Role of Generative Artificial Intelligence in Assisting Systematic Review Process in Health Research: A Systematic Review

Muhammed Rashid[1], Cheng Su Yi[2], Suwapat Lawin[3], Pongsapat Limhensin[3], Suppachai Insuk[3], Sajesh K Veettil[4], Nai Ming Lai[5], Xiangyang Ye[1], Nathorn Chaiyakunapruk[1], Teerapon Dhippayom[1,3]

[1]Department of Pharmacotherapy, College of Pharmacy, University of Utah, Salt Lake City, UT, USA; [2]Independent Researcher, Kuala Lumpur, Malaysia; [3]Faculty of Pharmaceutical Sciences, Naresuan University, Thailand; [4]Department of Pharmacy Practice, School of Pharmacy, IMU University, Malaysia; [5]School of Medicine, Faculty of Health and Medical Sciences, Taylor's University, Subang Jaya, Malaysia

Background and objective

- Generative artificial intelligence (GAI) is widely used in healthcare for various purposes, including the systematic review (SR) process[1,2].
- Understanding the strengths and limitations of GAI tools and their performance in evidence synthesis is crucial for incorporating them into practice[3,4].
- This study aims to summarize the evidence on performance metrics of GAI in SR process.

Methods

- We followed the PRISMA Guidelines to report this systematic review and protocol is registered in PROSPERO (CRD42024603732)
- PubMed, EMBASE, and ProQuest Dissertations & Theses Global were searched from their inception up to May 2024
- We considered all experimental studies that employed the application of GAI in any process of conduct of SR in healthcare that compared with another AI model, new AI version, non-AI or human reviewers.
- The outcomes of interest were performance metrics such as accuracy, sensitivity, specificity, recall, precision, inter-and-intra-rater reliability (agreement), Kappa and F-statistics.
- Modified Quality Assessment of Diagnostic Accuracy Studies 2 (QUADAS-2)[5] was employed to assess quality of studies that used GAI in study selection process.
- We summarized the findings of the included studies using a narrative approach.
- The study selection, data extraction and the quality assessment were performed by two independent reviewers and any disagreements were resolved in consultation with third reviewers
- The Methodology is depicted in **Figure 1**

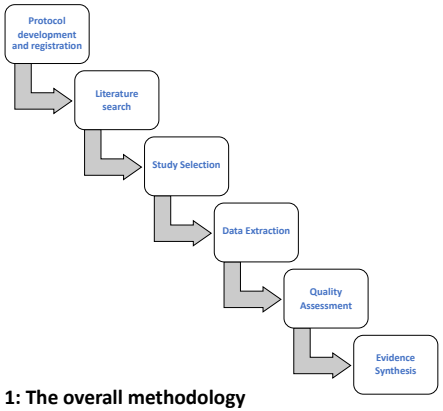

Figure 1: The overall methodology

Results

- A total of 8 studies[6-13] out of 3663 records published were included (**Figure 2**).
- The included studies used multiple methods of prompt development, evaluation, reliability and model training. Three studies used GAI for study selection alone (**Figure 3**).
- One study used GAI for PICO development, literature search, data extraction, risk of bias assessment, and both study selection and data extraction (**Figure 3**).
- GPT-3.5 and GPT-4 demonstrated good accuracy in PICO question formulation.
- The performance of GAI in the study selection process varied across studies.
- Though GPT-4 had a better performance in Tit/Abs screening, performance was low in full-text screening and combined Tit/Abs and full-text screening.
- This variation may be attributed to different prompts used, field of study, and nature of performance assessment.
- GPT-3.5 has good agreement with human reviewers in extracting simple information, but not with complex information.
- There was a lower agreement between the Cochrane SRs and GPT-4 in performing risk of bias assessment using ROBINS-I.
- GAI studies focused on selection process had a low risk of bias based on modified QUADAS-2

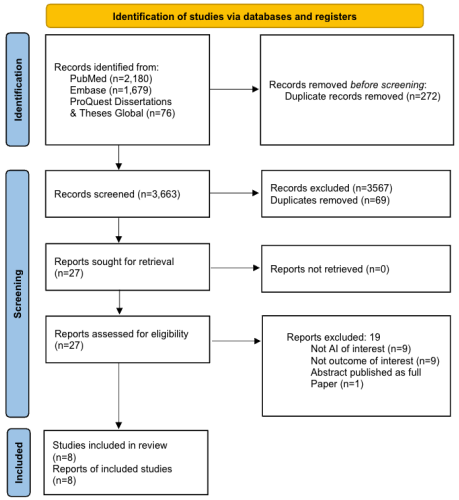

Figure 2: The PRISMA Flow diagram for study selection

Results

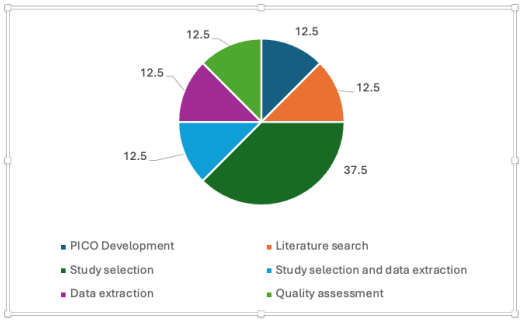

Figure 3: Role of AI in systematic review Process

Discussion and Conclusion

- This review suggests that GAI may be used in PICO question formulation and data extraction, especially simple information, as GAI cannot fully replace human reviewers in these tasks and we need more evidence to integrate the use of GAI alone in study screening in view of mixed findings.
- GAI performs well for extraction of simple information such as country and risk factors but appears to perform poorly in extracting complex details.
- The current evidence did not provide a promising result in quality assessment and none of the studies assessed the performance of GAI in data synthesis.
- A key finding across the included studies is the consensus that GAI functions best as an assistant to human reviewers rather than a standalone tool.
- Inadequate performance of GAI models hinters its standalone application in practice.
- There is no standardized methodology to develop and assess the GAI models, which limits the generalizability.

References

1. Halevi G et al., Publishing Research Quarterly. 2020; 36: 523-37.
2. Borah R, et al. BMJ Open. 2017; 7: e012545.
3. Blaizot A, et al., Res Synth Methods. 2022; 13: 353-62.
4. van Dijk SHB, et al., BMJ Open. 2023; 13: e072254.
5. Whiting PF et al., Ann Intern Med. 2011;155(8):529-536.
6. Chelli M, et al., J Med Internet Res. 2024; 26: e53164.
7. Demir GB, et al., Eur J Orthod. 2024; 46.
8. Gargari OK, et al., BMJ Evid Based Med. 2024; 29: 69-70.
9. Gue CCY, et al., Syst Rev. 2024; 13: 135.
10. Guo E, et al., J Med Internet Res. 2024; 26: e48996.
11. Gwon YN, et al., JMIR Med Inform. 2024; 12: e51187.
12. Hasan B, et al., BMJ Evid Based Med. 2024.
13. Khraisha Q, et al., Res Synth Methods. 2024; 15: 616-26.

muhammed.rashid@pharm.Utah.edu