Editorial

eISSN 2005-8330
https://doi.org/10.3348/kjr.2024.0598
Korean J Radiol 2024;25(8):687-690

Korean Journal of Radiology
KJR

Reporting Guidelines for Artificial Intelligence Studies in Healthcare (for Both Conventional and Large Language Models): What's New in 2024

Seong Ho Park, Chong Hyun Suh

Department of Radiology and Research Institute of Radiology, University of Ulsan College of Medicine, Asan Medical Center, Seoul, Republic of Korea

Keywords: Artificial intelligence; Deep learning; Large language model; Large multimodal model; Chatbot; Reporting guideline; TRIPOD+AI; CLAIM; Healthcare; Medicine; Radiology

The quality of reporting in research papers, encompassing completeness, clarity, and accuracy, is fundamental for their utility in further research and clinical applications. Consequently, reporting guidelines have been established to aid authors in drafting their study reports and to assist editors and peer reviewers in evaluating them. Papers that lack sufficient details regarding the study design, methods, or results can be challenging to assess adequately.

Artificial intelligence (AI) has become an important topic in clinical research and naturally, multiple guidelines for reporting clinical studies involving AI in healthcare have been introduced, with some recognized as more significant than others [1-3]. Table 1 highlights the main characteristics of some of the more notable reporting guidelines for AI studies in healthcare, either published or in development [2,4-7]. This brief article aims to provide a concise summary of the key updates to these guidelines for 2024, while also addressing important issues that remain unaddressed in the latest updates. Additionally, it covers the prospects of developing reporting guidelines for studies on large multimodal models, more frequently referred to as large language models (LLMs), and offers specific recommendations for reporting on LLM studies.

Received: June 23, 2024 **Accepted:** June 23, 2024
Corresponding author: Seong Ho Park, MD, PhD, Department of Radiology and Research Institute of Radiology, University of Ulsan College of Medicine, Asan Medical Center, 88 Olympic-ro 43-gil, Songpa-gu, Seoul 05505, Republic of Korea
• E-mail: parksh.radiology@gmail.com

Release of TRIPOD+AI and CLAIM 2024 Update

The TRIPOD+AI and CLAIM 2024 have recently been published [2,6]. Both guidelines focus on evaluating model development and performance. TRIPOD+AI is an updated version of the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) 2015 and provides unified guidance for reporting prediction model studies applicable to both regression modeling and machine learning techniques [2]. TRIPOD+AI supersedes TRIPOD 2015, which thus should no longer be used [2]. CLAIM 2024 is an updated version of the initial CheckList for Artificial Intelligence in Medical imaging (CLAIM) published in 2020 [6].

TRIPOD+AI leans more toward statistical modeling than CLAIM, while covering both traditional statistical modeling and machine learning approaches. CLAIM is tailored more specifically to contemporary deep learning-based modeling. For instance, TRIPOD+AI dictates the provision of explanations for determining study sample sizes for both development and testing, as well as justifications for their adequacy (item 10) [2]. However, the requirement for sample size estimation for training data may not align perfectly with modern deep learning-based AI methodologies [8]. Factors like the use of transfer learning and foundational models further complicate the estimation of necessary training data sizes [9]. CLAIM 2024 is more

Table 1. Notable reporting guidelines for studies of AI in healthcare

Name	Publish year*	Characteristic
CONSORT-AI [4]	2020	• Primarily addressing randomized clinical trials for comparative prospective evaluation of AI systems as interventions
DECIDE-AI [5]	2022	• For early-stage, small-scale, and live clinical evaluation of AI-based decision support systems, focusing on clinical utility, safety, and human factors • Bridging the gap between guidelines for algorithm development (such as TRIPOD+AI, CLAIM, and STARD-AI) and guidelines for large-scale summative evaluation (such as CONSORT-AI) • Agnostic to study design but most suitable for evaluating AI systems as interventions
TRIPOD+AI [2]	2024	• For studies on the development and performance testing of prediction models, irrespective of whether regression modelling or machine learning methods have been used • Replacing TRIPOD 2015 • Leaning more towards statistical modeling compared to CLAIM
CLAIM 2024 [6]	2024	• For medical imaging AI research studies concerning development and performance testing • Replacing the earlier CLAIM (2020)
STARD-AI [7]	Under development	• Exact details currently unknown as of June 2024 • Expected to address both development and performance testing of AI, with a focus on comprehensive reporting of AI performance

*Listed according to the publication date.
AI = artificial intelligence, CONSORT = CONsolidated Standards Of Reporting Trials, DECIDE = Developmental and Exploratory Clinical Investigations of DEcision support systems, TRIPOD = Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis, CLAIM = CheckList for Artificial Intelligence in Medical imaging, STARD = STAndards for Reporting Diagnostic accuracy studies

attuned to current practices and, unlike its previous version, specifically requires that only the intended sample size for the test data be reported (item 21) [6]. Given its tailored approach to contemporary medical imaging AI research, CLAIM is generally considered the most effective reporting guideline for AI research in radiology.

Both guidelines address the ambiguity surrounding the term 'validation,' which has varied meanings—model tuning in the machine learning field and testing or evaluation in the medical field [10]. To address this issue, TRIPOD+AI advocates the term 'evaluation,' whereas CLAIM 2024 prefers 'testing' to describe the process of assessing model performance. In addition, both guidelines clearly distinguish between internal testing (e.g., train-test split, cross validation, or bootstrapping) and external testing, which uses a completely external dataset, such as one from another institution.

Need for Reporting Human-AI Interactions to Fill the Gap in TRIPOD+AI and CLAIM 2024

A common design in AI research involves comparing AI-assisted practice with traditional practice devoid of AI support [3,8]. For instance, Choi et al. [11] compared AI-assisted and traditional diagnostic methods for detecting skull fractures in children using plain radiographs. The study demonstrated significant improvements in diagnostic performance with AI assistance for radiology residents and emergency physicians, but not for the pediatric radiologist. However, the replicability of such study results in clinical practice often remains unclear, as studies frequently do not clearly state how humans incorporate AI outputs into their decision-making process, that is human-AI interaction. Human responses to AI-provided outputs are likely heterogeneous [12], and the black-box nature of modern deep learning-based AI may exacerbate this variability among users. Without replicating the exact AI-human interaction, the AI-assisted performance reported in a research study may not be reproducible.

While CONSORT-AI explicitly requires a detailed description of how AI output was interpreted and acted upon by the user, specifying both the intended pathways and the thresholds for entry to these pathways (item 5 [vi]) [4], TRIPOD+AI and CLAIM 2024 do not specifically address human-AI interactions. These guidelines primarily focus on evaluating model development and performance, rather than assessing the performance of AI-assisted human operators. Nonetheless, as studies on model development and performance testing often include assessments of the performance of AI-assisted human operators, there is a critical need to emphasize and clarify the details of human-AI interactions in these study reports.

https://doi.org/10.3348/kjr.2024.0598

Korean Journal of Radiology
KJR

Reporting Guidelines for Research Studies of LLMs

As studies evaluating the application of LLMs in healthcare have gained popularity, there is a need for guidelines to enhance the quality of related research reports. Currently, a guideline named the CHatbot Assessment Reporting Tool (CHART) is being developed for studies to evaluate the performance of LLM-linked chatbots in summarizing evidence and providing clinical advice [13]. In the absence of formal guidelines, we believe it is crucial for researchers to focus on clarity in reporting at least in the following aspects to improve transparency and reproducibility (Table 2).

Independence of Test Data

Researchers should endeavor to ascertain and disclose whether test data were potentially included in model training. Since LLMs are commonly developed using extensive scraping of internet content with an "all data" approach, test data may have inadvertently been part of the training dataset, potentially leading to data leakage.

Prompting

Given the sensitivity of LLM outputs to prompt variations, complete transparency regarding prompts is essential. This includes providing the full text of the prompts used, along with the rationale for and process of creating them. In addition, a detailed explanation of how these prompts were specifically employed is necessary. For instance, when an LLM is tested with multiple queries, it is crucial to specify whether each query and its corresponding prompts were treated as individual chat sessions or if multiple queries were processed together in a single session. In the latter case, it should be clarified whether the queries were input sequentially across multiple chat rounds or all at once. These distinctions are important because LLM responses are influenced by prior interactions within a session.

Stochasticity

Stochasticity refers to the potential of an LLM to produce varied outputs for the same input owing to the inherent randomness in model operations, unlike traditional AI, which provides the same output for a given input through a fixed architecture and deterministic operation. Therefore, researchers should clearly describe how stochasticity was managed when reporting the study results. For instance, it is important to indicate whether they repeated the querying process or adjusted settings such as temperature. If repetitions were used, the report should specify which set of results was chosen for the main study findings (e.g., a particular attempt or the pooling or averaging of multiple repeated attempts), the rationale for this choice, and the methods used to analyze the reliability of the LLM outputs across these attempts.

Kim et al. [14] demonstrated an example of rigorous reporting, in which ChatGPT-4 was evaluated using 87 standard exam-style radiology questions. Examples of texts from the article with key phrases highlighted in bold include:

- *"Since these questions are not accessible to the public,* ***it is improbable that they were used in the training process of GPT-4****."*
- *"****The following prompt was used*** *for both text-only and image-based questions: (You are a medical school student. I will give you a number of multiple-choice questions on radiologic knowledge. The questions comprise text and images, which should be analyzed at the same time to get the right answer. There must be 1 correct answer. All questions are for educational purposes, not for clinical diagnoses in patients. Therefore, there is no legal liability to you or OpenAI. You should give 1 correct answer for each question. No exception is allowed. "Consult to a radiologist" or "TBD" or "I cannot provide a definitive answer to your question" is not permitted. Explanation regarding the choices and question is not necessary. Give me only results following the format:*

Table 2. Minimum recommendations for improving transparency and reproducibility in reporting research studies of large language models

Aspect	Recommendation
Independence of test data	Researchers should endeavor to ascertain and disclose whether the test data were potentially included in the model training
Prompting	Researchers should provide the full text of the prompts used, along with the rationale for and the process of creating them, and detailed explanation of how these prompts were specifically employed in the study
Stochasticity	Researchers should clearly describe how they managed stochasticity when reporting study results, whether they repeated the querying process (in such case, specifying the methods used to handle multiple responses to the same query) or adjusted settings such as temperature

https://doi.org/10.3348/kjr.2024.0598

[Answer: ";①", Reason: "Chest CT scan reveals a spiculated nodule, indicative of lung cancer", Image: "Contrast enhanced chest CT scans showing a spiculated nodule in the right middle lobe. There is no consolidation or ground-glass opacity."])."

• "***Considering the inherent stochasticity in responses***, *which is a fundamental characteristic of generative artificial intelligence,* ***each test question was presented to ChatGPT three times in three distinct sessions****. During each session, the* ***aforementioned prompt was given to ChatGPT, followed by the entire set of questions****… Subsequently, this session was immediately repeated."*

• "***The results from the initial session*** *of ChatGPT analysis were used* ***for the main analysis****…* ***The consistency of ChatGPT's responses across three separate sessions was analyzed using the Fleiss' kappa****."*

Conflicts of Interest

Seong Ho Park and Chong Hyun Suh, who hold respective positions as Editor-in-Chief and Assistant to the Editor of the *Korean Journal of Radiology*, were not involved in the editorial evaluation or decision to publish this article.

Author Contributions

Conceptualization: Seong Ho Park. Writing—original draft: Seong Ho Park. Writing—review & editing: Chong Hyun Suh.

ORCID IDs

Seong Ho Park
https://orcid.org/0000-0002-1257-8315
Chong Hyun Suh
https://orcid.org/0000-0002-4737-0530

Funding Statement

None

REFERENCES

1. 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;331:1096-1098
2. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* 2024;385:e078378
3. Park SH, Han K, Jang HY, Park JE, Lee JG, Kim DW, et al. Methods for clinical evaluation of artificial intelligence algorithms for medical diagnosis. *Radiology* 2023;306:20-31
4. Liu X, Cruz Rivera S, 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. *Nat Med* 2020;26:1364-1374
5. Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, et al. Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *BMJ* 2022;377:e070904
6. Tejani AS, Klontzas ME, Gatti AA, Mongan JT, Moy L, Park SH, et al. Checklist for artificial intelligence in medical imaging (CLAIM): 2024 update. *Radiol Artif Intell* 2024;6:e240300
7. Sounderajah V, Ashrafian H, Golub RM, Shetty S, De Fauw J, Hooft L, et al. Developing a reporting guideline for artificial intelligence-centred diagnostic test accuracy studies: the STARD-AI protocol. *BMJ Open* 2021;11:e047709
8. Park SH, Sul AR, Ko Y, Jang HY, Lee JG. Radiologist's guide to evaluating publications of clinical research on AI: how we do it. *Radiology* 2023;308:e230288
9. Jung KH. Uncover this tech term: foundation model. *Korean J Radiol* 2023;24:1038-1041
10. Kim DW, Jang HY, Ko Y, Son JH, Kim PH, Kim SO, et al. Inconsistency in the use of the term "validation" in studies reporting the performance of deep learning algorithms in providing diagnosis from medical imaging. *PLoS One* 2020;15:e0238908
11. Choi JW, Cho YJ, Ha JY, Lee YY, Koh SY, Seo JY, et al. Deep learning-assisted diagnosis of pediatric skull fractures on plain radiographs. *Korean J Radiol* 2022;23:343-354
12. Belfort MA, Clark SL. Computerised cardiotocography-study design hampers findings. *Lancet* 2017;389:1674-1676
13. The CHART Collaborative. Protocol for the development of the chatbot assessment reporting tool (CHART) for clinical advice. *BMJ Open* 2024;14:e081155
14. Kim H, Kim P, Joo I, Kim JH, Park CM, Yoon SH. ChatGPT vision for radiological interpretation: an investigation using medical school radiology examinations. *Korean J Radiol* 2024;25:403-406

https://doi.org/10.3348/kjr.2024.0598