JAMA Network Open

Special Communication | Medical Journals and Publishing

Reporting Guideline for Chatbot Health Advice Studies
The CHART Statement

The CHART Collaborative

Abstract

IMPORTANCE The rise in chatbot health advice (CHA) studies is accompanied by heterogeneity in reporting standards, impacting their interpretability.

OBJECTIVE To provide reporting recommendations for studies evaluating the performance of generative artificial intelligence (AI)–driven chatbots when summarizing clinical evidence and providing health advice.

DESIGN, SETTING, AND PARTICIPANTS CHART was developed in several phases after performing a comprehensive systematic review to identify variation in the conduct, reporting, and methodology in CHA 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, 3 synchronous panel consensus meetings of 48 stakeholders, and subsequent pilot testing of the checklist.

RESULTS CHART includes 12 items and 39 subitems to promote transparent and comprehensive reporting of CHA studies. These include title (subitem 1a), abstract or summary (subitem 1b), background (subitems 2ab), model identifiers (subitem 3ab), model details (subitems 4abc), prompt engineering (subitems 5ab), query strategy (subitems 6abcd), performance evaluation (subitems 7ab), sample size (subitem 8), data analysis (subitem 9a), results (subitems 10abc), discussion (subitems 11abc), disclosures (subitem 12a), funding (subitem 12b), ethics (subitem 12c), protocol (subitem 12d), and data availability (subitem 12e).

CONCLUSIONS AND RELEVANCE The CHART checklist and corresponding methodological diagram were designed to support key stakeholders including clinicians, researchers, editors, peer reviewers, and readers in reporting, understanding, and interpreting the findings of CHA studies.

JAMA Network Open. 2025;8(8):e2530220. doi:10.1001/jamanetworkopen.2025.30220

+ Supplemental content

Author affiliations and article information are listed at the end of this article.

Introduction

Artificial intelligence (AI) has made great strides toward clinical applications in health care, with deep learning algorithms performing comparably with current reference standards in several areas in patient care.[1,2] With the introduction of large language models (LLMs) into mainstream use,[3] there has been an explosive rise in the number of studies evaluating the performance of generative artificial intelligence (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 gastroesophageal 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 studies elaborate on the development of their prompts, while fewer than 40% of studies report key elements of 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 studies 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 health care, building trust in the use of generative AI models in clinical practice among clinicians, patients, and the 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 study describes the methodology used to identify, evaluate, and gain consensus on the checklist items and diagram that are included in CHART. We aimed to develop robust guidance to promote high methodological rigor 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] Ethics approval was submitted to and waived by the Hamilton integrated research ethics board.

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.[12] The systematic review was devised according to methodological guidance from the Joanna Briggs Institute.[13] 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 October 27, 2023. Full search syntax from all database searches are provided in the supplementary section of our systematic review.[6] We screened 7752 studies to identify 137 eligible studies of interest. Considerable variation in methodology and reporting was observed, and we identified 120 candidate checklist items for CHART (eTable 1 in Supplement 1 and Supplement 2). 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 to be individuals who have made important contributions academically to their discipline, with an emphasis on individuals who 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, including 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 10 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 4 editors from the top journals identified. We invited members by email 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 of 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).[14] 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

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 preexisting 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 (CHA) study	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 with specificity; methods include but are not limited to reinforcement learning or distillation.
Generative AI–driven chatbot	A program that permits users to interact with an AI model (such as an LLM) that is 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 AI model including 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.
Parameter	A variable that is tuned iteratively or 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).
Postimplementation/deployment	Refers to alteration of the generative AI model following its release.
Pre-implementation/deployment	Refers to alteration of the generative AI model before its release.
Prompt	The input provided by users when interfacing with a generative AI–driven chatbot, leading to input interaction with the AI model.
Prompt engineering	An iterative testing phase where various pieces of text are input 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, including but not limited to retrieval augmented generation, which seeks to alter performance rather than the model.
Zero shot	A machine learning paradigm in which the task (such as classification) is performed without explicit training, fine-tuning, or other optimization.

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 members (50.9%) participated in both Delphi consensus rounds, rating a total of 140 candidate checklist items for review by the expert panel (eTable 1 in Supplement 1 and Supplement 2).

Expert Panel Consensus

The steering group assembled an international, multidisciplinary panel including 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 4 patient partners. The distribution of stakeholders among the panel is presented in eTable 4 in Supplement 1 and Supplement 2. The steering group used a prespecified threshold of 80% agreement for inclusion to show majority consensus based on prior work.[7,15] 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 frequencies to the expert panel to support their interpretation and decision-making. We held synchronous discussions over 3 separate panel consensus meetings on Zoom spanning over 12 collective hours on June 30, August 5, and September 2, 2024. Items on which the expert panel disagreed with the advisory committee as well as items voted as *unsure* 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 (eTable 2 in Supplement 1 and Supplement 2) and 9 abstract checklist items (eTable 3 in Supplement 1 and Supplement 2). A fillable methodological diagram can be found in the eFigure in Supplement 3. A list of panel members can be found in eTable 4 in Supplement 1 and Supplement 2. 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 5 authors used the draft CHART checklist to evaluate 10 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 studies 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 reevaluation 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 2 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 expert multidisciplinary 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, 2 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 3 separate synchronous, 4-hour panel consensus meetings rather than 2, as initially planned in our protocol.[7] Further guidance and points of emphasis are detailed in the CHART explanation and elaboration article.[16]

Results

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

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 they evaluated (items 3 and 4). This

Figure. The CHART Methodological Diagram

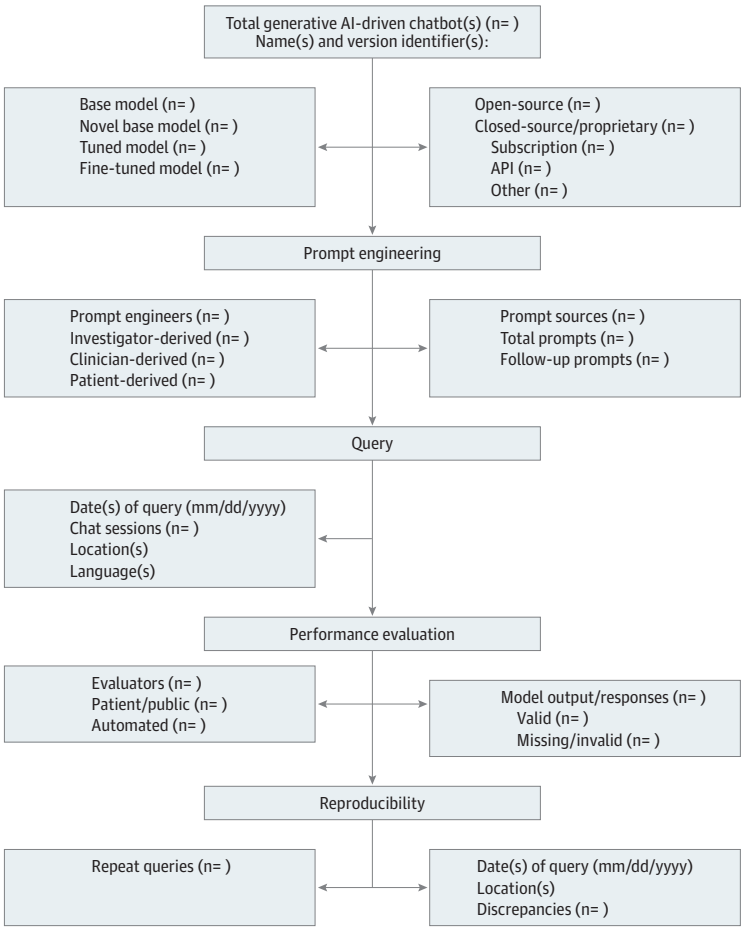

Table 2. CHART Checklist

Heading	No.	Chart checklist item	Page No.
Title and abstract			
Title	1a	State that the study is assessing 1 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 health care 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 postimplementation/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.	

(continued)

Table 2. CHART Checklist (continued)

Heading	No.	Chart checklist item	Page No.
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.	

Table 3. The CHART Abstract Checklist

Heading	CHART checklist No.	Item	Page No.
Background	2a	State the scientific background, rationale, and health care 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 vs 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.	

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 includes 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 evaluation process (item 7). Stakeholders emphasized the importance of providing a sample size, which includes the number of independent responses from 1 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 review of CHA studies, a modified asynchronous Delphi process conducted by an international, multidisciplinary advisory committee, and 3 synchronous international, multidisciplinary expert panel consensus meetings.[7] Detailed rationale for each subitem are described in our explanation and elaboration article.[16] 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, to which our reporting checklist—developed in the context of studies evaluating LLM performance—may not fully apply.[17] 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.[18]

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 clinical 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[19] or STROBE.[20] Authors using applications in the field of artificial intelligence more broadly (but not generative AI) are encouraged to use more generic reporting guidelines.[15,21,22] Authors using generative AI models for medical writing are encouraged to apply the CANGARU[23] reporting guidelines, which are in development. 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 spreadsheet can be found in Supplement 2. 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 article accompanies the CHART checklist and explains why the reporting of each item is recommended.[16]

Copyright Protections and Fair Use Doctrine

The accuracy of LLMs is influenced by the nature of the data on which they were trained.[10,24] This principle is the first of 4 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.[10] The second principle is the nature of the original training data.[10,25] While many LLMs will be trained on nonmedical data, it is essential that factual, evidence-based information must

be prioritized in the health care 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 preexisting base model rather than one that is a novel base model, a tuned model, or a fine-tuned model (items 3-4). The CHART checklist promotes open science practices and calls for authors to share their code and training datasets 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 and/or 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 output of generative AI models, such as LLMs, is further impacted by the presence of bias in their training datasets.[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.[26,27] 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 or mitigated by open coding and training data sharing.[27] 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.[10,28] 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.[29,30] 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.[18] The update interval for this reporting guideline will apply to individual checklist items, rather than the entire guideline.[18] Core members of the steering group will perform a systematic search of the literature to continuously survey the literature per living guideline best practices,[18] 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,[18] and it includes panel members committed to making themselves available to meet virtually at very short notice.[18] 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 and explanation publication.[18] 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,[18] 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.[18] If applicable, the updated manuscript will be copublished 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 rigor. 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 eTables 2, 3, and 4 in Supplement 1, Supplement 2, and Supplement 3 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.[31,32] 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,[33] and the reporting of validation results using the CHART tool. 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 health care 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 rigor, 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 this remains 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 methodologic rigor of studies in this field.[34] 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.

Conclusions

The transparent reporting of CHA studies is crucial for their interpretation as we move toward the clinical integration of AI technologies. The CHART reporting guideline consists of a 12-item checklist and corresponding methodological diagram to support key stakeholders including clinicians, researchers, editors, peer reviewers, and readers in reporting, understanding, and interpreting the findings of CHA studies.

ARTICLE INFORMATION

Accepted for Publication: July 8, 2025.

Published: August 1, 2025. doi:10.1001/jamanetworkopen.2025.30220

Corresponding Author: Bright Huo, MD, BScPharm, Division of General Surgery, Department of Surgery, McMaster University, 50 Charlton Ave E, Hamilton, ON L8N 1Y2, Canada (brighthuo@dal.ca).

The CHART Collaborative: Bright Huo, PhD; Gary S. Collins, PhD; David Chartash, PhD; Arun J. Thirunavukarasu, MD; Annette Flanagin, BScN, MA; Alfonso Iorio, MD, PhD; Giovanni Cacciamani, MD, MSc; Xi Chen, MD, PhD; Nan Liu, MD, PhD; Piyush Mathur, MD; An-Wen Chan, MD, DPhil; Christine Laine, MD, PhD; Daniela Pacella, PhD; Michael Berkwits, MD, MScE; Stavros A. Antoniou, MD, PhD; Jennifer C. Camaradou; Carolyn Canfield; Michael Mittelman, MBA; Timothy Feeney, MD, PhD; Elizabeth W. Loder, MD, MPH; Riaz Agha, MBBS, MSc; Ashirbani Saha, PhD; Julio Mayol, MD, PhD; Anthony Sunjaya, MD, PhD; Hugh Harvey, MD; Jeremy Y. Ng, MD, PhD; Tyler McKechnie, MD; Yung Lee, MD, MPH; Nipun Verma, MBBS; Gregor Stiglic, PhD; Melissa McCradden, PhD; Karim Ramji, MD, MBA; Vanessa Boudreau, MD, MSc; Monica Ortenzi, MD, PhD; Joerg J. Meerpohl, MD, MSc; Per Olav Vandvik, MD, PhD; Thomas Agoritsas, MD, PhD; Diana Samuel, PhD; Helen Frankish, PhD; Michael Anderson, MBBS, PhD; Xiaomei Yao, MD, MSc; Stacy Loeb, MD, PhD; Cynthia Lokker, PhD; Xiaoxuan Liu, MBChB, PhD; Eliseo Guallar, MD, DrPH; Gordon H. Guyatt, MD, MSc.

Affiliations of The CHART Collaborative: Division of General Surgery, Department of Surgery, McMaster University, Hamilton, Ontario, Canada (Huo, McKechnie, Lee, Boudreau); UK EQUATOR Centre, University of Oxford, Oxford, United Kingdom (Collins); Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, Botnar Research Centre, University of Oxford, Oxford, United Kingdom (Collins); Department of Biomedical Informatics and Data Science, Yale University School of Medicine, New Haven, Connecticut (Chartash); Nuffield Department of Clinical Neurosciences, Medical Sciences Division, University of Oxford, Oxford, United Kingdom (Thirunavukarasu); *JAMA* and JAMA Network, American Medical Association, Chicago, Illinois (Flanagin); Department of Health Research Methods, Evidence, and Impact; Department of Medicine; McMaster University, Hamilton, Ontario, Canada (Iorio, Agoritsas, Lokker, Guyatt); USC Institute of Urology and Catherine and Joseph Aresty Department of Urology, Keck School of Medicine, University of Southern California, Los Angeles (Cacciamani); Artificial Intelligence Center at USC Urology, USC Institute of Urology, University of Southern California, Los Angeles, (Cacciamani); Sports Medicine Center, West China Hospital, Sichuan University, Chengdu, China (Chen); Department of Orthopedics and Orthopedic Research Institute, West China Hospital, Sichuan University, Chengdu, China (Chen); Duke-NUS Medical School, National University of Singapore, Singapore, Singapore. (N. Liu); Cleveland Clinic, Case Western Reserve University, Cleveland, Ohio (Mathur); Department of Medicine, Women's College Research Institute, University of Toronto, Toronto, Ontario, Canada (Chan); Annals of Internal Medicine, American College of Physicians, Philadelphia, Pennsylvania (Laine); American College of Physicians, Philadelphia, Pennsylvania (Laine); Department of Public Health, University of Naples Federico II, Naples, Italy (Pacella); Director, Office of Science Dissemination, Office of Science, Centers for Disease Control and Prevention, Atlanta, Georgia (Berkwits); Department of General Surgery, Papageorgiou General Hospital, Thessaloniki, Greece (Antoniou); British Psychological Society, University of Plymouth, Plymouth, United Kingdom (Camaradou); Innovation Support Unit, Department of Family Practice, University of British Columbia, Vancouver, British Columbia, Canada (Canfield); Patient SME, Independent Cybersecurity Professional, Philadelphia, Pennsylvania (Mittelman); The BMJ, London, United Kingdom (Feeney); Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina (Feeney); The BMJ, London, United Kingdom (Loder); Department of Neurology, Brigham and Women's Hospital, Boston, Massachusetts (Loder); International Journal of Surgery, London, United Kingdom (Agha); Eworkflow Ltd, London, United Kingdom (Agha); Department of Oncology, McMaster University, Hamilton, Ontario, Canada (Saha, Yao); Hospital Clinico San Carlos, Instituto de Investigación Sanitaria San Carlos, Facultad de Medicina Universidad Complutense de Madrid, Spain (Mayol); The George Institute for Global Health; Tyree Institute of Health Engineering, UNSW Engineering; School of Population Health, University of New South Wales Medicine and

Health, Sydney, Australia (Sunjaya); Hardian Health, Haywards Heath, United Kingdom (Harvey); Centre for Journalology, Ottawa Hospital Research Institute, Ottawa, Canada (Ng); Digestive Diseases Institute, Cleveland Clinic, Cleveland, Ohio (Lee); Postgraduate Institute of Medical Education and Research, Chandigarh, India (Verma); University of Maribor, Maribor, Slovenia (Stiglic); Australian Institute for Machine Learning, Adelaide, South Australia, Australia (McCradden); Phelix AI, Hamilton, Ontario, Canada (Ramji); Università Politecnica delle Marche, Clinica di Chirurgia Generale e d'Urgenza, Ancona, Italy (Ortenzi); Institute for Evidence in Medicine, Medical Center and Faculty of Medicine, University of Freiburg, Germany (Meerpohl); Cochrane Germany, Cochrane Germany Foundation, Freiburg, Germany (Meerpohl, Vandvik); MAGIC Evidence Ecosystem Foundation, Oslo, Norway (Vandvik, Agoritsas, Guyatt); University Hospitals of Geneva, Geneva, Switzerland (Agoritsas); The Lancet Digital Health, London, United Kingdom (Samuel); The Lancet, London, United Kingdom (Frankish); NIHR Clinical Lecturer, Health Organisation, Policy, Economics (HOPE), Centre for Primary Care and Health Services Research, The University of Manchester (Anderson); Senior Visiting Fellow, LSE Health, London School of Economics and Political Science, London, United Kingdom (Anderson); New York University Langone Health (Loeb); College of Medicine and Health, University of Birmingham, United Kingdom (X. Liu); School of Public Health, New York University (Guallar).

Author Contributions: Dr Huo had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Concept and design: Huo, Collins, Thirunavukarasu, Flanagin, Iorio, Cacciamani, Mathur, Chan, Laine, Pacella, Berkwits, Camaradou, Canfield, Feeney, Mayol, Sunjaya, Harvey, Ng, McKechnie, Verma, McCradden, Ramji, Boudreau, Meerpohl, Vandvik, Agoritsas, Frankish, Lokker, X. Liu, Guyatt.

Acquisition, analysis, or interpretation of data: Huo, Collins, Chartash, Thirunavukarasu, Flanagin, Iorio, Cacciamani, Chen, N. Liu, Mathur, Chan, Laine, Pacella, Berkwits, Antoniou, Camaradou, Canfield, Mittelman, Feeney, Loder, Agha, Saha, Mayol, Sunjaya, Harvey, Ng, McKechnie, Lee, Verma, Štiglic, McCradden, Ramji, Boudreau, Ortenzi, Meerpohl, Vandvik, Agoritsas, Samuel, Frankish, Anderson, Yao, Loeb, Lokker, X. Liu, Guyatt.

Drafting of the manuscript: Huo, Collins, Chartash, Thirunavukarasu, Cacciamani, Mathur, Pacella, Berkwits, Antoniou, Canfield, Mayol, Harvey, McCradden, Ortenzi, Frankish, Guyatt.

Critical review of the manuscript for important intellectual content: Huo, Collins, Chartash, Thirunavukarasu, Flanagin, Iorio, Cacciamani, N. Liu, Mathur, Chan, Laine, Pacella, Berkwits, Antoniou, Camaradou, Canfield, Feeney, Loder, Saha, Mayol, Sunjaya, Harvey, Ng, McKechnie, Lee, Verma, Štiglic, McCradden, Ramji, Boudreau, Ortenzi, Meerpohl, Vandvik, Agoritsas, Samuel, Anderson, Yao, Loeb, Lokker, X. Liu, Guyatt.

Statistical analysis: Huo, Pacella, Sunjaya, McCradden, Yao.

Obtained funding: Huo.

Administrative, technical, or material support: Huo, N. Liu, Canfield, Feeney, Saha, Sunjaya, McKechnie, Lee, Štiglic, Ramji, Ortenzi, Frankish.

Supervision: Cacciamani, Mathur, Antoniou, Harvey, Lee, Ramji, Boudreau, Vandvik, Loeb, Guyatt.

Conflict of Interest Disclosures: Dr Collins reported being a National Institute for Health and Care Research senior investigator. Dr Thirunavukarasu reported receiving funding from HealthSense. Mr Mathur reported being cofounder of BrainX LLC. Dr Saha reported receiving research funding from the Australian government and being cofounder of BantingMed Pty Ltd. Dr Samuel reported being the acting deputy editor for *The Lancet Digital Health*. Mr Mittelman reported receiving research funding from The Hospital Research Founding Group. Dr Feeney reported being on the executive committee of MDEpiNet. Dr Frankish reported being a senior executive editor for *The Lancet*. Dr Lokker reported being the editor in chief of *Annals of Internal Medicine*. Ms Flanagin reported being executive managing editor and vice president of editorial operations of *JAMA* and The JAMA Network. Drs Feeney and Loder reported being journal editors for *BMJ*. Dr Agha reported being the editor in chief of *International Journal of Surgery.* Dr Stiglic reported being executive editor of *Artificial Intelligence in Medicine*. Dr Loeb reported receiving consulting fees from Astellas. Dr Pacella reported receiving research funding from the Italian Ministry of University and Research. Prof Ortenzi reported being a paid consultant for Theator. Drs Agoritsas, Vandvik, and Guyatt reported being board members of the MAGIC Evidence Ecosystem Foundation. No other conflicts were reported.

Funding/Support: The Chatbot Assessment Reporting Tool (CHART) was funded by the *First Cut* competition and the Postgraduate Medical Education Committee at McMaster University.

Role of the Funder/Sponsor: The funder had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Additional Information: This article is being published jointly in *Artificial Intelligence in Medicine, Annals of Family Medicine, BJS, BMC Medicine, BMJ Medicine, JAMA Network Open, The Lancet, NEJM-AI, and Surgical Endoscopy.* 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.

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(1):71. doi:10.1038/s43856-024-00492-0

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(5):e367-e373. doi:10.1016/S2589-7500(24)00047-5

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(12):2988. doi:10.1038/s41591-023-02656-2

4. Huo B, McKechnie T, Ortenzi M, 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(3):463-469. doi:10.1007/s12553-024-00836-9

5. Huo B, Marfo N, Sylla P, 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(10):5668-5677. doi:10.1007/s00464-024-11155-5

6. Huo B, Boyle A, Marfo N, et al. Large language models for chatbot health advice studies: a systematic review. *JAMA Netw Open*. 2025;8(2):e2457879. doi:10.1001/jamanetworkopen.2024.57879

7. CHART Collaborative. Protocol for the development of the Chatbot Assessment Reporting Tool (CHART) for clinical advice. *BMJ Open*. 2024;14(5):e081155. doi:10.1136/bmjopen-2023-081155

8. Moher D, Schulz KF, Simera I, Altman DG. Guidance for developers of health research reporting guidelines. *PLoS Med*. 2010;7(2):e1000217. doi:10.1371/journal.pmed.1000217

9. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*. 2024;385:e078378. doi:10.1136/bmj-2023-078378

10. Ong JCL, Chang SYH, William W, et al. Ethical and regulatory challenges of large language models in medicine. *Lancet Digit Health*. 2024;6(6):e428-e432. doi:10.1016/S2589-7500(24)00061-X

11. Altman DG, Simera I, Hoey J, Moher D, Schulz K. EQUATOR: reporting guidelines for health research. *Lancet*. 2008;371(9619):1149-1150. doi:10.1016/S0140-6736(08)60505-X

12. Protocol for a scoping review of chatbot assessment studies: guidance for the CHART tool. OSF Registries. Accessed July 16, 2025. https://osf.io/cxsk3

13. Munn Z, Peters MDJ, Stern C, Tufanaru C, McArthur A, Aromataris E. Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. *BMC Med Res Methodol*. 2018;18(1):143. doi:10.1186/s12874-018-0611-x

14. Decision Eyes. Welphi. Accessed July 10, 2025. https://www.welphi.com/

15. Liu X, Rivera SC, Moher D, Calvert MJ, Denniston AK; SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI Extension. *BMJ*. 2020;370:m3164. doi:10.1136/bmj.m3164

16. The CHART Collaborative. Reporting guidelines for chatbot health advice studies: explanation and elaboration for the Chatbot Assessment Reporting Tool (CHART). *BMJ*. 2025;390:e083305. doi:10.1136/bmj-2024-083305

17. Yin S, Fu C, Zhao S, et al. A survey on multimodal large language models. *arXiv*. Preprint published online June 23, 2023. http://arxiv.org/abs/2306.13549

18. Akl EA, Meerpohl JJ, Elliott J, Kahale LA, Schünemann HJ; Living Systematic Review Network. Living systematic reviews: 4. Living guideline recommendations. *J Clin Epidemiol*. 2017;91:47-53. doi:10.1016/j.jclinepi.2017.08.009

19. Begg C, Cho M, Eastwood S, et al. Improving the quality of reporting of randomized controlled trials. The CONSORT statement. *JAMA*. 1996;276(8):637-639. doi:10.1001/jama.276.8.637

20. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP; STROBE Initiative. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *BMJ*. 2007;335(7624):806-808. doi:10.1136/bmj.39335.541782.AD

21. Rivera SC, Liu X, Chan AW, Denniston AK, Calvert MJ; SPIRIT-AI and CONSORT-AI Working Group. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI Extension. *BMJ*. 2020;370:m3210. doi:10.1136/bmj.m3210

22. Vasey B, Nagendran M, Campbell B, et al; DECIDE-AI expert group. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *Nat Med*. 2022;28(5):924-933. doi:10.1038/s41591-022-01772-9

23. Cacciamani GE, Collins GS, Gill IS. ChatGPT: standard reporting guidelines for responsible use. *Nature*. 2023; 618(7964):238. doi:10.1038/d41586-023-01853-w

24. Xie SM, Pham H, Dong X, et al. DoReMi: optimizing data mixtures speeds up language model pretraining. *arXiv*. Preprint published online May 17, 2023. https://arxiv.org/abs/2305.10429

25. 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;4(10):101230. doi:10.1016/j.xcrm.2023.101230

26. 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(6):e333-e335. doi:10.1016/S2589-7500(23)00083-3

27. The Lancet Digital Health. Large language models: a new chapter in digital health. *Lancet Digit Health*. 2024; 6(1):e1. doi:10.1016/S2589-7500(23)00254-6

28. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. *Nat Med*. 2023;29(8):1930-1940. doi:10.1038/s41591-023-02448-8

29. 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(1):183. doi:10.1038/s41746-024-01157-x

30. Thirunavukarasu AJ. Large language models will not replace healthcare professionals: curbing popular fears and hype. *J R Soc Med*. 2023;116(5):181-182. doi:10.1177/01410768231173123

31. Kane RL, Wang J, Garrard J. Reporting in randomized clinical trials improved after adoption of the CONSORT statement. *J Clin Epidemiol*. 2007;60(3):241-249. doi:10.1016/j.jclinepi.2006.06.016

32. 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(1):60. doi:10.1186/2046-4053-1-60

33. de Hond A, Leeuwenberg T, Bartels R, et al. From text to treatment: the crucial role of validation for generative large language models in health care. *Lancet Digit Health*. 2024;6(7):e441-e443. doi:10.1016/S2589-7500(24)00111-0

34. Logullo P, MacCarthy A, Kirtley S, Collins GS. Reporting guideline checklists are not quality evaluation forms: they are guidance for writing. *Health Sci Rep*. 2020;3(2):e165. doi:10.1002/hsr2.165

SUPPLEMENT 1.

eTable 1. Candidate CHART Checklist Items
eTable 2. The CHART Checklist
eTable 3. The CHART Abstract Checklist
eTable 4. The CHART Panel

SUPPLEMENT 2.

eTable 1. Candidate CHART Checklist Items (editable spreadsheet)
eTable 2. The CHART Checklist (editable spreadsheet)
eTable 3. The CHART Abstract Checklist (editable spreadsheet)
eTable 4. The CHART Panel

SUPPLEMENT 3.

Methodological Diagram