Reporting guidelines for chatbot health advice studies: explanation and elaboration for the Chatbot Assessment Reporting Tool (CHART)

On behalf of: The CHART Collaborative

Correspondence to: B Huo (https://orcid.org/0000-0003-4999-4328), McMaster University, Hamilton, ON L8N 1Y3, Canada brighthuo@dal.ca (or @brighthuo on X; ORCID 0000-0003-4999-4328)

Additional material is published online only. To view please visit the journal online.

Cite this as: *BMJ* 2025;390:e083305
http://dx.doi.org/10.1136/bmj-2024-083305

Accepted: 16 June 2025

The Chatbot Assessment Reporting Tool (CHART) reporting guideline promotes transparent and comprehensive reporting of studies evaluating the performance of generative artificial intelligence (AI)-driven chatbots for the purposes of summarising clinical evidence and providing health advice, referred to here as chatbot health advice (CHA) studies. CHART is the product of an international, multi-phase, consensus based initiative involving various stakeholders and comprises a 12-item checklist with 39 subitems. The checklist includes items on open science, title and abstract, introduction, model identification, model details, prompt engineering, query strategy, performance definition and evaluation, statistical analysis, results, discussion, with an accompanying flow diagram. Each item includes distinct subitems. This explanation and elaboration article discusses each subitem and provides a detailed rationale for its inclusion in the CHART checklist.

Generative artificial intelligence (AI) refers to a class of AI algorithms that can process and generate text, images, or other data forms, exhibiting human-like comprehension and output.*1* Unlike traditional AI models designed to perform specific tasks, generative AI can produce novel outputs that are not represented within its training data, enabling it to execute a diverse range of tasks in healthcare.[2 3] There is a considerable interest in applying generative AI to support clinical decision making in healthcare.[4]

Clinical decision making is a complex, structured process where physicians and other clinicians integrate patient data, medical knowledge, clinical judgment, and patient values to formulate diagnoses and treatment plans, often following established frameworks for systematic evaluation and intervention.[5] Its complexity arises from the need to interpret extensive data, consider multiple diagnoses, manage uncertainties, account for variable patient presentations, and integrate evolving medical evidence.[6] Despite their expertise, clinicians could still make errors in decision making due to cognitive biases, uncertainty, incomplete information, and the inherent complexity of clinical cases.[7 8] Moreover, humans and systems are prone to fault, because medical errors are the third leading cause of death in the US and represent a major cause of preventable patient harm.[9]

Given these challenges, interest is growing in the use of generative AI models to assist clinicians in clinical decision making and optimise evidence based care. These AI tools, functioning as chatbots, offer advice to clinicians by providing health information, interpreting test results, and suggesting diagnostic and treatment recommendations.[10-15] Generative AI-driven chatbots are being quickly adopted as sources of health advice, and thus investigators have begun to evaluate the quality and accuracy of the information which they provide to both clinicians and patients.[16-19] We refer to these studies as generative, AI-driven, chatbot health advice (CHA) studies. However, the completeness of reporting among these studies can vary widely,[4] which could impair the interpretation of study findings and prevent readers from assessing their reliability.[20]

Existing reporting guidelines offer recommendations for evaluating AI in healthcare but are designed for studies aimed at specific tasks such as disease outcome prediction,[21] diagnostic test accuracy,[22] early stage clinical evaluation of decision support systems,[23] imaging studies,[24] health economic evaluations,[25] or general interventions.[26 27] These guidelines focus on AI models based on structured or unstructured data after training with similar data labelled by humans and generally apply deterministic algorithms aimed at clear, predefined outputs.[28 29] In contrast, generative AI models produce novel outputs that are not represented in their training data, and are not limited to predefined answer categories or tasks.[30] This capability allows generative AI to be deployed as chatbots in healthcare, offering dynamic and conversational health advice.[16 17] Given these substantial differences in underlying

SUMMARY POINTS

CHART was developed according to robust methodological standards following a systematic review, Delphi consensus, and panel consensus meetings among international, multidisciplinary stakeholders

The CHART checklist contains 12 key reporting items with 39 subitems, while the abstract checklist contains nine key reporting items with 16 subitems

The CHART reporting guideline also includes a fillable methodological diagram

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 (CHA) study	Any research study evaluating the performance of chatbots when summarising 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 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 or automatically to optimise the intended outcome of the algorithm. Parameters may be at the model level to optimise tuning (hyperparameters) or so-called weights within the model linking layer to layer (parameters).
Post-implementation/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	Text input by a user into the chatbot for the purpose of communicating with the LLM.
Prompt engineering	The input provided by users when interfacing with a generative AI-driven chatbot, leading to input interaction with the AI model.
Query	The act of communicating with a generative AI-driven chatbot by inputting a prompt into the chatbot that 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 optimisation.

principles and applications, current AI research guidelines are insufficient for CHA studies, which evaluate the clinical accuracy of health advice from generative AI-driven chatbots.

To address this gap, we developed the Chatbot Assessment Reporting Tool (CHART) reporting guideline for CHA studies.[31 32] This guideline resolves several critical reporting gaps.[4] Generative AI outputs vary considerably because of context,[33] requiring detailed documentation of the prompts used to elicit model output. Additionally, given the diversity of models and their continuous learning and adaptation,[2] changes in model development and deployment can substantially influence performance,[34] necessitating the disclosure of technical details. Lastly, our reporting guideline standardises the reporting of performance evaluation to optimise the comprehensiveness and transparency in the reporting of CHA studies. This explanation and elaboration article describes the rationale behind each subitem included in the CHART checklist. All terminology applied in the CHART reporting guideline can be found in the glossary (table 1).

Methods

Our methodology is described in detail in the accompanying statement article.[35] The CHART study protocol further outlines the development of this reporting guideline: https://osf.io/cxsk3.[32] Ethical approval was submitted to the Hamilton Integrated Research Ethics Board, and the need for review and approval was waived (HiREB 17025).

CHART development

In brief, we completed a comprehensive scoping review and identified 137 CHA studies published before 27 October 2023.[4] We evaluated the methods and reporting standards across these studies to draft candidate checklist items. An international, multidisciplinary advisory committee comprised of experts in generative AI, regulation, medical ethics, policy, and various other disciplines[32] voted on the candidate checklist items via an online Delphi platform Welphi, *Decision Eyes* (https://www.welphi.com/) from 6 May to 9 June 2024. Following the modified Delphi consensus, an international, multidisciplinary expert panel comprised of 48 members reviewed the anonymised voting results on the candidate checklist as well as additionally suggested items from the advisory committee (appendix 1). The panel reviewed these items over three panel consensus meetings and agreed on a list of 12 checklist items, as well as a flow diagram on 30 June, 5 August, and 2 September 2024. The expert panel reviewed, discussed, and approved all items and subitems on the final checklist with at least 80% agreement. All items excluded from the checklist were unanimously removed by the expert panel after extensive discussion and input from relevant content experts. Authors of prior CHA studies then performed pilot testing by using the checklist to evaluate new, separate CHA studies. After extensive feedback, we created a final list of 12 checklist items comprising 39 distinct subitems as well as a methodological diagram for the CHART reporting guideline (table 2).

Table 2 | CHART checklist

Heading	Item 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 masked 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 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 assessment 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; CHART=Chatbot Assessment Reporting Tool.
*If in supplementary appendix, indicate "supp" and appendix number, if applicable.

Full methodological details are available in corresponding publications.[32 35] An editable version is available for the full checklist (appendix 2) as well as the abridged abstract checklist (appendix 3). Table 3 lists examples of possible excerpts from the manuscript text corresponding to each subitem of the CHART checklist. Note that these are not examples of high-quality methodology but rather serve as examples for illustrative purposes to demonstrate use of the CHART checklist. Table 4 contains the CHART abstract checklist. Figure 1 outlines the methodological diagram. Appendix 4 contains a fillable version of the methodological diagram.

CHART checklist items

Item 1: Title and abstract

Subitem 1a—Title: state that the study is assessing one or more generative AI-driven chatbots for clinical evidence or health advice

Explanation: To ensure clarity for all readers, authors are encouraged to indicate that the study is evaluating the performance of a generative AI-driven chatbot.

Table 3 | CHART checklist examples

CHART checklist item	Item No	Possible manuscript excerpts*
Title and abstract		
Title	1a	Surgical management recommendations to patients via LLM-driven chatbot: a pragmatic study 3
Abstract/summary	1b	—
Introduction		
Background	2a	—
	2b	—
Methods		
Model identifiers	3a	We evaluated ChatGPT-4o (gpt-4o-2024-08-06), with a knowledge cut-off date of 30 September 2023.
	3b	Both its chatbot and LLM are proprietary, closed-source entities. No additional software license was needed.
Model details	4a	We used the base model of gpt-4o.
	4b	Additional information can be found online: https://platform.openai.com/docs/api-reference/introduction.
	4c	We used gpt-4o (gpt-4o-2024-08-06) set to the following parameters: temperature 0.8; context window 128 000; total tokens 16 384; presence penalty 0; frequency penalty: 0.
Prompt engineering	5a	The surgical resident drafted prompts in English that were revised by the patient partner to ensure plain language. We applied a semi-structured, iterative approach and revised our prompts on the basis of the model output, with the goal of using each successive revision and test prompt to optimise the response from the model. We applied follow-up prompts when needed. We focused on the adjustment of phrasing by using synonyms or using layperson terminology.
	5ai	We used recommendations from the 2024 SAGES guideline on the management of appendicitis to derive our initial prompts.
	5aii	Neither the surgical resident nor patient partner had prior experience with developing prompts and were supervised by an attending surgeon with two prior publications evaluating the performance of generative AI-driven chatbots when providing health advice.[4 5]
	5aiii	Addressed above.
	5b	Study prompts are included in supplementary appendix 1.
Query strategy	6a	We accessed the model via OpenAI's ChatGPT-4o web interface on chat.openai.com.
	6b	We queried gpt-4o-2024-08-06 on 25 April 2025 in Halifax, NS, Canada.
	6c	We inputted prompts in separate chat sessions.
	6d	Chatbot/model output is available in supplementary appendix 2.
Performance evaluation	7a	We defined the ground truth using response alignment with the 2024 SAGES guideline recommendations on the surgical management of appendicitis.[3]
	7b	We considered responses to be correct only if they produced an answer that matched the SAGES guideline recommendation. If responses also included incorrect information, this was regarded as incorrect. Responses not aligning with guideline recommendations, including responses that did not provide a clinical decision, were considered to be incorrect.
	7bi	The surgical resident and patient partner evaluated responses, and the attending surgeon provided clarification as needed to reconcile areas of disagreement. Only the attending surgeon had experience with performance evaluation in the context of developing prior CHA studies.
	7bii	Addressed by above. A patient partner was involved in evaluating responses. They did not have prior experience with performance evaluation for CHA studies.
	7biii	Authors were not blinded to the identity of the generative AI-driven chatbot.
Sample size	8	This exploratory study did not use a comparator group. However, based on a prior CHA study that obtained 121 responses and found a difference between ChatGPT and LLAMA in their ability to provide advice on the management of cholecystitis, we aimed to at minimum obtain 60 responses to distinct patient cases.
Data analysis	9a	We used descriptive statistics in the form of counts and percentages. We did not assess the reproducibility of responses.
	9ai	We reported the number of correct responses by ChatGPT-4o.
Results		
	10a	ChatGPT-4o was able to correctly advise on 58/60 patient cases of acute appendicitis.
	10b	For the two incorrect responses, ChatGPT-4o recommended against surgical intervention. In one of these cases, we sought the expert opinion of two additional general surgeons who thought that considerable harm in the form of inadvertent morbidity could have resulted from the advice given by ChatGPT as patients might have delayed seeking care rather than undergoing early surgical intervention when indicated.
	10c	We did not detect responses that may lead to physical, emotional, or psychological patient harm.
Discussion		
	11a	—
	11b	—
	11c	—
Open science		
Disclosures	12a	None of the authors had conflicts of interest to disclose.
Funding	12b	This study was unfunded.
Ethics	12c	We applied for ethical approval from research ethics body A. The need for ethical approval was waived.
	12ci	We did not use patient data.
	12cii	We did not use original copyrighted data.
Protocol	12d	Full details related to our methodology are listed in our protocol, which was registered a priori on osf.io/chart.
Data availability	12e	Neither data nor code data were needed.

AI=artificial intelligence; CHA=chatbot health advice; CHART=Chatbot Assessment Reporting Tool; LLM=large language model; SAGES=Society of American Gastrointestinal and Endoscopic Surgeons.
*Example and references are fabricated for illustrative purposes.

Table 4 | 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; CHART=Chatbot Assessment Reporting Tool.

Doing so facilitates indexing for subsequent record collection and article searching. Authors should avoid broadly stating that the article discusses artificial intelligence.

Subitem 1b—Abstract/summary: Apply a structured format, if applicable

Explanation: Authors may be writing an abstract or summary for the purposes of conference submission, grant submission, or for manuscript submission. The abridged checklist for CHA studies' abstracts looks at relevant sections that generally include background, methods, results, and conclusion, capturing core information from subitems in the full version of the checklist. The full rationale for each subitem is discussed in subsequent sections of this article following the title and abstract.

The background section should provide the context of the problem or question being addressed by the research (item 2a). The abstract should clearly state that the study is a CHA study, which will be evaluating the performance of a model when summarising clinical evidence or providing health advice. Authors should state the type of advice being evaluated, including health prevention, screening, differential diagnosis, diagnosis, treatment, prognosis, and/or general information (item 2b). In the methods section, authors should identify the generative AI model(s) and associated chatbot(s) under evaluation by stating the model name, version number, and the date(s) of the query or queries (item 3a).

The abstract should also include details on model accessibility, referring to the use of open-source/accessible models versus closed-source/inaccessible models (item 3b) and whether authors are evaluating a novel base model, tuned model, or fine-tuned model (item 4a). Authors should state prompt engineering if applicable (item 5a) and identify the sources of their prompts (item 5ai). Route of access to the model should also be described, which might range from application programming interfaces to direct interaction with chatbot platforms (item 6a).

Additionally, the definition of successful performance, or ground truth, should be stated explicitly (item 7a), and the process for performance evaluation should be outlined (item 7b). Details regarding the sample size (item 8) and data analysis (item 9a) should also be reported. Authors should state the alignment of model responses with the ground truth in the results section (item 10a) before summarising the implications for stakeholders and key takeaway points in their conclusion. We encourage conference organisers, journal editors, or grant/award committees requesting abstracts for CHA studies to use an appropriate outline and word count to facilitate the fulfilment of the minimum reporting standards outlined here (table 4).

Item 2: Introduction

Subitem 2a—Background: state the scientific background, rationale, and healthcare context for evaluating the generative AI-driven chatbot(s), referencing relevant literature when applicable

Explanation: Investigators should describe the context in which this research is occurring, including the current state of relevant literature. This foundational knowledge should include only the information necessary for readers to understand the potential role of the generative AI-driven chatbot in a clinical setting. For example, whether the work being undertaken is preclinical (the role of the gut microbiome on diverticulitis) or clinical (the treatment

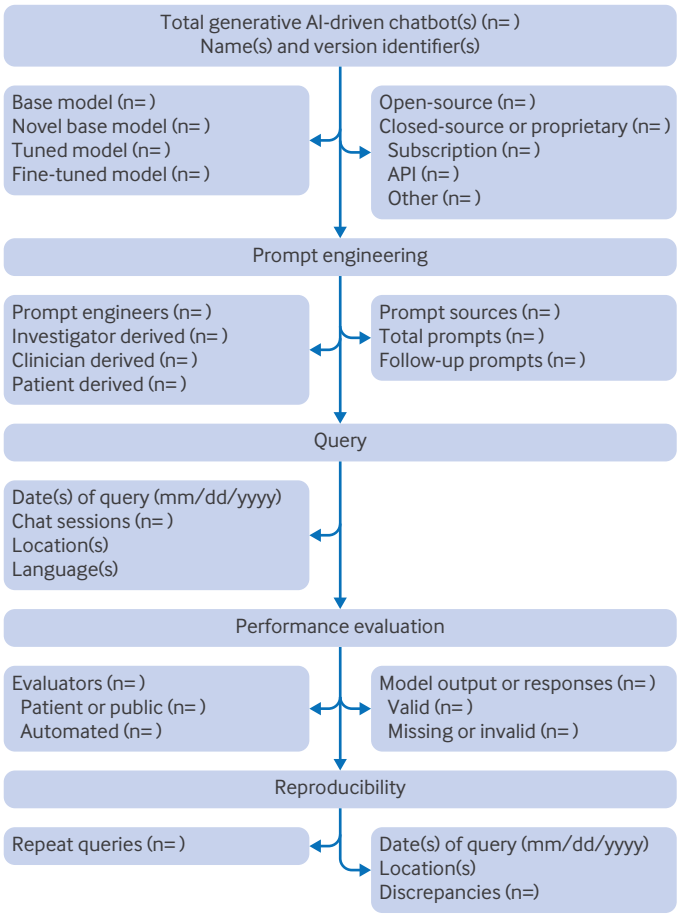

Fig 1 | CHART methodological diagram. AI=artificial intelligence; API=application programming interfaces; CHART=Chatbot Assessment Reporting Tool

of diverticulitis), investigators are encouraged to outline relevant advances made in the field thus far. Authors should further clarify whether applications are designed to provide advice according to specific resources (such as societal guidelines) for any range of conditions, specialties, or scenarios. Overall, investigators are encouraged to signpost the use-case for which the generative AI-driven chatbot(s) is being evaluated. Authors should describe the gap in current knowledge that their work addresses.

Subitem 2b—Background: state the aims and research questions including the target audience, intervention, comparator(s), and outcome(s)

Explanation: We encourage authors to specify the purpose of their study by explicitly outlining the aims of their work. Investigators are encouraged to clarify whether the generative AI-driven chatbot(s) is/are directed at clinicians, patients, members of the general public, and/or CHA researchers as the end user. Authors should further clarify whether they evaluated the performance of a single generative AI-driven chatbot as an intervention, additional chatbot/model comparators, other AI models, and/or humans (such as clinicians). Furthermore, investigators should state whether the aim of the study was to evaluate the ability of the generative AI-driven chatbot(s) to summarise clinical evidence, provide health advice (in the form of guidance or recommendations), or both summarise clinical evidence and provide health advice. Finally, authors should specifically outline the research questions being investigated. These questions might include but are not limited to whether the work relates to health prevention, screening, diagnosis, treatment, prognosis, and/or general health information. Investigators should also clearly state the primary and secondary outcome(s) of the study as applicable, described in further detail in subitem 9ai.

For instance, investigators could have evaluated the ability of chatbot A versus chatbot B to summarise the clinical evidence supporting the use of drug X in lowering blood pressure among patients with hypertension, with physicians as the end user. Alternatively, authors could have evaluated the ability of chatbot A versus B to provide clinical recommendations to help physicians reduce HbA_1C (glycated haemoglobin) among patients with type 2 diabetes. Equally, investigators might wish to know whether chatbot A can advise a patient on lifestyle changes they can undertake to mitigate the symptoms of heartburn from gastroesophageal reflux disease, or whether generative AI-driven chatbots can advise surgeons on the surgical management of their patient with gastroesophageal reflux disease to mitigate dysphagia.[13]

Methods

Item 3: Model identifier

Subitem 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

Explanation: To facilitate the identification of the intervention being evaluated,[36] authors are encouraged to identify the generative AI model and chatbot under assessment. Notably, both generative AI models and chatbots are typically separate entities.[37] For example, when authors use chatbots to engage with large language models (LLMs) in CHA studies, they are effectively evaluating the performance of the ensemble of both the chatbot and the generative AI model, as the software architecture of the chatbot may impact their overall performance and functionality.[37] CHA researchers should be aware that accessing generative AI models through application programming interface keys still constitutes the use of the chatbot component to access the model itself. Both chatbots and generative AI models can be version controlled, and readers must understand which iteration of the model and chatbot is being evaluated to properly interpret the study findings. For instance, authors querying a more recent version of ChatGPT on 5 September 2024 would be using ChatGPT-4o as a chatbot, with the "4o" specifying the version. This chatbot allows authors to access GPT-4o as an LLM for the ChatGPT interface ensemble. Notably, the model GPT-4o has been updated several times since its original release. In addition to identifying the name and version number of the model and chatbot

being assessed, investigators should specify the date of release or last update of both the model and the chatbot where possible. One example illustrated for ChatGPT would be "gpt-4o-2024-08-06," which helps to ensure clarity for readers, as iterative updates might be implemented without announcements or recognisable changes to version identifiers.

Subitem 3b—State whether the generative AI model(s) and chatbot(s) are open-source or closed-source/proprietary

Explanation: Both generative AI models and chatbots may be open-source or closed-source.[38] Closed-source approaches conceal source code, training data, model architecture, and/or fine-tuning protocols, whereas open-source offers more transparency and control to the user by facilitating model and/or chatbot adaptation and customisation.[38 39] Proprietors of closed-source models and chatbots might also alter their product or its training data and subsequently change their output at any time without the ability of users to recognise these changes, which affects the reproducibility of scientific research evaluating generative AI-driven chatbots.[39] For these reasons, authors should specify whether they are evaluating open-source or closed-source models and chatbots. For instance, ChatGPT-4o is a proprietary/closed-source chatbot, while GPT-4o is a proprietary generative AI model.

Additionally, specific software licenses might be needed for users to access the model(s)/chatbot(s) under evaluation, which could affect the ability of users to modify the model(s)/chatbot(s). If applicable, investigators are encouraged to report the details of each license granted.

Item 4: Model details

Subitem 4a—State whether the generative AI model was a base model or a novel base model, tuned model, or fine-tuned model

Explanation: Investigators might choose to evaluate the performance of one or more generative AI models:

- An out-of-the-box model that has already been developed and described (a base model).
- A novel base model that has not been previously described and is proposed as part of the study itself.
- A pre-existing model that has been tuned (a tuned model)—which has been customised for some functionality.[14]
- A fine-tuned or adapted model (through the alteration of whole or parts of model parameters, weights, or the addition or removal of parameters and layers) that has been revised either through algorithmic tuning to alter its performance (such as through reinforcement learning or retrieval-augmented generation), or through additional training with new data.[40]

It is important to explicitly specify which of these model types are being evaluated in order for readers to contextualise the study findings.

Subitem 4b—If a base model is used, cite its development in sufficient detail to identify the model

Explanation: Investigators might evaluate generative AI models whose development have been described in great detail, such as a base model.[3] In these cases, it is sufficient for authors to provide readers with a citation or link to an existing resource where this information is readily available. We encourage investigators to reference peer reviewed sources where possible. Authors might update pre-existing generative AI models such as LLMs, and thus various iterations of a given model could exist. Investigators should be specific to enable readers to understand the exact base model under evaluation, with the aim of reporting methodology that is both transparent and reproducible. Model descriptions are further addressed in subitem 4c. Where investigators evaluate closed-source models and chatbots, we recognise that the information available might be limited. In these situations, authors are encouraged to report as much detail as is accessible and acknowledge this lack of identifying information as a limitation.

Subitem 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

Explanation: Investigators might further revise generative AI models such as LLMs before their release for user engagement (pre-implementation or pre-deployment) or following their release (post-implementation or post-deployment) through tuning such as customisation of model parameters such as temperature or token length. Authors may also refine their models via fine-tuning through reinforcement learning, retrieval-augmented generation, further training with new datasets, or the adjustment of model parameters such as penalties, add-on availability, and/or layers (table 1).[41 42] This list is not meant to be exhaustive, and authors are encouraged to report what is minimally necessary to allow others to reproduce their study methodology, with a focus on enabling the replicability of experiments. Although organisations may be incentivised to restrict access to their models or conceal the model architecture and/or training/source data,[41] it is imperative that we move towards open-source generative AI models to facilitate the transparent evaluation and reporting of CHA studies that is necessary to establish the role of generative AI-driven chatbots in the clinical workflow.

Item 5: Prompt engineering

Subitem 5a—Describe the evolution of study prompt development

Explanation: Prompt engineering refers to the development and optimisation of prompt words and sentences to optimise the stability and appropriateness of model output.[10] Prompt engineering affects the performance of generative AI models such as LLMs, particularly in the context of clinical questions.[10]

Thus, investigators should report the overall process undertaken during study prompt development to facilitate the interpretation of study findings. We encourage authors to report the number and nature of test prompts used to elicit model responses. Additionally, investigators might encounter barriers during prompt development. For example, chatbots could refuse to answer a medical question or might simply present multiple options without committing to a clinical decision.[15] Thus, further test iterations of prompts might be needed to elicit guidance from the chatbot.[15] These barriers can be circumvented with the use of follow-up prompts that might be applied in a standardised way across all models under evaluation.[15] Authors might also comment on whether prompts were reviewed for grammatical accuracy, and whether approaches were taken to mitigate biased responses from the generative AI-driven chatbot. Subitems 5ai-iii further look at specific elements of prompt development that should be clearly outlined. If applicable, authors may report any approaches taken to mitigate potentially harmful output from the model during prompt development. Biased, potentially harmful, or misleading responses are described in more detail in subitem 10c.

Subitem 5ai—Describe the sources of prompts

Explanation: Prompts might originate from a wide variety of sources. Although investigators frequently derive prompts on their own (based on expert opinion), other sources such as professional society/organisation websites, non-evidence based websites including patient forums, social media, textbooks, and clinical practice guidelines have been used previously.[4] Alternatively, investigators can collect prompts from patients or clinicians through prospective or retrospective review of patient records or communication systems used for healthcare work. Providing readers with a clear summary of the source of experimental prompts helps readers contextualise the clinical relevance of the chatbot system under evaluation and might also reveal where the system is likely to perform well or fail.

Subitem 5aii—State the number and characteristics of the individual(s) involved in prompt engineering

Explanation: In addition to describing the source of prompts, the number of individuals overseeing the development of prompts should be clearly stated. Furthermore, we encourage investigators to describe the demographic characteristics and/or background of the individuals involved in study prompt development. Relevant details might vary depending on the aim of the study and the type of advice being examined. Individuals involved in prompt engineering could include study investigators such as clinicians, data scientists, or other researchers. In particular, the CHART expert panel voiced the importance of identifying the expertise of those involved in study prompt development to support readers in judging the trustworthiness of study findings. This knowledge might include prior experience in publishing CHA studies, relevant expertise in the clinical performance evaluation of generative AI models, specific clinical expertise related to the topic being examined, general experience in AI research. Authors without expertise in these areas should not be precluded from initiating CHA studies but might be less likely to develop optimal systems for generating accurate, pragmatic, and safe responses to clinical queries. In all scenarios, the prior experience and expertise of prompt engineers should be expressly stated.

Subitem 5aiii—Provide details of any patient and public involvement during prompt engineering

Explanation: Prompt engineering could have a role in optimising model output in response to user queries, but the responses from generative AI-driven chatbots have been shown to be very sensitive to prompt phrasing.[10] Owing to a discrepancy between the medical language used by clinicians and patients, differences in phrasing between study investigators and patients might affect the generalisability of study findings of CHA studies.[43] In the spirit of patient centred care, authors may use a pragmatic approach by including patients or other members of the public as study investigators to enhance the external validity of their study. We encourage authors to report whether patients or other members of the public were involved in study prompt development to facilitate the interpretation of their findings. Further, we encourage investigators to report how these individuals were involved in prompt engineering. For comprehensive guidance on how to report patient and public involvement in research, authors are advised to consult the GRIPP2 (guidance for reporting involvement of patients and the public) statement.[44]

Subitem 5b—Provide study prompts

Explanation: By providing the prompts used in experiments, investigators will improve the reliability and trustworthiness of their findings. The phrasing of prompts and the manner in which queries are conducted directly impact model output,[10] so study investigators of CHA studies are strongly encouraged to provide the raw transcript of prompts used for model query in either the manuscript body or appendix. This information enables readers to examine prompts for factors that influence model output. For instance, these factorsmight include the use of standardised prompts, one or more languages, follow-up prompts, and checks for the reproducibility of responses. Additionally, authors should consider the importance of inclusive design and testing, because English versus non-English prompts could prompt language or cultural barriers in generative AI-derived health advice depending on the nature and scope of the study.[45 46]

Item 6: Query strategy

Subitem 6a—State route of access to generative AI model

Explanation: Users can query generative AI-driven chatbots through a variety of interfaces. For instance,

GPT-4o is a closed-source, proprietary LLM that be accessed through ChatGPT, an online chatbot, or via an application programming interface. For application programming interfaces, investigators might not interact solely via a user-facing chatbot using natural language prompts but rather by coding a script to implement many queries and response processing with minimal delay or human involvement during trials. In these cases, all used code should be reported as per subitem 12e. Bing, Enterprise Co-Pilot, and other LLMs also leverage the same model but through their own interfaces. For the purpose of transparency and reproducibility, investigators should report the manner in which they access the model under evaluation, to provide further context for readers.

Subitem 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

Explanation: As generative AI models are frequently updated, their performance can vary depending on the date of query. Authors of CHA studies should report the day, month, and year in which the chatbot was queried.[47] Additionally, the availability of generative AI models might vary depending on location, while model performance could differ on the basis of local recommendations and best practices, described further in item 7a. For these reasons, investigators of CHA studies should also report the city and country of the query.

Subitem 6c—Describe whether prompts were inputted into separate chat session(s)

Explanation: Users can engage with generative AI-driven chatbots by entering a prompt in a new chat session. Users may continue to enter more prompts in response to model outputs, creating continuous dialogue. However, as users enter successive prompts, subsequent LLM responses may become influenced by prior discourse, and thus investigators may instead enter their prompts in distinct chat sessions.[15] We encourage authors to optimise the reproducibility of their methodology by stating whether separate chat sessions or continuous dialogue in a single chat session were used.

Subitem 6d—Provide all generative AI-driven chatbot output/responses

Explanation: As with prompts, we encourage study investigators to transparently report the model responses from their experiments. By having access to the raw discourse between investigator prompts and model output, readers will be better positioned to both understand and reproduce the study methodology. This added information will improve the generalisability of study findings, because readers can judge the practicality of the study queries used to elicit chatbot responses. Authors may elect to present these data in the main body of the manuscript, or in the appendix or supplementary file as appropriate or as specified by the applicable journal. If investigators are evaluating closed-source or proprietary models, their ability to share transcripts might be limited. In these cases—as well as other situations where the raw transcripts of model responses are not accessible—authors should report this as a major limitation of their study.

Item 7: Performance definition

Subitem 7a—Define the ground truth or reference standard used to define successful generative AI-driven chatbot performance

Explanation: In addition to stating the primary outcome of the study, authors should explicitly define what is considered to be successful performance by the chatbot, a process otherwise known as defining the reference standard or ground truth.[40]

To evaluate the clinical performance of generative AI-driven chatbots, study authors should state their primary (and secondary) outcome(s). Additionally, authors should explicitly define what is considered to be successful performance by the chatbot(s) under evaluation, a process otherwise known as defining the reference standard or ground truth.[40] This standard can be used to gauge whether chatbot responses are accurate. Investigators are encouraged to apply a pragmatic approach to performance evaluation as CHA studies are often evaluated in the context of supporting clinical decision making.[48] Thus, authors can define the ground truth using various sources, the most important of which are evidence-informed clinical practice guidelines. These guidelines use a systematic approach to determining the certainty of the evidence as well as a structured approach to transitioning from the evidence to a decision and/or recommendation such as the GRADE (grading of recommendations, assessment, development, and evaluation) framework. Other sources that might be used to define the ground truth in CHA studies include non-evidence informed clinical practice guidelines, or guidelines that do not use a systematic approach such as GRADE.[49] Additional sources used by authors of CHA studies to define the ground truth might include but are not limited to evidence based systematic reviews or meta-analyses, non-evidence based systematic reviews (similarly defined as the lack of a systematic approach to rating the certainty of the evidence), traditional textbooks, electronic compendiums (such as UpToDate), organisation or society websites, study investigators, or primary articles such as randomised controlled studies or cohort studies.[4]

Subitem 7b—Describe the process undertaken for generative AI-driven chatbot performance evaluation

Explanation: Once the ground truth is established, authors may further clarify how responses are classified. For instance, a study evaluating the clinical accuracy of colorectal cancer screening recommendations derived by generative AI-driven chatbots might consider responses to be correct provided that they align with local guideline recommendations on whether they are for or against colorectal cancer screening.[15] However,

model output might not simply recommend or advise against screening, but instead state that a clinical decision is "reasonable" or "appropriate."[15] Moreover, models could still provide disclaimers rather than commit to clinical guidance despite the use of follow-up prompts, or list a myriad of separate screening options. Whether authors are evaluating topics of health prevention, differential diagnosis, diagnosis, treatment, and/or general information,[4] authors should clearly outline how they plan to handle these types of circumstances, preferably determined a priori and documented in a study protocol as in subitem 1d. Subitems 7bi-iii further outline specific elements of performance evaluation which should be reported.

Subitem 7bi—State the number and characteristics of team members involved in performance evaluation

Explanation: Similar to subitem 5aii for prompt engineering, CHA authors are encouraged to describe the number of individuals involved in the evaluation of chatbot responses. Demographic characteristics of the team members performing the evaluation including any relevant expertise (as discussed in subitem 6aii) should be reported. However, specific clinical expertise of the team members should be clearly stated. While intellectual conflicts of interests are covered in subitem 12a, any performance evaluators with relevant expertise in the topic of interest should be clearly described to enable readers to interpret study findings and assess for the presence of biased evaluations of model responses. For instance, in a CHA study evaluating the clinical accuracy of recommendations for the surgical management of diverticulitis derived by a generative AI-driven chatbot, the use of two surgeons with expertise in the surgical management of diverticulitis as performance evaluators could be beneficial. But if the ground truth is defined as a particular guideline, readers may examine the evaluations more carefully in case evaluations deviate from the ground truth defined in the paper, which could suggest that the expert evaluators are biased by their own practice.

Subitem 7bii—Provide details of any patients and public involvement during the evaluation process

Explanation: As with prompt engineering (subitem 5aii), performance evaluation might differ when rated by study investigators (especially clinicians) compared to other stakeholders. For instance, in a study evaluating the clinical accuracy of LLMs in deriving patient level recommendations for the management of gastroesophageal reflux disease, the use of clinician investigators as performance evaluators could weaken the generalisability of study findings if the target users are patients with gastroesophageal reflux disease. Authors should therefore clearly state whether patients or members of the public were involved in the performance evaluation process where applicable and provide a comprehensive description of their involvement.

Subitem 7biii—State whether evaluators were masked to the identity of the generative AI-driven chatbot(s) under assessment

Explanation: Just as masking of outcome adjudicators is essential to mitigate measurement bias in a randomised controlled trial,[50] investigators of CHA studies can consider masking the identity of models during performance evaluation. Individuals evaluating model performance might have personal biases for or against certain models, particularly as they develop experience and expertise with CHA studies or more broadly, with machine learning. Thus, investigators of CHA studies should clearly state whether blinding was applied during performance evaluation. Blindingmight also apply to any potential comparators evaluated in the study, including other models/chatbots and/or human (clinician) controls. To demonstrate that blinding was effective and that adjudicators were fair in their assessments, studies might compare the alignment of judgements by measuring concordance, or design examples of better or worse responses and observe adjudicator assessments of these.

Item 8: Report how the sample size was determined

Explanation: When appropriate for comparative studies, an adequate sample size is essential in ensuring that study findings are reliable. In the context of CHA studies, investigators might compare the accuracy of health advice given by one or more chatbots, or in comparison to clinician advice. In this setting, the sample consists of the number of independent responses from one or more generative AI-driven chatbots or clinicians, to user queries. Authors of CHA studies are encouraged to describe how the sample size was obtained. The choice of the sample size requires careful planning and should be determined a priori based on both medical and statistical considerations in the context of the primary outcome under evaluation, described in further detail in subitem 9ai. If possible, the sample size should be determined using statistical methods, and all components needed for the calculation as well as prior knowledge or assumptions about the outcome, should be stated. The sample size should be reported before and after any predicted inflation for expected missing or invalid responses, as applicable.

Equally, if the generative AI model was developed or revised through additional training, the extent of training could affect model performance. The determination of the sample size related to the data used for model training should be justified as well, which might include any number of considerations depending on factors including but not limited to the quality and nature of the data.

Item 9a: Data analysis

Subitem 9a—Describe statistical analysis methods, including any evaluation of reproducibility of generative AI-driven chatbot responses

Explanation: The reporting of complete, detailed, and clear statistical analysis methods helps to ensure that

the study methodology is replicable, lending validity to study findings. Statistical analysis procedures should be described for the primary and any secondary outcome. In some scenarios, models might be unable to generate usable responses. If applicable, the statistical analysis methods should explain how any missing data were handled and the justification for any imputation procedures adopted. Authors of CHA studies should report any a priori or post hoc sensitivity analyses performed to assess the robustness and reproducibility of the findings. In the context of CHA studies and, more generally, generative AI models for alignment tasks, reproducibility is a particularly challenging objective given the impact on the responses of, among other factors, intrinsic stochasticity, prompt structure, prior context of queries, and model behavioural parameters.[51] For this reason, authors should detail any statistical methodology adopted for the evaluation of the reproducibility of the generative AI-driven chatbot if undertaken. For illustrative purposes, examples include but are not limited to measures of consistency, stability and factual accuracy.[52]

Subitem 9ai—Report the measures used for performance evaluation

Explanation: In the context of CHA studies, performance pertains to the ability of the generative AI-driven chatbot to provide responses that are aligned to the ground truth or reference standard (subitem 7a), which represents the optimal or true response to each query in relation to the primary or secondary outcome(s). Authors should clearly state which performance measures were used for the evaluation of such responses. As the field is dynamically evolving, we make no recommendation on what specific measures investigators should use, but authors should explain why they are applying the measures chosen for their study. A myriad of different measures have been used, including quantitative or mixed methods approaches.[53] For illustrative purposes, some contemporary examples have included the adoption of weighted or unweighted Cohen's κ coefficient, as appropriate, for categorical data, or Lin's concordance correlation coefficient for continuous variables. Other commonly reported diagnostic performance measures used in the context of AI performance evaluation include the F1 score, sensitivity (recall), specificity, positive predictive value (precision), negative predictive value, or area under the receiver operating characteristic curve (AUC/AUROC). If applicable, authors should state how true positives, false positives, true negatives, and false negatives were defined and how the confusion matrix was calculated as constructed. Other examples may include the relevancy, consistency, and meaningfulness of the generative AI-driven chatbot responses (intrinsic evaluation) as well as the effectiveness of the chatbot in real world applications (extrinsic evaluation).[53] These examples are not an exhaustive list, and as the field continues to grow, authors may choose to report increasingly different measures, so long as they are rationalised as described here.

Item 10: Results

Subitem 10a—Report the alignment between generative AI-driven chatbot output and ground truth/reference standard using quantitative or mixed methods approaches as applicable

Explanation: As stated in subitem 9a, authors should present the results of all planned and post hoc analyses, indicating which results apply to each category. If applicable, authors should also report the results of any analyses of reproducibility, showing whether generative AI chatbot(s) were consistent in their responses or whether variability might introduce uncertainty in expected performance. Authors also might report whether a generative AI-driven chatbot(s) presented evidence justifying its advice or recommendations. Authors should always report the estimate of the intervention (ie, the generative AI-driven chatbot) effect along with a confidence interval, thus providing a range within which the true effect is included with a specified level of confidence. Results should be reported comprehensively, using supplementary files to present granular data for every output if experiment scale necessitates summarisation in the results section. Authors should present sufficient data to replicate analyses and enable readers to explore specific strengths and weaknesses of the trialled application(s).

Subitem 10b—For responses deviating from the ground truth or reference standard, state the nature of the difference(s)

Explanation: Where applications provide advice or guidance that differs from the ground truth used to define accuracy, desirability, and safety, it is important to explore the qualitative nature of these differences. Authors may provide readers with a dataset containing each individual model output recorded during experiments with deviations from the ground truth indicated, facilitating independent analysis. At a minimum, authors should present a summary of generative AI application deviations, describing frequency, type of deviation, and magnitude of deviation. Types of deviation might relate to broad categories such as empirical factors, bias, harmful language, or be categorised with regard for the specific context of a study. For instance, studies trialling applications across a broad range of clinical specialties or subspecialties could stratify that rate of undesirable responses by those categories to explore relative strengths and weaknesses.[54] Alternatively, thematic analysis might be undertaken to explore any common denominators between deviations in application responses, such as logical inconsistency, use of incorrect external information (eg, hallucination), or failure of interpretation of the query.[55]

The magnitude of deviation can be interpreted and presented in a variety of formats. For illustrative purposes, authors might use Likert scales to quantify qualitative judgments by human or AI evaluators,[16] or provide a clear delineation between errors with minimal clinical consequences and errors with serious potential ramifications that could affect patient safety.

This list is not exhaustive, and authors should ideally prospectively define how judgements were made in a study protocol as in subitem 12d, alongside other performance metrics described in subitem 9ai.

Subitem 10c—Report the assessment for potentially harmful, biased, or misleading responses

Explanation: Evaluation of generative AI output for potentially harmful, biased, or misleading content helps readers gauge the safety and reliability of provided health advice. Results might help direct subsequent research and development to ensure that issues of equitable care provision and safety requirements (eg, oversight) are dealt with before wider deployment.[40 45 56] Harmful responses include advice that might lead to physical, emotional, or psychological harm if received by users. Biased responses reflect unfair prejudice or favouritism towards particular demographics, treatments, interventions, diagnostic procedures, or medical perspectives.[56] Misleading responses might include inaccurate information, outdated medical advice, medical advice not supported by evidence, or responses that could be misinterpreted by users.

Authors should describe the specific methods used to identify and evaluate these problematic responses in sufficient depth to replicate their experiments and analyses; to the same standard as performance evaluation (subitem 7b). Methods might include stress testing or using challenging prompts to simulate real world variation and possible user behaviour and observing whether trialled applications have the adaptability and flexibility to provide useful responses and avoid generating dangerous or harmful content. An additional example is red teaming, where humans or AI models are used to circumvent model safeguards, which is a more intensive form of stress testing that can help improve model robustness.[57] Model responses might be scored using bias detection algorithms, user feedback analysis, qualitative appraisal by researchers, or scenario based testing with sensitive or high risk queries.[54 58] As with other performance assessments, results should be quantified where possible to facilitate comparative analyses. For example, authors might report the percentage of responses flagged as potentially harmful, biased, or misleading, or Likert scale ratings from researchers or users tasked with scoring non-empirical characteristics of model output. If applicable, authors should report any strategies implemented to address identified issues. Authors are also encouraged to use a systematic approach to classifying potential harm. We endorse no specific classification system, but for illustrative purposes, one such approach might be the World Health Organization's International Classification for Patient Safety, which outlines five degrees of harm severity, from no harm to death.[59]

Item 11: Discussion

Subitem 11a—Interpret study findings in the context of relevant evidence.

Explanation: The findings of the study should be summarised and discussed in the context of the existing evidence. Authors might refer to other studies with a similar objective that have used the same or other AI tools for clinical advice; alternatively, it should report that such evidence could not be identified. The discussion should highlight the contribution of the study to the body of literature that assesses the usefulness of generative AI-driven chatbots as decision support tools for health purposes. By contextualising the study findings with existing literature, the authors can enhance the understanding of the role that generative AI might have in the dissemination of health advice in the given context.

Furthermore, the authors should compare similarities and differences in performance metrics across various contexts. Comparisons will vary depending on the purpose of the study, as well as the outcome measures used by authors (as in subitem 9ai). For illustrative purposes, one hypothetical example is a study that reports that chatbots exhibit high accuracy in symptom assessment for a given condition, but another study reports that user engagement metrics of the chatbot might not match those of traditional health advice delivery methods. Thus, the discussion of findings might need to highlight not only the effectiveness of generative AI, but also areas for improvement.

Subitem 11b—Describe the strengths and limitations of the study

Explanation: Authors should elaborate on the strengths of the study. This may be in the context of the robustness of study methodology. For instance, strengths might include the interdisciplinarity of the collaboration, or the nature of participants involved in study development. Regardless of the advantages, it is essential that authors of CHA studies clearly outline the unique elements of their study that enables readers to interpret what their study adds to the literature.

Furthermore, authors should make readers aware of study limitations in an honest, detailed, and comprehensive manner.[60] These limitations might include insufficient design and planning or problems during conduct and execution of the study such as limited sample size, unclear definitions of ground truth(s), or minimal public or patient involvement. Limitations might be substantial with important potential for impact on the internal or external validity of the study, or they could be of lesser importance, with limited expected impact on study validity. This subitem will give readers a better understanding of the credibility of the study findings and help other researchers who are performing similar research to avoid problems that they might encounter in this process.[60] In addition, investigators should discuss any limitations related to the generalisability of study findings across populations and results of analyses conducted to address applicability to vulnerable or under-represented subgroups.[61]

Subitem 11c—Describe the potential implications for practice, education, policy, regulation, and research

Explanation: Authors should provide a comprehensive explanation of how their study findings might

impact various aspects of the healthcare ecosystem. Depending on the nature and scope of the study, authors might discuss practice implications, including the integration of generative AI models into existing workflows; their effects on patient-provider interactions; potential changes in healthcare delivery and financing models; implementation barriers; and ethical, medicolegal, and regulatory considerations.[62] Educational implications should be explored, focusing on how AI chatbots might be incorporated into medical education and training,[62] implications for curriculum design, opportunities for patient empowerment, and potential advancements in patient education.[63] Research implications should also be outlined, including future areas of study, methodological improvements, the necessity for clinical validation, or potential collaborations between AI researchers and healthcare professionals.[62]

Item 12: Open science

Subitem 12a—Disclosures: report any relevant conflicts of interest for all authors

Explanation: Much like how conflicts of interest are an important source of bias in clinical outcomes research, these conflicts affect the trustworthiness of CHA studies. Conflicts of interest exist when a past, current, or expected interest imposes a substantial risk of influencing an individual's judgment, decision, or action when performing a specific duty.[64] In academia, examples of these duties might include but are not limited to rationalising the training or coding data to share or evaluating the clinical accuracy of a generative AI model.[64] Conflicts of interest refer to a benefit at the individual level, or through institutional affiliation and include financial, intellectual, or personal interests such as having specific cultural or religious beliefs.[64] Having a conflict of interest does not automatically result in an inappropriate judgment, decision, or action taking place.[64] The International Committee of Medical Journal Editors (ICMJE) avoids asking authors to judge what relationships constitute a conflict, but rather recommends that authors simply disclosure their relationships for readers to judge for themselves.[65] Authors are encouraged to report the process used to identify conflicts of interest, and state whether conflicts exist among each named coauthor. Detailed reporting of how relevant conflicts of interest were handled, as well as their real or perceived impact in the conduct and reporting of the study are expected, if applicable.

Subitem 12b—Funding: report sources of funding and their role in the conduct and reporting of the study

Explanation: Financial conflicts of interest are among the most common types of conflict of interest and often arise due to author relationships with or payment by industry.[66] Furthermore, much like how pharmaceutical companies frequently fund drug trials,[66] generative AI companies can choose to fund CHA studies. Financial conflicts of interest could be particularly relevant in the context of CHA studies, because authors may partner with industry owing to the resource-intensive nature of LLM development.[67] Indeed, beyond the level of study funding, some investigators might be subject to financial bias on an individual level because they work with industry.[66] These factors could influence the generation and dissemination of scientific knowledge in the clinical AI community, and thus should be expressly declared by authors.

Subitem 12c—Ethics: describe the process undertaken for ethical approval

Explanation: There are many potential bioethical ramifications of implementing machine learning systems in clinical decision making processes.[68] Concerns regarding the impact of include data security, accountability, biased training, and harmful model decision making (cognitive and automation bias), as well as the epistemic role of clinicians.[41 68 69] The responsible use of generative AI models such as LLMs in healthcare calls for patient centred innovation by designing studies for the purpose of benefiting patients while preserving data privacy, patient autonomy, and emphasising fairness during model training and development.[69] However, preclinical studies that do not involve patient data or involve participants recruited to generate prompts might be exempt from typical ethical approval processes associated with clinical research. Authors should therefore name the institutional research board or ethics committee that approved the study or waived a requirement for ethical approval. If applicable, authors should also describe how informed consent was obtained from participants or provide the rationale if this requirement was waived by the institutional research board or ethics committee.

Subitem 12ci—Describe the measures taken to safeguard data privacy of patient health information, as applicable

Explanation: Researchers should apply the same protection to ensure patient and study participant rights to privacy in CHA studies as they do in other types of studies. Particular concerns arise where authors access generative AI applications. For instance, authors might use patient data to train the model(s), or patient data might be used after deployment in the experiments themselves to query the generative AI-driven chatbot(s). It is often unclear whether and how inputted data in queries and material for customisation and fine-tuning are stored and used for other, undefined purposes.[70] Mitigation strategies can range from anonymisation of patient data to generation and use of synthetic data in experiments. If and when possible, authors should document any methods undertaken to safeguard data privacy if patient data were used in their study, providing sufficient detail for readers to appraise and replicate the safeguarding procedures. Equally, there might be circumstances when this documentation is not possible. For example, if a commercially available, proprietary generative AI-

driven chatbot is used, authors might not have access to the data used to train the model(s), which might or might not include patient data. If authors plan to use patient data, they should carefully consider how safeguarding can be implemented, as applicable.

Subitem 12cii—State whether permission/licensing was obtained for the use of original, copyrighted data

Explanation: Researchers should not enter works protected by copyright into generative AI models such as LLMs without permission or license to do so.[70] For instance, authors may train a model using licensed data.[70] Equally, users might inadvertently prompt these models with references to copyrighted or trademarked works.[70] Because generative AI models learn from the language and data entered by users, model outputs for one user might contain or make reference to the copyrighted or trademarked information included in the prompts inputted by another user. Authors of CHA studies are encouraged to explicitly state when permission or licensing is obtained to use copyrighted or trademarked work, information, or data, when applicable.

Subitem 12d—Protocol: provide a study protocol

Explanation: To avoid methodologically weak or biased studies, study plans must be established a priori.[71] Prospective registration of a study protocol aids in the identification and improvement of methodologically flawed study designs, and limits opportunities to spin research findings and/or manipulate analyses to generate positive findings to ameliorate the chances of publication.[71] The practice of committing to a research plan also helps to diminish the likelihood that methodological deviations occur but are under-reported, which might otherwise affect the interpretation of the plausibility of study findings.[71] Any methodological deviations from the protocol or major changes to the protocol introduced during the study should be described to help readers interpret the study findings in the context of these changes. If no study protocol was developed, investigators should clearly state this.

Subitem 12e—Data availability: state where study data, code repository, and model parameters can be accessed

Explanation: In the traditional approach to medical research, it is essential to clearly define an intervention to objectively explore the association between an intervention and any given outcome.[36] Authors of CHA studies are encouraged to clearly describe the generative AI-driven chatbot (intervention) that they are implementing to thoroughly study this association. Authors should report the full data used to train, customise, or develop generative AI models. CHA researchers should provide all available data, code, and model parameters in a freely accessible format (such as within an open repository or supplementary material) to facilitate the replicability of study methodology. Investigators should provide all datasets used in pre-training, fine-tuning or adaptation, as well as code or script files showing how these processes were implemented. For proprietary models, much of this information is not expected to be available, and authors might instead cite methods papers where details about model development are presented (if available). To replicate experiments, any code used to automate trials should be provided, in addition to prompt data (subitem 5b). To facilitate the replicability of data analysis, authors should present their code or spreadsheets used for analysis and should also provide all response data for independent replication (subitem 10b). It is understandable that authors might be able to report only part of the information included in this subitem, and in these circumstances, they are encouraged to report this limitation and explain why.

The CHART checklist consists of 12 unique items and 39 subitems for the transparent reporting of CHA studies. Items relate to the 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. Authors should report the page numbers where the requirements of each item and subitem are fulfilled, or simply direct readers to an appendix or supplement. Our CHART statement article describes how to use the CHART checklist in further detail.[35]

Discussion

CHART is an international, multidisciplinary, expert informed reporting guideline that provides guidance for researchers in the transparent reporting of CHA studies. In contrast to generic AI reporting tools, CHART applies specifically to generative AI models. CHART provides several specific checklist items related to the development of CHA studies, which typically evaluate the performance of generative AI-driven chatbots when summarising clinical evidence or providing health advice.[31] LLMs have frequently featured in CHA studies to date,[4] but technological innovation is very likely to shift this trend. Large multimodal models that combine raw text data with images or tabular data are being widely used as successors to flagship LLM/chatbots, and alternative approaches such as through liquid neural networks might yield applications with even greater abilities.[72] To respond to these changes, CHART is intended to be a living document that can be updated as needed to capture these advancements.

Checklist applicability

In CHA studies, authors usually query generative AI-driven chatbots with prompts to evaluate their ability to summarise evidence or provide clinical advice for specific topics including but not limited to health prevention, screening, differential diagnosis, diagnosis, treatment, prognosis, and/or general information.[31] Currently, user engagement with patients or clinicians in CHA studies has frequently been limited to the study investigators owing to concerns with privacy of patient health information. For instance, investigators

often develop hypothetical patient cases rather than using real patient data.[13-15] However, authors of CHA studies are beginning to move towards evaluating their use prospectively, sometimes with the use of real patient data.[48] For example, patients in a hernia clinic might be randomised to receiving preoperative advice from a clinician or a generative AI-driven chatbot for general health advice for weight loss, with the goal of reducing BMI at the time of surgery. Alternatively, patients might interact with a generative AI-driven chatbot over a period of time to evaluate its impact on their health behaviours and/or health outcomes, such as progression to diabetes among patients with a diagnosis of prediabetes. In these examples, while CHA researchers and readers of these studies might certainly require an understanding of several key elements of the CHART checklist as they relate to generative AI-driven chatbots, other methodological issues related to study design would be detailed by checklists intended for randomised controlled trials and prospective studies which are not included in the CHART checklist. Thus, future CHART extensions or implementations of relevant checklists for other study designs are planned and could be developed as the need arises.

In the interim for such study designs, CHA researchers are encouraged to use the CHART checklist to report key elements related to chatbots as an intervention relating to model access and identification, prompt engineering, query strategy, and performance evaluation while also applying dedicated reporting guidelines for methodological features related to the relevant study design components.[73 74] There is further interest in using generative AI models with clinical electronic medical records (EMRs) whereby models intake a patient's EMR profile to produce machine-understandable output.[75] Studies might apply these AI models to extract data to be used to predict patient outcomes via a traditional or AI based predictive model to predict perioperative risk.[76 77] Where authors use generative AI models for medical writing, we recommend that authors use the CANGARU (ChatGPT, generative artificial neural generative artificial reporting and use) guidelines.[78] Still, CHART applies to the current landscape of CHA studies and deals with the evolving integration of generative AI into clinical pathways for summarising evidence and providing clinical advice.[31]

Patient safety

Most published CHA studies have failed to deal with patient safety and harm.[4] Thus, several CHART checklist items look at key elements of patient safety, including the potential impact of study findings on patient harm (subitem 10c). Although reader evaluation has been widely adopted to assess the potential harm of generative AI,[40 79] this method has notable limitations owing to its inherent subjectivity.[34 40] When generative AI serves as an intervention, randomised controlled trials might be necessary to demonstrate whether it truly represents a safe alternative to conventional standards of care. Furthermore, the integration of generative AI in healthcare contexts presents considerable challenges to patient privacy, particularly when using closed-source or proprietary commercial models. In these cases, patient data might be transmitted to service providers, potentially compromising data security.[80 81] Additional CHART checklist items intended to address these issues include whether authors have taken ethical implications of conducting their study (subitem 12ci), and whether authors have implemented robust measures to safeguard the privacy of patient health information (subitem 12ci).

Limitations

One of the primary limitations of any Delphi process is expert selection bias. To reduce the risk that the group completing the Delphi process was non-representative of the target audience for this tool, we ensured that a large, multidisciplinary group of participants was assembled. We invited and included experts in LLMs, reporting guideline development, and health research methods from across the world, as well as patient partners to ensure that the resultant tool was generalisable not only to health researchers that would be using CHART, but also to the general public, who may be consuming the research output. However, we did not capture the professional background of each Delphi member, which remains a limitation of our study.

Additionally, the Delphi process is criticised for its over-reliance on consensus in handling complex issues, where imperfect understanding of the underlying issues could detract from the value of consensus. We reviewed each potential item via multiple multidisciplinary Zoom conferences with panel members to mitigate the risk that any of the potential checklist items was overlooked or misunderstood. All participants had the opportunity to review potential checklist items at length, which helped navigate some of the complex issues that arose during this process. While the possibility of response bias exists during large forums such as this, three separate consensus meetings were held to provide as much time and space to discuss each potential issue at length. Consensus meetings were also run by senior investigators with substantial experience in navigating these processes to mitigate biases arising from social desirability and difficulty in having in-depth discussions about complex issues, especially as they pertain to the rapidly evolving field of generative AI.

Another limitation involves potential for selection bias in identifying relevant studies for the scoping review that informed the initial checklist items. The 137 studies included in the scoping review might not fully represent all CHA studies, especially in the rapidly evolving body of literature in the field of generative AI. To minimise this bias, we followed a systematic approach for study identification and selection, yet some relevant work might have been overlooked. This approach could limit the generalisability of the guideline to certain types of studies, particularly those involving cutting-edge or proprietary technologies.

Furthermore, as generative AI models are continuously updated, variables such as model version, data sources, and training algorithms could affect performance outcomes. These changes could obscure the association between the chatbot's design and its clinical effectiveness, potentially affecting the data and publications derived from their use. To manage this, the CHART guideline includes specific recommendations for documenting model versions and updates, but this does not fully eliminate the risk that external validity might be compromised owing to evolving technology. Another limitation is that CHART guideline, in its current form, is leaning towards a unimodal data type (ie, text), both in the form of prompts and outputs. However, considering its living nature, it could be balanced for both unimodal and multimodal generative AI that offers clinical advice, in the future.

The CHART checklist has some intrinsic limitations. Firstly, it may be daunting to apply when reporting a CHA study. A first barrier could be the absence of information to provide, should the study plan have been suboptimally detailed with no plan for relevant components of CHA studies. To minimise this issue, we suggest that authors in the field familiarise themselves with this reporting checklist early in the process, ensuring that all the relevant research components needed to support optimal reporting are included. A second barrier is the space requirement to be able to provide all the details mandated by the CHART checklist. While many journals offer unlimited space for online only appendices and supplements, space in the main manuscript might still be an issue. Thirdly, high quality reporting of an experiment is a time-consuming activity, although we consider it to be a valuable investment to optimise the interpretability and reliability of study findings. In the space of clinical trials and systematic reviews, reporting guidelines have undoubtedly improved methodological standards. Still, the clinical application of generative AI is a nascent field, and familiarity with the use of reporting guidelines could take time. We have done our best to only keep essential indicators, and we have provided this explanation document exactly to help understand its rationale and anticipated value.

Conclusion

Generative AI chatbots occupy a unique position in healthcare AI, offering interactive, language based interfaces for health information and advice, with the potential to substantially change ways of delivering medical education and patient care.[2 41] This novel approach necessitates specialised reporting guidelines. CHART distinguishes itself from generic AI reporting tools by looking at the specific nuances of generative AI in healthcare contexts. Detailed guidance is provided on elements such as prompt engineering, query strategies, and evaluation of AI-generated responses in healthcare settings. By promoting standardised reporting, CHART enhances transparency and reproducibility in this field. This standardisation is crucial for building trust among healthcare professionals, patients, and regulatory bodies. Additionally, CHART is positioned to evolve alongside AI advancements. Future expansions may address other types of generative AI applications in healthcare,[82] or hybrid approaches combining generative AI with other AI technologies.[83] CHART's adaptability as a living document is key to its relevance in the rapidly evolving landscape of AI in healthcare. This flexibility ensures that CHART will continue to provide valuable guidance for researchers, enhancing the quality and reliability of studies on generative AI chatbots in healthcare.

We thank the *First Cut* competition organisers and the postgraduate medical education committee at McMaster University for financially supporting the development of this project; and members of the advisory committee for their invaluable time and effort devoted to the development of the Chatbot Assessment Reporting Tool.

Contributors: BH, DC, A-WC, CL, and GG 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. BH is the guarantor and is responsible for the overall content, 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.

Funding: We received support from the *First Cut* competition and the postgraduate medical education committee at McMaster University. Neither funding sources were involved in the design, conduct, or reporting of this reporting guideline.

Competing interests: All authors have completed the ICMJE uniform disclosure form at www.icmje.org/disclosure-of-interest/ and declare: support from the *First Cut* competition and the postgraduate medical education committee at McMaster University for the submitted work. GSC is a National Institute for Health and Care Research (NIHR) Senior Investigator. The views expressed in this article are those of the author(s) and not necessarily those of the NIHR, or the Department of Health and Social Care; AJT has received funding from HealthSense to investigate evidence based medicine applications of large language models. PM is the co-founder of BrainX; AS has received research funding from the Australian government and is co-founder of BantingMed Pty; DS is the acting deputy editor for *The Lancet Digital Health*; MM has received research funding from the Hospital Research Founding Group; TF sits on the executive committee of MDEpiNet; HF is a senior executive editor for *The Lancet*; CL is editor-in-chief of *Annals of Internal Medicine*; AF is executive managing editor and vice president, editorial operations, at *JAMA* and the JAMA Network; TF and EL are journal editors for *The BMJ*; RA is the editor-in-chief of the *International Journal of Surgery*; GS is an executive editor of *Artificial Intelligence in Medicine*; SL is a paid consultant for Astellas; DP has received research funding from the Italian Ministry of University and Research; MO is a paid consultant for Theator; TA, POV, and GG are board member of the MAGIC Evidence Ecosystem Foundation (www.magicproject.org), a non-for profit organisation 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.

Dissemination to participants and related patient and public communities: To increase the dissemination of this reporting guideline to CHA study investigators, CHART will be shared with relevant clinical journals for endorsement. CHART will also be presented at national and international meetings to increase awareness and uptake of the checklist and methodological diagram.

Provenance and peer review: Not commissioned; externally peer reviewed.

The CHART Collaborative authors: Bright Huo (Division of General Surgery, Department of Surgery, McMaster University, Hamilton, ON, Canada), Arun J Thirunavukarasu (Nuffield Department of Clinical Neurosciences, Medical Sciences Division, University of Oxford, Oxford, UK), Gary S 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), 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 Center, 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 (Director, 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*; 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, Madrid, Spain), Anthony Sunjaya (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, 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 (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).

1 Noy S, Zhang W. *Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence*. https://www.science.org
2 Shah NH, Entwistle D, Pfeffer MA. Creation and Adoption of Large Language Models in Medicine. *JAMA* 2023;330:866-9. doi:10.1001/jama.2023.14217
3 Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. *Nat Med* 2023;29:1930-40. doi:10.1038/s41591-023-02448-8
4 Huo B, Boyle A, Marfo N, et al. Large Language Models for Chatbot Health Advice Studies: A Systematic Review. *JAMA Netw Open* 2025;8:e2457879. doi:10.1001/jamanetworkopen.2024.57879
5 Elstein AS, Schwarz A. *Evidence Base Of Clinical Diagnosis: Clinical Problem Solving And Diagnostic Decision Making: Selective Review Of The Cognitive Literature*. Vol 324; 2002. https://about.jstor.org/terms
6 Kanjee Z, Crowe B, Rodman A. Accuracy of a Generative Artificial Intelligence Model in a Complex Diagnostic Challenge. *JAMA* 2023;330:78-80. doi:10.1001/jama.2023.8288
7 Berner ES, Graber ML. Overconfidence as a cause of diagnostic error in medicine. *Am J Med* 2008;121(Suppl):S2-23. doi:10.1016/j.amjmed.2008.01.001
8 Scott IA. Errors in clinical reasoning: causes and remedial strategies. *BMJ* 2009;338:b1860. doi:10.1136/bmj.b1860
9 Thomas Rodziewicz AL, Houseman B, Vaqar S, Hipskind Affiliations JE. *Medical Error Reduction and Prevention Continuing Education Activity*. https://www.ncbi.nlm.nih.gov/books/NBK499956/?report=printable
10 Wang L, Chen X, Deng X, et al. Prompt engineering in consistency and reliability with the evidence-based guideline for LLMs. *NPJ Digit Med* 2024;7:41. doi:10.1038/s41746-024-01029-4
11 Kresevic S, Giuffrè M, Ajcevic M, Accardo A, Crocè LS, Shung DL. Optimization of hepatological clinical guidelines interpretation by large language models: a retrieval augmented generation-based framework. *NPJ Digit Med* 2024;7:102. doi:10.1038/s41746-024-01091-y
12 Bhayana R. Chatbots and Large Language Models in Radiology: A Practical Primer for Clinical and Research Applications. *Radiology* 2024;310:e232756. doi:10.1148/radiol.232756
13 Huo B, Calabrese E, Sylla P, et al. The performance of artificial intelligence large language model-linked chatbots in surgical decision-making for gastroesophageal reflux disease. *Surg Endosc* 2024;38:2320-30. doi:10.1007/s00464-024-10807-w
14 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;38:5668-77. doi:10.1007/s00464-024-11155-5
15 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;14:463-9. doi:10.1007/s12553-024-00836-9.
16 Ayers JW, Poliak A, Dredze M, et al. Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. *JAMA Intern Med* 2023;183:589-96. doi:10.1001/jamainternmed.2023.1838
17 Pan A, Musheyev D, Bockelman D, Loeb S, Kabarriti AE. Assessment of Artificial Intelligence Chatbot Responses to Top Searched Queries About Cancer. *JAMA Oncol* 2023;9:1437-40. doi:10.1001/jamaoncol.2023.2947
18 Bernstein IA, Zhang YV, Govil D, et al. Comparison of Ophthalmologist and Large Language Model Chatbot Responses to Online Patient Eye Care Questions. *JAMA Netw Open* 2023;6:e2330320. doi:10.1001/jamanetworkopen.2023.30320
19 Yalamanchili A, Sengupta B, Song J, et al. Quality of Large Language Model Responses to Radiation Oncology Patient Care Questions. *JAMA Netw Open* 2024:E244630. doi:10.1001/jamanetworkopen.2024.4630
20 Moher D, Schulz KF, Simera I, Altman DG. Guidance for developers of health research reporting guidelines. *PLoS Med* 2010;7:e1000217. doi:10.1371/journal.pmed.1000217
21 Moons KGM, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med* 2015;162:W1-73. doi:10.7326/M14-0698
22 Sounderajah V, Ashrafian H, Aggarwal R, et al. Developing specific reporting guidelines for diagnostic accuracy studies assessing AI interventions: The STARD-AI Steering Group. *Nat Med* 2020;26:807-8. doi:10.1038/s41591-020-0941-1

23 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;28:924-33. doi:10.1038/s41591-022-01772-9
24 Tejani AS, Klontzas ME, Gatti AA, et al, CLAIM 2024 Update Panel. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): 2024 Update. *Radiol Artif Intell* 2024;6:e240300. doi:10.1148/ryai.240300
25 Elvidge J, Hawksworth C, Avşar TS, et al, CHEERS-AI Steering Group. Consolidated Health Economic Evaluation Reporting Standards for Interventions That Use Artificial Intelligence (CHEERS-AI). *Value Health* 2024;27:1196-205. doi:10.1016/j.jval.2024.05.006
26 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
27 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
28 Haug CJ, Drazen JM. Artificial Intelligence and Machine Learning in Clinical Medicine, 2023. *N Engl J Med* 2023;388:1201-8. doi:10.1056/NEJMra2302038
29 Beam AL, Kohane IS. Big data and machine learning in health care. *JAMA* 2018;319:1317-8. doi:10.1001/jama.2017.18391
30 Howell MD, Corrado GS, DeSalvo KB. Three Epochs of Artificial Intelligence in Health Care. *JAMA* 2024;331:242-4. doi:10.1001/jama.2023.25057
31 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. doi:10.1038/s41591-023-02656-2
32 CHART Collaborative. Protocol for the development of the Chatbot Assessment Reporting Tool (CHART) for clinical advice. *BMJ Open* 2024;14:e081155. doi:10.1136/bmjopen-2023-081155
33 Shi F, Chen X, Misra K, et al. Large Language Models Can Be Easily Distracted by Irrelevant Context. Published online 31 January 2023. http://arxiv.org/abs/2302.00093
34 Van Veen D, Van Uden C, Blankemeier L, et al. Adapted large language models can outperform medical experts in clinical text summarization. *Nat Med* 2024;30:1134-42. doi:10.1038/s41591-024-02855-5
35 The CHART Collaborative. Reporting guideline for chatbot health advice studies: the Chatbot Assessment Reporting Tool (CHART). *BMJ MED* 2025;4:e001632. doi:10.1136/bmj-2025-001632
36 Huo B, Andreou A, Onos L, Francis NK, Antoniou SA. Methods of quality assurance in multicenter trials in laparoscopic fundoplication for gastroesophageal reflux disease. *Surg Endosc* 2023;37:6711-7. doi:10.1007/s00464-023-10325-1
37 Freyer O, Wiest IC, Kather JN, Gilbert S. A future role for health applications of large language models depends on regulators enforcing safety standards. *Lancet Digit Health* 2024;6:e662-72. doi:10.1016/S2589-7500(24)00124-9
38 Tian S, Jin Q, Yeganova L, et al. Opportunities and challenges for ChatGPT and large language models in biomedicine and health. *Brief Bioinform* 2023;25:bbad493. doi:10.1093/bib/bbad493
39 Spirling A. Why open-source generative AI models are an ethical way forward for science. *Nature* 2023;616:413. doi:10.1038/d41586-023-01295-4
40 Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. *Nature* 2023;620:172-80. doi:10.1038/s41586-023-06291-2
41 Thirunavukarasu AJ. Large language models will not replace healthcare professionals: curbing popular fears and hype. *J R Soc Med* 2023;116:181-2. doi:10.1177/01410768231173123
42 Tian S, Jin Q, Yeganova L, et al. Opportunities and challenges for ChatGPT and large language models in biomedicine and health. *Brief Bioinform* 2023;25:bbad493. doi:10.1093/bib/bbad493
43 Gotlieb R, Praska C, Hendrickson MA, et al. Accuracy in Patient Understanding of Common Medical Phrases. *JAMA Netw Open* 2022;5:e2242972. doi:10.1001/jamanetworkopen.2022.42972
44 Staniszewska S, Brett J, Simera I, et al. GRIPP2 reporting checklists: tools to improve reporting of patient and public involvement in research. *BMJ* 2017;358:j3453. doi:10.1136/bmj.j3453
45 Omiye JA, Lester JC, Spichak S, Rotemberg V, Daneshjou R. Large language models propagate race-based medicine. *NPJ Digit Med* 2023;6:195. doi:10.1038/s41746-023-00939-z
46 Wang W, Tu Z, Chen C, et al. All Languages Matter: On the Multilingual Safety of Large Language Models. Published online 2 October 2023. http://arxiv.org/abs/2310.00905
47 Chen L, Zaharia M, Zou J. *How Is ChatGPT's Behavior Changing over Time?*https://github.com/lchen001/LLMDrift
48 Hager P, Jungmann F, Holland R, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nat Med* 2024;30:2613-22. doi:10.1038/s41591-024-03097-1
49 Guyatt G, Oxman AD, Akl EA, et al. GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. *J Clin Epidemiol* 2011;64:383-94. doi:10.1016/j.jclinepi.2010.04.026
50 Karanicolas PJ, Farrokhyar F, Bhandari M, Karanicolas PJ. *345 Continuing medical education formation médicale continue practical tips for surgical research*. Vol 53; 2010.
51 Wei H, He S, Xia T, Wong A, Lin J, Han M. Systematic Evaluation of LLM-as-a-Judge in LLM Alignment Tasks: Explainable Metrics and Diverse Prompt Templates. Published online 23 August 2024. http://arxiv.org/abs/2408.13006
52 Biderman S, Schoelkopf H, Sutawika L, et al. Lessons from the Trenches on Reproducible Evaluation of Language Models. Published online May 23, 2024. http://arxiv.org/abs/2405.14782
53 Abbasian M, Khatibi E, Azimi I, et al. Foundation metrics for evaluating effectiveness of healthcare conversations powered by generative AI. *NPJ Digit Med* 2024;7:82. doi:10.1038/s41746-024-01074-z
54 Thirunavukarasu AJ, Mahmood S, Malem A, et al. Large language models approach expert-level clinical knowledge and reasoning in ophthalmology: A head-to-head cross-sectional study. *PLOS Digit Health* 2024;3:e0000341. doi:10.1371/journal.pdig.0000341
55 Gilson A, Safranek CW, Huang T, et al. How Does ChatGPT Perform on the United States Medical Licensing Examination (USMLE)? The Implications of Large Language Models for Medical Education and Knowledge Assessment. *JMIR Med Educ* 2023;9:e45312. doi:10.2196/45312
56 Zack T, Lehman E, Suzgun M, et al. Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study. *Lancet Digit Health* 2024;6:e12-22. doi:10.1016/S2589-7500(23)00225-X
57 Perez E, Huang S, Song F, et al. Red Teaming Language Models with Language Models. Published online 7 February 2022. http://arxiv.org/abs/2202.03286doi:10.18653/v1/2022.emnlp-main.225
58 Chen F, Wang L, Hong J, Jiang J, Zhou L. Unmasking bias in artificial intelligence: a systematic review of bias detection and mitigation strategies in electronic health record-based models. *J Am Med Inform Assoc* 2024;31:1172-83. doi:10.1093/jamia/ocae060
59 Cooper J, Williams H, Hibbert P, et al. Classification of patient-safety incidents in primary care. *Bull World Health Organ* 2018;96:498-505. doi:10.2471/BLT.17.199802
60 Ross PT, Bibler Zaidi NL. Limited by our limitations. *Perspect Med Educ* 2019;8:261-4. doi:10.1007/S40037-019-00530-X
61 Flanagin A, Pirracchio R, Khera R, Berkwits M, Hswen Y, Bibbins-Domingo K. Reporting Use of AI in Research and Scholarly Publication-JAMA Network Guidance. *JAMA* 2024;331:1096-8. doi:10.1001/jama.2024.3471
62 *Ethics and Governance of Artificial Intelligence for Health*. World Health Organization; 2024.
63 Bragazzi NL, Garbarino S. Toward Clinical Generative AI: Conceptual Framework. *JMIR AI* 2024;3:e55957. doi:10.2196/55957
64 Akl EA, Hakoum M, Khamis A, Khabsa J, Vassar M, Guyatt G. A framework is proposed for defining, categorizing, and assessing conflicts of interest in health research. *J Clin Epidemiol* 2022;149:236-43. doi:10.1016/j.jclinepi.2022.06.001
65 Taichman DB, Backus J, Baethge C, et al. A Disclosure Form for Work Submitted to Medical Journals: A Proposal From the International Committee of Medical Journal Editors. *JAMA* 2020;323:1050-1. doi:10.1001/jama.2019.22274
66 Mathew A, Clase CM. Conflicts of Interest and the Trustworthiness of Clinical Practice Guidelines. *Clin J Am Soc Nephrol* 2022;17:771-3. doi:10.2215/CJN.04640422
67 Truhn D, Eckardt JN, Ferber D, Kather JN. Large language models and multimodal foundation models for precision oncology. *NPJ Precis Oncol* 2024;8:72. doi:10.1038/s41698-024-00573-2
68 Grote T, Berens P. On the ethics of algorithmic decision-making in healthcare. *J Med Ethics* 2020;46:205-11. doi:10.1136/medethics-2019-105586
69 Ong JCL, Chang SYH, William W, et al. Medical Ethics of Large Language Models in Medicine. *NEJM AI* 2024;1. doi:10.1056/Alra2400038.
70 Ong JCL, Chang SYH, William W, et al. Ethical and regulatory challenges of large language models in medicine. *Lancet Digit Health* 2024;6:e428-32. doi:10.1016/S2589-7500(24)00061-X
71 Bradley SH, DeVito NJ, Lloyd KE, et al. Reducing bias and improving transparency in medical research: a critical overview of the problems, progress and suggested next steps. *J R Soc Med* 2020;113:433-43. doi:10.1177/0141076820956799
72 Bidollahkhani M, Atasoy F, Abdellatef H. *LTC-SE: Expanding the Potential of Liquid Time-Constant Neural Networks for Scalable AI and Embedded Systems*.

73 Begg C, Cho M, Eastwood S, et al. *Improving the Quality of Reporting of Randomized Controlled Trials The CONSORT Statement*. https://jamanetwork.com/
74 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;335:806-8. doi:10.1136/bmj.39335.541782.AD
75 Raza MM, Venkatesh KP, Kvedar JC. Generative AI and large language models in health care: pathways to implementation. *NPJ Digit Med* 2024;7:62. doi:10.1038/s41746-023-00988-4
76 Chung P, Fong CT, Walters AM, Aghaeepour N, Yetisgen M, O'Reilly-Shah VN. Large Language Model Capabilities in Perioperative Risk Prediction and Prognostication. *JAMA Surg* 2024;159:928-37. doi:10.1001/jamasurg.2024.1621
77 Gallifant J, Afshar M, Ameen S, et al. The TRIPOD-LLM Statement: A Targeted Guideline For Reporting Large Language Models Use 2 3 Affiliations.doi:10.1101/2024.07.24.24310930
78 Cacciamani GE, Collins GS, Gill IS. ChatGPT: standard reporting guidelines for responsible use. *Nature* 2023;618:238. doi:10.1038/d41586-023-01853-w
79 Pfohl SR, Cole-Lewis H, Sayres R, et al. A Toolbox for Surfacing Health Equity Harms and Biases in Large Language Models. Published online March 18, 2024. doi:10.1038/s41591-024-03258-2
80 Brown H, Lee K, Mireshghallah F, Shokri R, Tramèr F. What Does it Mean for a Language Model to Preserve Privacy? Published online February 11, 2022. http://arxiv.org/abs/2202.05520doi:10.1145/3531146.3534642
81 Das BC, Amini MH, Wu Y. Security and privacy challenges of large language models: A survey. *ArXiv* 2022;1:1-26.
82 Wan P, Huang Z, Tang W, et al. Outpatient reception via collaboration between nurses and a large language model: a randomized controlled trial. *Nat Med* 2024;30:2878-85. doi:10.1038/s41591-024-03148-7
83 Li J, Guan Z, Wang J, et al. Integrated image-based deep learning and language models for primary diabetes care. *Nat Med* 2024;30:2886-96. doi:10.1038/s41591-024-03139-8

Web appendix 1: Candidate CHART checklist items
Web appendix 2: Fillable version of the CHART Checklist
Web appendix 3: Fillable version of the CHART Abstract Checklist
Web appendix 4: Fillable version of the CHART methodological diagram