npj | digital medicine

Comment

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-025-02113-z

Reporting guidelines for studies involving generative artificial intelligence applications: what do I use, and when?

Bright Huo, Gary S. Collins, Giovanni E. Cacciamani & Gordon Guyatt

With a growing number of studies applying generative artificial intelligence (GAI) models for health purposes, reporting standards are being developed to guide authors in this space. We describe the currently available reporting guidelines that apply to GAI models and provide an overview of upcoming reporting standards. Investigators must remain up-to-date with the most applicable tools to guide the comprehensive reporting of their research as we integrate GAI in healthcare.

The rise in publications addressing the use of general artificial intelligence (GAI), namely large language models (LLMs), for health purposes has generated the need to guide authors on transparent reporting practices[1,2]. Although LLMs currently dominate, other GAI applications such as diffusion models and large multimodal models are gaining popularity[3]. One key distinction between GAI and conventional AI is the ability of GAI to create new information based on its training data. Varying methodology and incomplete reporting among studies applying GAI for health purposes compromise the ability of readers to accurately interpret the study findings[3], which is a particularly relevant issue when evaluating the effectiveness of complex GAI platforms in a healthcare context.

GAI models are now used to address a variety of research questions across alternative study designs, which require novel reporting guidelines[4]. While more than 25 reporting guidelines address studies applying artificial intelligence or machine learning in a healthcare context, very few reporting standards apply to studies involving GAI applications in healthcare, while fewer adhere to contemporary methodological standards[5–8]. As journal editors adopt these reporting standards, investigators may be encouraged to complete and submit checklists and methodological diagrams to accompany their submissions to optimize the transparent reporting of their methods. Authors applying GAI models in healthcare must therefore carefully identify the most appropriate reporting guideline for their study, as these standards contain tailored items for studies involving GAI models[5–8]. The purpose of this article is to summarize current GAI reporting guidelines of contemporary rigor and highlight those that are in development.

Reporting guidelines for GAI. Selecting the most suitable reporting guideline will generally depend on the research aims. Figure 1 provides a list of potential research aims currently addressed by reporting guidelines. At the time of writing, LLMs are the predominant GAI model being evaluated in the healthcare context, though other popular examples include diffusion models and large multimodal models[9]. Studies involving LLMs are addressed by the Chatbot Assessment Reporting Tool (CHART), the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD)-LLM, or the Generative Artificial intelligence tools in Medical Research (GAMER)[5–8].

Clinical evidence summaries and health advice. CHART provides reporting recommendations for studies evaluating any GAI model or GAI-driven chatbot that summarizes clinical evidence and provides health advice—termed Chatbot Health Advice (CHA) studies[6,8]. CHART can also be applied to studies of standalone GAI models, provided the model interacts with users in natural language, such as through an application programming interface. Investigators should apply CHART for CHA studies evaluating a single GAI model or GAI-driven chatbot, as well as in comparative studies between multiple GAI models or chatbots[6,8]. The framework is also relevant for evaluations of tuned or fine-tuned GAI models or chatbots for tailored evidence summaries or health advice. While examples are provided in Fig. 1, CHART's scope includes clinical evidence or health advice related to health prevention, screening, diagnosis, treatment, prognosis, and general health information[6,8].

Model development, document generation, and outcome prediction. Authors may apply TRIPOD-LLM across a wide range of use cases, from de novo LLM development to using LLMs for generating medical documents or predicting outcomes using patient data[5]. The TRIPOD-LLM authors also recommend its use for studies assessing an LLM's capability in tasks such as:

- Text processing (e.g., identifying predefined categories of objects in a body of data, or named entity recognition)[5].
- Classification (e.g., determining whether a clinic note uses a patient's pronouns correctly).
- Information retrieval (e.g., training a GAI model to respond to user queries using relevant publications)[5].
- Summarization (e.g., translating clinical documents into specific languages for patients).

Figure 1 outlines further use cases, as does the original TRIPOD-LLM publication[5]. The reporting recommendations are suitable for evaluations of a single LLM or comparisons among multiple LLMs.

Applying GAI for manuscript writing. Studies discussed thus far have evaluated GAI model performance for specific study objectives. However, there is growing interest in applying GAI models to assist in manuscript writing across traditional research designs[7]. Rather than focusing on model performance, the GAMER reporting guideline provides recommendations that address studies where all or portions of a manuscript are written by a

Fig. 1 | Overview of GAI reporting guidelines[17–25].

GAI model for medical research[7]. For example, authors may apply GAMER if they apply a GAI model to assist in writing a case report. Figure 1 lists additional examples.

Strengths and limitations of current reporting guidelines. All reporting guidelines described above followed methodological guidance from the Enhancing the QUAlity and Transparency Of health Research Network; an international initiative to improve the transparency of health research[10,11]. These reporting guidelines currently apply to LLMs, while CHART and TRIPOD-LLM are designed as living documents which will be updated periodically to respond to advances in the field[5,6,8]. Authors applying conventional study designs such as randomized controlled trials or cohort studies should continue to adhere to relevant tools such as the CONsolidated Standards Of Reporting Trials (CONSORT) and the STrengthening the Reporting of OBservational studies in Epidemiology (STROBE) reporting guidelines in addition to those described here[5,12].

One strength of the CHART reporting guideline was its input from broad representation of interdisciplinary stakeholders through 531 members during the Delphi consensus. Though it is highly applicable to CHA studies, its scope is narrow. In contrast, TRIPOD-LLM applies to a multitude of use cases involving LLMs, though the applicability of each checklist item may depend on the specific use case. While the GAMER checklist is concise and specifically relevant for medical research, it may lack important items included in other reporting guidelines.

Reporting guidelines in development. There are multiple reporting guidelines in development including the ChatGPT and Artificial Intelligence Natural Large Language Models for Accountable Reporting and Use (CANGARU) reporting guideline[13]. CANGARU is being developed according to robust methodological standards involving a living systematic review, Delphi consensus, and panel consensus meetings among international, multidisciplinary stakeholders[14]. Once published, investigators may be interested in the CANGARU guidelines when using LLMs in academic research and scientific writing. The CANGARU guidelines will apply to studies within medicine, but also to those studies using LLMs for manuscript writing in other non-medical scientific sectors[14].

In health economics, investigators have initiated the ELEVATE-GenAI framework with 10 preliminary checklist items after a targeted literature review, iterative discussion, and usability testing for both systematic reviews and health economic modeling[15]. It currently consists of a structured framework and a checklist for practical implementation which uses a scoring system, with a maximum of 3 points awarded per domain. The authors are planning stakeholder consultation across various disciplines through a Delphi consensus to improve the validity of the tool[15].

In contrast, the Consolidated Criteria for Reporting Qualitative Research (COREQ) extension for LLMs (COREQ-LLM) will address studies employing LLMs for qualitative research[16]. COREQ-LLM will be developed following a systematic scoping review and Delphi consensus to identify checklist items to aid in the transparent reporting of qualitative research involving LLMs. It is anticipated that this reporting guideline will address current trends in qualitative research in which LLMs are used to support research design, data processing, analysis, interpretation, and direct interaction with qualitative data[16].

These represent the first iterations of reporting guidelines addressing the landscape of GAI research in healthcare. They address the development of GAI models as well as the use of GAI models for manuscript writing, summarizing clinical evidence, providing health advice, or predicting health outcomes using electronic health records. Clinicians, researchers, journal editors, and publishers should note that these reporting guidelines and apply to any studies evaluating the use of GAI models for health purposes. Future iterations, extensions, and/or new reporting guidelines will keep pace with the dynamically transforming nature of the field. Researchers must remain up-to-date with the literature and continue to apply the most applicable reporting standards to their work as we work toward the safe and responsible integration of GAI technology in healthcare. Journal editors and publishers must also be alert to updates in the GAI field and continue to encourage authors to adhere to relevant reporting standards. We will perform a living systematic survey of GAI-oriented reporting guidelines to help readers remain up-to-date with the dynamically evolving environment of GAI literature.

Data availability

Data sharing is not applicable to this article as no datasets were generated or analysed during the current study.

Bright Huo[1,2] ✉, Gary S. Collins[3,4], Giovanni E. Cacciamani[5,6] & Gordon Guyatt[7,8]

[1]Division of General Surgery, Department of Surgery, McMaster University, Hamilton, Canada. [2]Guidelines Committee, European Association for Endoscopic Surgery, Eindhoven, The Netherlands. [3]Department of Applied Health Sciences, School of Health Sciences, College of Medicine and Health, University of Birmingham, Birmingham, UK. [4]NIHR Birmingham Biomedical Research Centre, University Hospitals Birmingham NHS Foundation Trust and University of Birmingham, Birmingham, UK. [5]USC Institute of Urology and Catherine and Joseph Aresty Department of Urology, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA. [6]AI Center at USC Urology, University of Southern California, Los Angeles, CA, USA. [7]Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, Canada. [8]Department of Medicine, McMaster University, Hamilton, Canada. ✉e-mail: brighthuo@dal.ca

Received: 16 August 2025; Accepted: 23 October 2025;
Published online: 07 November 2025

References

1. Huo, B. et al. Reporting standards for the use of large language model-linked chatbots for health advice. *Nat. Med.* **29**, 2988 (2023).
2. Huo, B. et al. Large language models for chatbot health advice studies: a systematic review. *JAMA Netw. Open* **8**, e2457879 (2025).
3. The CHART Collaborative. Protocol for the development of the Chatbot Assessment Reporting Tool (CHART) for clinical advice. *BMJ Open* **14**, 1–7 (2024).
4. Kolbinger, F. R., Veldhuizen, G. P., Zhu, J., Truhn, D. & Kather, J. N. Reporting guidelines in medical artificial intelligence: a systematic review and meta-analysis. *Commun. Med.* **4**, 71 (2024).
5. Gallifant, J. et al. The TRIPOD-LLM reporting guideline for studies using large language models. *Nat. Med.* **31**, 60–69 (2025).
6. The CHART Collaborative. Reporting guidelines for chatbot health advice studies: explanation and elaboration for the Chatbot Assessment Reporting Tool (CHART). *BMJ* **390**, e083305 (2025).
7. Luo, X. et al. Reporting guideline for the use of Generative Artificial intelligence tools in MEdical Research: the GAMER statement. *BMJ Evid. Based Med.* **0**, 1–11 (2025).
8. The CHART Collaborative. Reporting guideline for chatbot health advice studies: the Chatbot Assessment Reporting Tool (CHART) statement. *BMJ Med.* **0**, e083305 (2025).
9. Rouzrokh, P. et al. A current review of generative AI in medicine: core concepts, applications, and current limitations. *Curr. Rev. Musculoskelet. Med.* **18**, 246–266 (2025).
10. Simera, I., Moher, D., Hoey, J., Schulz, K. F. & Altman, D. G. The EQUATOR Network and reporting guidelines: helping to achieve high standards in reporting health research studies. *Maturitas* **63**, 4–6 (2009).
11. Moher, D., Schulz, K. F., Simera, I. & Altman, D. G. Guidance for developers of health research reporting guidelines. *PLoS Med.* **7**, e1000217 (2010).
12. Yoon, W. J. et al. LCD benchmark: long clinical document benchmark on mortality prediction for language models. *J. Am. Med. Inf. Assoc.* **32**, 285–295 (2025).

13. Cacciamani, G., Gill, I. & Collins, G. ChatGPT: standard reporting guidelines for responsible use. *Nature* **618**, 1–1 (2023).
14. Cacciamani, G.E. Development of the ChatGPT, Generative Artificial Intelligence and Natural Large Language Models for ACcountable Reporting and Use (CANGARU) Guidelines. Preprint at https://arxiv.org/abs/2307.08974 (2023).
15. Fleurence, R. L. et al. ELEVATE-GenAI: reporting guidelines for the use of large language models in health economics and outcomes research: an ISPOR Working Group on Generative AI Report. *Value Health* **11**, 1–57 (2025).
16. Fehring, L. et al. Reporting of qualitative research using large language models (COREQ+LLM): protocol for an extension of the consolidated criteria for reporting qualitative research guideline. *JMIR Res. Protoc.* **14,** 1–19 (2025).
17. Washington, C. J. et al. Evaluating the effectiveness of ChatGPT and Google Gemini in providing lung cancer screening recommendations for vulnerable communities. *CHEST Pulm.* **3**, 100167 (2025).
18. Alomari, E. Evaluating ChatGPT for disease prediction: a comparative study on heart disease and diabetes. *BioMedInformatics* **5**, 33 (2025).
19. Huo, B. 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.* **38**, 5668–5677 (2024).
20. Pastrak, M. et al. Evaluation of ChatGPT performance on emergency medicine board examination questions: observational study. *JMIR AI* **4**, e67696 (2025).
21. Elliott, D. B. An editorial on myopia control, mainly written by ChatGPT. *Optom. Vis. Sci.* **101**, 233–235 (2024).
22. O'Connor, S. Open artificial intelligence platforms in nursing education: tools for academic progress or abuse? *Nurse Educ. Pr.* **66**, 103537 (2023).
23. Peng, C. et al. A study of generative large language model for medical research and healthcare. *NPJ Digit. Med.* **6**, 210 (2023).
24. Chen, S. et al. The effect of using a large language model to respond to patient messages. *Lancet Digit Health* **6**, e379–e381 (2024).
25. Han, L., Erofeev, G., Sorokina, I., Gladkoff, S. & Nenadic, G. Examining large pre-trained language models for machine translation: what you don't know about it. Preprint at https://arxiv.org/abs/2209.07417 1–12 (2022).

Author contributions

B.H. conceptualized and drafted the manuscript; G.C., G.S. and G.G. edited the manuscript.

Competing interests

B.H., G.C., G.S.C. and G.G. are all authors of the Chatbot Assessment Reporting Tool (CHART). G.C. declares equity in EditorAIPro.

Additional information

Correspondence and requests for materials should be addressed to Bright Huo.