton, Canada), Gary Collins (UK EQUATOR Centre, University of Oxford, Oxford, UK; Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, Botnar Research Centre, University of Oxford, Oxford, UK), David Chartash (Department of Biomedical Informatics and Data Science, Yale University School of Medicine, New Haven, CT, USA), Arun Thirunavukarasu (Nuffield Department of Clinical Neurosciences, Medical Sciences Division, University of Oxford, Oxford, UK), Annette Flanagin (*JAMA* and *JAMA Network*, American Medical Association, Chicago, IL, USA), Alfonso Iorio (Department of Health Research Methods, Evidence, and Impact; Department of Medicine; McMaster University, Hamilton, ON, Canada), Giovanni Cacciamani (USC Institute of Urology and Catherine and Joseph Aresty Department of Urology, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA; Artificial Intelligence Center at USC Urology, USC Institute of Urology, University of Southern California, Los Angeles, CA, USA), Xi Chen (Sports Medicine Centre, West China Hospital, Sichuan University, Chengdu, China; Department of Orthopaedics and Orthopaedic Research Institute, West China Hospital, Sichuan University, Chengdu, China), Nan Liu (Duke-NUS Medical School, National University of Singapore, Singapore, Singapore), Piyush Mathur (Cleveland Clinic, Case Western Reserve University, Cleveland, OH, USA), An Wen Chan (Department of Medicine, Women's College Research Institute, University of Toronto, Toronto, ON, Canada), Christine Laine (Annals of Internal Medicine, American College of Physicians, Philadelphia, PA, USA; American College of Physicians, Philadelphia, PA, USA), Daniela Pacella (Department of Public Health, University of Naples Federico II, Naples, Italy), Michael Berkwits (Office of Science Dissemination, Office of Science, Centers for Disease Control and Prevention, Atlanta, GA, USA), Stavros A Antoniou (Department of General Surgery, Papageorgiou General Hospital, Thessaloniki, Greece), Jennifer C Camaradou (British Psychological Society, University of Plymouth, Plymouth, UK), Carolyn Canfield (Innovation Support Unit, Department of Family Practice, University of British Columbia, Vancouver, BC, Canada), Michael Mittelman (Patient subject matter expert, Independent Cybersecurity Professional), Timothy Feeney (*The BMJ*, London, UK; Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA), Elizabeth Loder (*The BMJ*, London, UK; Department of Neurology, Brigham and Women's Hospital, Boston, MA, USA), Riaz Agha (*International Journal of Surgery*, London, UK; Eworkflow, London, UK), Ashirbani Saha (Department of Oncology, McMaster University, Hamilton, ON, Canada), Julio Mayol (Hospital Clinico San Carlos, Instituto de Investigación Sanitaria San Carlos, Facultad de Medicina Universidad Complutense de Madrid, Spain), Anthony Sunjaya (The George Institute for Global Health; Tyree Institute of Health Engineering, UNSW Engineering; School of Population Health, UNSW Medicine and Health, Sydney, NSW, Australia), Hugh Harvey (Hardian Health, Haywards Heath, UK), Jeremy Y Ng (Centre for Journalology, Ottawa Hospital Research Institute, Ottawa, ON, Canada), Tyler McKechnie (Division of General Surgery, Department of Surgery, McMaster University, Hamilton, ON, Canada), Yung Lee (Division of General Surgery, Department of Surgery, McMaster University, Hamilton, ON, Canada; Digestive Diseases Institute, Cleveland Clinic, Cleveland, OH, USA), Nipun Verma (Postgraduate Institute of Medical Education and Research, Chandigarh, India), Gregor Stiglic (University of Maribor, Maribor, Slovenia), Melissa McCradden (Australian Institute for Machine Learning, Adelaide, SA, Australia), Karim Ramji (Phelix AI, Hamilton, ON, Canada), Vanessa Boudreau (Division of General Surgery, Department of Surgery, McMaster University, Hamilton, ON, Canada), Monica Ortenzi (Università Politecnica delle Marche, Clinica di Chirurgia Generale e d'Urgenza), Joerg Meerpohl (Institute for Evidence in Medicine, Medical Centre and Faculty of Medicine, University of Freiburg, Germany; Cochrane Germany, Cochrane Germany Foundation, Freiburg, Germany), Per Olav Vandvik (Cochrane Germany, Cochrane Germany Foundation, Freiburg, Germany; MAGIC Evidence Ecosystem Foundation, Oslo, Norway), Thomas Agoritsas (Department of Health Research Methods, Evidence, and Impact; Department of Medicine; McMaster University, Hamilton, ON, Canada; MAGIC Evidence Ecosystem Foundation, Oslo, Norway; University Hospitals of Geneva, Geneva, Switzerland), Diana Samuel (*The Lancet Digital Health*, London, UK), Helen Frankish (*The Lancet*, London, UK), Michael Anderson (NIHR clinical lecturer, Health Organisation, Policy, Economics (HOPE), Centre for Primary Care and Health Services Research, University of Manchester, Manchester, UK; LSE Health, London School of Economics and Political Science, London, UK), Xiaomei Yao (Department of Oncology, McMaster University, Hamilton, ON, Canada), Stacy Loeb (New York University Langone Health, New York, NY, USA), Cynthia Lokker (Department of Health Research Methods, Evidence, and Impact; Department of Medicine; McMaster University, Hamilton, ON, Canada), Xiaoxuan Liu (College of Medicine and Health, University of Birmingham, Birmingham, UK), Eliseo Guallar (School of Global Public Health, New York University, New York, NY, USA), Gordon Guyatt (Department of Health Research Methods, Evidence, and Impact; Department of Medicine; McMaster University, Hamilton, ON, Canada; MAGIC Evidence Ecosystem Foundation, Oslo, Norway).

Funding

The Chatbot Assessment Reporting Tool (CHART) was funded by the *First Cut* competition and the Postgraduate Medical Education Committee at McMaster University. The funders had no role in considering the study design or in the collection, analysis, interpretation of data, writing of the report, or decision to submit the article for publication.

Acknowledgements

The authors thank the *First Cut* competition organizers and the Postgraduate Medical Education Committee at McMaster University for financially supporting the development of this project. The authors would also like to thank members of the advisory committee for their invaluable time and effort devoted to the development of the Chatbot Assessment Reporting Tool.

Author contributions

B.H., D.C., A.-W.C., C.L., and G.G. contributed to the conception and design of the work. All authors contributed to data analysis, interpretation, manuscript drafting, revising, and approve of the final version to be published. B.H. acted as guarantor and is responsible for the overall content (as guarantor), and all authors are accountable for the accuracy of the checklist and methodological diagram. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

Disclosure

All authors have completed the ICMJE uniform disclosure form at www.icmje.org/disclosure-of-interest/ and declare support from McMaster University for the submitted work. G.S.C. is a National Institute for Health and Care Research (NIHR) Senior Investigator. The views expressed in this article are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care. A.J.T. has received funding from HealthSense to investigate evidence-based medicine applications of large language models. P.M. is the co-founder of BrainX. A.S. has received research funding from the Australian government and is co-founder of BantingMed Pty. D.S. is the Acting Deputy Editor for the *Lancet Digital Health*. M.M. has received research funding from The Hospital Research Founding Group. T.F. sits on the executive committee of MDEpiNet. H.F. is a Senior Executive Editor for *The Lancet*. C.L. is the Editor-in-Chief of *Annals of Internal Medicine*. A.F. is Executive Managing Editor and Vice President, Editorial Operations, at *JAMA* and the *JAMA Network*. T.F. and E.L. are journal editors for *The BMJ*. R.A. is the Editor-in-Chief of *International Journal of Surgery*. G.S. is an Executive Editor of *Artificial Intelligence in Medicine*. S.L. is a paid consultant for Astellas. D.P. has received research funding from the Italian Ministry of University and Research. M.O. is a paid consultant for Theator. T.A., P.O.V. and G.G. are board members of the MAGIC Evidence Ecosystem Foundation (www.magicproject.org), a not-for-profit organization that conducts research and evidence appraisal and guideline methodology and implementation, and provides authoring and publication software (MAGICapp) for evidence summaries, guidelines, and decision aids.

Supplementary material

Supplementary material is available at *BJS* online.

Ethics approval

Ethics approval was submitted to and waived by the Hamilton Integrated Research Ethics Board (HiREB #17025).

Data availability

All data relevant to the study are included in the article or uploaded as *supplementary material*.

References

1. Kolbinger FR, Veldhuizen GP, Zhu J, Truhn D, Kather JN. Reporting guidelines in medical artificial intelligence: a systematic review and meta-analysis. *Commun Med (Lond)* 2024;**4**:71
2. Han R, Acosta JN, Shakeri Z, Ioannidis JPA, Topol EJ, Rajpurkar P. Randomised controlled trials evaluating artificial intelligence in clinical practice: a scoping review. *Lancet Digit Health* 2024;**6**: e367–e373
3. Huo B, Cacciamani GE, Collins GS, McKechnie T, Lee Y, Guyatt G. Reporting standards for the use of large language model-linked chatbots for health advice. *Nat Med* 2023;**29**:2988–2988
4. Huo B, McKechnie T, Ortenzi M, Lee Y, Antoniou S, Mayol J *et al.* Dr. GPT will see you now: the ability of large language model-linked chatbots to provide colorectal cancer screening recommendations. *Health Technol (Berl)* 2024;**14**:463–469
5. Huo B, Marfo N, Sylla P, Calabrese E, Kumar S, Slater BJ *et al.* Clinical artificial intelligence: teaching a large language model to generate recommendations that align with guidelines for the surgical management of GERD. *Surg Endosc* 2024;**38**: 5668–5677
6. Huo B, Boyle A, Marfo N, Tangamornsuksan W, Steen JP, McKechnie T *et al.* Large language models for chatbot health advice studies: a systematic review. *JAMA Netw Open* 2025;**8**: e2457879
7. CHART Collaborative. Protocol for the development of the Chatbot Assessment Reporting Tool (CHART) for clinical advice. *BMJ Open* 2024;**14**:e081155
8. Moher D, Schulz KF, Simera I, Altman DG. Guidance for developers of health research reporting guidelines. *PLoS Med* 2010;**7**:e1000217
9. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B *et al.* TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* 2024;**385**:e078378
10. Ong JCL, Chang SYH, William W, Butte AJ, Shah NH, Chew LST *et al.* Ethical and regulatory challenges of large language models in medicine. *Lancet Digit Health* 2024;**6**:e428–e432
11. Altman DG, Simera I, Hoey J, Moher D, Schulz K. EQUATOR: reporting guidelines for health research. *Open Med* 2008;**2**:e49–e50
12. Munn Z, Peters MDJ, Stern C *et al.* Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. *BMC Med Res Methodol* 2018;**18**:143. doi:10.1186/s12874-018-0611-x
13. Liu X, Rivera SC, Moher D, Calvert MJ, Denniston AK. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *BMJ* 2020; **370**:m3164
14. The CHART Collaborative. Reporting guidelines for chatbot health advice studies: explanation and elaboration for the Chatbot Assessment Reporting Tool (CHART). *BMJ* 2025;**390**:e083305
15. Yin S, Fu C, Zhao S, Li K, Sun X, Xu T *et al.* A survey on multimodal large language models. *Natl Sci Rev* 2024;**11**:nwae403

16. Akl EA, Meerpohl JJ, Elliott J, Kahale LA, Schünemann HJ, Agoritsas T *et al.* Living systematic reviews: 4. Living guideline recommendations. *J Clin Epidemiol* 2017;**91**:47–53
17. Begg C, Cho M, Eastwood S, Horton R, Moher D, Olkin I *et al.* Improving the quality of reporting of randomized controlled trials. The CONSORT statement. *JAMA* 1996;**276**:637–639
18. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. Strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. *BMJ* 2007;**335**: 806–808
19. Rivera SC, Liu X, Chan AW, Denniston AK, Calvert MJ. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *BMJ* 2020;**370**:m3210
20. Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S *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
21. Cacciamani GE, Collins GS, Gill IS. ChatGPT: standard reporting guidelines for responsible use. *Nature New Biol* 2023;**618**:1. doi: 10.1038/d41586-023-01853-w
22. Xie SM, Pham H, Dong X, Du N, Liu H, Lu Y *et al.* DoReMi: optimizing data mixtures speeds up language model pretraining. arXiv:2305.10429. doi:10.48550/arXiv.2305.10429
23. Ng FYC, Thirunavukarasu AJ, Cheng H, *et al.* Artificial intelligence education: an evidence-based medicine approach for consumers, translators, and developers. *Cell Rep Med* 2023; **17**:101230. doi:10.1016/j.xcrm.2023.101230
24. Li H, Moon JT, Purkayastha S, Celi LA, Trivedi H, Gichoya JW. Ethics of large language models in medicine and medical research. *Lancet Digit Health* 2023;**5**:e333–e335
25. The Lancet Digital Health. Large language models: a new chapter in digital health. *Lancet Digit Health* 2024;**6**:e1
26. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. *Nat Med* 2023;**29**: 1930–1940
27. Haltaufderheide J, Ranisch R. The ethics of ChatGPT in medicine and healthcare: a systematic review on large language models (LLMs). *NPJ Digit Med* 2024;**7**:183
28. Thirunavukarasu AJ. Large language models will not replace healthcare professionals: curbing popular fears and hype. *J R Soc Med* 2023;**116**:181–182
29. Kane RL, Wang J, Garrard J. Reporting in randomized clinical trials improved after adoption of the CONSORT statement. *J Clin Epidemiol* 2007;**60**:241–249
30. Turner L, Shamseer L, Altman DG, Schulz KF, Moher D. Does use of the CONSORT statement impact the completeness of reporting of randomised controlled trials published in medical journals? A Cochrane review. *Syst Rev* 2012;**1**:60
31. de Hond A, Leeuwenberg T, Bartels R, van Buchem M, Kant I, Moons KGM *et al.* From text to treatment: the crucial role of validation for generative large language models in health care. *Lancet Digit Healt* 2024;**6**:e441–e443
32. Logullo P, MacCarthy A, Kirtley S *et al.* Reporting guideline checklists are not quality evaluation forms: they are guidance for writing. *Health Sci Rep* 2020;**3**:e165. doi:10.1002/ hsr2.165