Letter to the Editor

Critical Limitations in Systematic Reviews of Large Language Models in Health Care

Zvi Weizman, MD, Prof Dr Med

Faculty of Health Sciences, Ben-Gurion University, Tel-Aviv, Israel

Corresponding Author:

Zvi Weizman, MD, Prof Dr Med
Faculty of Health Sciences
Ben-Gurion University
8 Balfour Street
Tel-Aviv 6521120
Israel
Phone: 972 544888686
Email: wzvi@bgu.ac.il

Related Articles:
Comment in: https://www.jmir.org/2025/1/e82729
Comment on: https://www.jmir.org/2025/1/e71916

J Med Internet Res 2025;27:e81769; doi: 10.2196/81769

Keywords: letter; large language models; AI; health care; review; LLM; clinical; artificial intelligence; digital health

Introduction

I read with interest the study by Li et al [1] on the implementation of large language models (LLMs) in health care, which provides clinicians with guidance for selecting appropriate models for specific tasks. Although it provides a comprehensive overview, several limitations undermine its utility for clinical decision-making.

Citation Threshold Bias

The authors exclude journals below a citation threshold of 13,000, which introduces a publication bias. It excludes innovative research from emerging or specialized journals, as documented in the methodology literature. This is problematic in a rapidly evolving field where important innovations may first appear in newer venues. While the authors note that only 8.9% (24/270) of studies reported negative results, which could affect the overall perception of their clinical effectiveness, they do not adequately account for this publication bias.

Flawed Performance Definition

The definition of "best performance" is problematic. They acknowledge that performance level in one context does not guarantee similar performance in different contexts, and therefore, they state that the frequency of "best performance" should not be interpreted as a metric for comparing models. This acknowledgment undermines their quantitative analysis. The heterogeneity in evaluation metrics, datasets, and contexts across studies renders their performance comparisons essentially meaningless, a problem well-documented in AI literature [2].

Limited Quality Assessment

The review lacks assessment of the included studies. A recent meta-analysis in medical AI has emphasized the importance of evaluating study design, validation approaches, and statistical rigor [3]. The authors' approach of simply counting "best performance" instances without considering study quality, sample sizes, or validation rigor represents a significant methodological weakness.

Conceptual and Analytical Limitations

The 5-stage linear workflow model, while organizationally useful, oversimplifies the complex and iterative nature of clinical decision-making. Modern health care delivery involves parallel processes, feedback loops, and multidisciplinary coordination that this model fails to capture, thereby limiting the practical utility of its recommendations [4].

Insufficient Discussion of Clinical Validation

They inadequately address the critical gap between research performance and clinical validation. As noted in recent systematic reviews of AI in health care, models trained and validated on research datasets face substantial deployment challenges in medical institutions due to significant differences between laboratory and clinical settings. While the authors mention this limitation, they do not adequately weigh it in their analysis.

Limited Safety and Risk Analysis

Although the authors discuss ethical concerns, their analysis of patient safety remains superficial. Recent literature emphasizes the critical importance of comprehensive risk assessment in implementing medical AI, including analysis of failure modes, error propagation, and impacts on clinical decision-making [5].

Absence of Economic Evaluation

The review lacks a comprehensive economic evaluation of LLM implementation, including cost-effectiveness analyses, resource allocation considerations, and return-on-investment assessments. These limitations significantly impact the review's clinical applicability and highlight the need for more rigorous methodological approaches in evaluating AI in health care.

Conflicts of Interest

None declared.

References

1. Li H, Fu JF, Python A. Implementing large language models in health care: clinician-focused review with interactive guideline. J Med Internet Res. Jul 11, 2025;27:e71916. [doi: 10.2196/71916] [Medline: 40644686]
2. Chang Y, Yin JM, Li JM, Liu C, Cao LY, Lin SY. Applications and future prospects of medical LLMs: a survey based on the M-KAT conceptual framework. J Med Syst. Dec 27, 2024;48(1):112. [doi: 10.1007/s10916-024-02132-5] [Medline: 39725770]
3. Liu X, Cruz Rivera S, Moher D, et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nat Med. Sep 2020;26(9):1364-1374. [doi: 10.1038/s41591-020-1034-x]
4. Sittig DF, Singh H. A new sociotechnical model for studying health information technology in complex adaptive healthcare systems. Qual Saf Health Care. Oct 2010;19 Suppl 3(Suppl 3):i68-74. [doi: 10.1136/qshc.2010.042085] [Medline: 20959322]
5. Sendak MP, Ratliff W, Sarro D, et al. Real-World Integration of a sepsis deep learning technology into routine clinical care: implementation study. JMIR Med Inform. Jul 15, 2020;8(7):e15182. [doi: 10.2196/15182] [Medline: 32673244]

Abbreviations

LLM: large language model

Edited by Tiffany Leung; This is a non–peer-reviewed article; submitted 03.08.2025; final revised version received 05.08.2025; accepted 29.08.2025; published 24.09.2025

Please cite as:
Weizman Z
Critical Limitations in Systematic Reviews of Large Language Models in Health Care
J Med Internet Res 2025;27:e81769
URL: https://www.jmir.org/2025/1/e81769
doi: 10.2196/81769