

INTERNATIONAL GRADUATE COLLOQUIUM

i-SPEAK

2026

**SPORTS AND PHYSICAL EXERCISE
ASSEMBLY OF KNOWLEDGE SHARING**

SERIES 1. 2026

COLLOQUIUM PROCEEDINGS

EXTENDED ABSTRACT



Malaysia, 10 February 2026



eISBN 978-629-7989-01-3



AI Chatbot in Healthcare: An Umbrella Review of Evidence Synthesis

Muhammad Ariq Izzuddin Abdul Jalil^{1*}, & Adam Feizrel Linoby¹

¹Faculty of Sports Science and Recreation, Universiti Teknologi MARA Negeri Sembilan, Seremban Campus, Malaysia

*Corresponding author: 2022453526@student.uitm.edu.my

Keywords: AI Chatbots; Healthcare; Umbrella Review; Evidence Synthesis; Transparency and Safety

I. INTRODUCTION

AI chatbots are increasingly used in healthcare, yet evidence remains fragmented across populations, conditions, and implementation contexts. Key gaps include limited inclusivity and low-resource representation, weak workflow and electronic health record (EHR) integration, inconsistent engagement, and uneven evaluation standards for safety, privacy, and cost-effectiveness. This umbrella review synthesises evidence to identify success factors for adoption and to appraise study quality, transparency, and whether user satisfaction aligns with clinical outcomes.

II. METHODS

We conducted an umbrella review following PRISMA 2020. Three databases (PubMed, Web of Science, and Scopus) were searched from inception to 31 January 2026 with no language restrictions. Peer-reviewed systematic reviews and meta-analyses on ChatGPT or comparable large language model (LLM) chatbots in healthcare were included. Titles, abstracts, and full texts were screened, data were extracted into structured tables, and findings were synthesised narratively by theme.

III. RESULTS AND DISCUSSION

Across reviews, effectiveness signals were generally promising but heterogeneous: 77% (23/30) of studies reported positive or mixed outcomes, with one synthesis noting $a > 20\%$ improvement in treatment compliance ($p = .04$). Benefits appeared more consistent for depressive symptoms than for anxiety or well-being, while evidence for sustained glycaemic control remained preliminary. Acceptability was often high, yet sustained engagement was limited. A large prospective sample reported 94% satisfaction but only 31% retention, suggesting that early positive user perceptions do not ensure continued use. Reported barriers included limited understanding, repetitive dialogues, voice recognition and error-handling issues, and weak integration with care processes and EHRs.

Safety and governance concerns were prominent despite limited direct evaluation. Reviews highlighted hallucinations, reliability limits, bias,

privacy and security risks, and medico-legal accountability issues. Proposed mitigations converged on expert-in-the-loop models, clinician oversight for core clinical content, explicit escalation pathways, and stronger regulatory and validation frameworks. Methodological quality and transparency constrained confidence. Reporting was inconsistent; 99.3% of evaluations involved closed-source LLMs, 65% used subjective success definitions, and 14 of 15 empathy studies used non-validated measures. Many studies had small samples ($19/31 < 100$) and design limitations, limiting replicability and comparability across syntheses.

IV. CONCLUSIONS

Evidence syntheses suggest that healthcare chatbots can improve selected outcomes and are initially well accepted, but effects are heterogeneous and long-term engagement is inconsistent. Safety, privacy, and accountability remain under-evaluated, while methodological weaknesses limit confidence. Future work should prioritise rigorous, transparent trials and implementation studies with validated measures and clear risk controls.

ACKNOWLEDGMENT

The authors gratefully acknowledge the guidance and support of supervisor Adam Feizrel Linoby Ronny Linoby throughout the conception, conduct, and reporting of this umbrella review.

REFERENCES

- [1] Grassini, E., Buzzi, M., Loporini, B., & Vozna, A. (2025). A systematic review of chatbots in inclusive healthcare: Insights from the last 5 years. *Universal Access in the Information Society*, 24, 195–203. <https://doi.org/10.1007/s10209-024-01118-x>.
- [2] Griot, M., Vanderdonckt, J., & Yuksel, D. (2025). Implementation of large language models in electronic health records. *PLOS Digital Health*, 4(12), e0001141. <https://doi.org/10.1371/journal.pdig.0001141>.
- [3] Hua, Y., Xia, W., Bates, D., Hartstein, G. L., Kim, H. T., Li, M., Nelson, B. W., Stromeyer, C. IV, King, D., Suh, J., Zhou, L., & Torous, J. (2025). Standardizing and scaffolding health care AI-chatbot evaluation: Systematic review. *JMIR AI*, 4, e69006. <https://doi.org/10.2196/69006>.
- [4] Huo, B., Boyle, A., Marfo, N., et al. (2025). Large language models for chatbot health advice studies: A systematic review. *JAMA Network Open*, 8(2), e2457879. <https://doi.org/10.1001/jamanetworkopen.2024.57879>