Artificial Intelligence as a New Research Ally?

Performing AI-Assisted Systematic Literature Reviews in Health Economics

van Mossel, Sietse; Oude-Wolcherink, Martijn Johan; de Feria Cardet, Rafael Emilio; de Geus-Oei, Lioe Fee; Vriens, Dennis; Koffijberg, Hendrik; Saing, Sopany

DOI
10.1007/s40273-025-01481-4

Publication date
2025

Document Version
Final published version

Published in
PharmacoEconomics

Citation (APA)
van Mossel, S., Oude-Wolcherink, M. J., de Feria Cardet, R. E., de Geus-Oei, L. F., Vriens, D., Koffijberg, H., & Saing, S. (2025). Artificial Intelligence as a New Research Ally? Performing AI-Assisted Systematic Literature Reviews in Health Economics. *PharmacoEconomics*, *43*(6), 647-650. Article e072254. https://doi.org/10.1007/s40273-025-01481-4

PharmacoEconomics
https://doi.org/10.1007/s40273-025-01481-4

SHORT COMMUNICATION

Artificial Intelligence as a New Research Ally? Performing AI-Assisted Systematic Literature Reviews in Health Economics

Sietse van Mossel[1,2] · Martijn Johan Oude-Wolcherink[3] · Rafael Emilio de Feria Cardet[4] · Lioe-Fee de Geus-Oei[1,2,5] · Dennis Vriens[6] · Hendrik Koffijberg[3] · Sopany Saing[3,4]

Accepted: 5 March 2025

1 Introduction

Systematic literature reviews (SLRs) are fundamental for aggregating published evidence, identifying knowledge gaps and informing health economic evaluations, especially in the field of novel diagnostics where randomised controlled trials are commonly absent [1]. The global publishing output within the field of novel diagnostics is increasing rapidly owing to increasing efforts towards precision medicine. Simultaneously, the global publishing output within the field of health economics is increasing rapidly owing to increasing efforts towards value-based healthcare and present budget constraints. The number of publications focussing on the health economic impact of diagnostics almost doubled from 86,244 in 2010 to 152,404 in 2025 (PubMed search using MeSH terms 'Diagnostic Techniques and Procedures' and 'Health Care Economics and Organizations'). Consequently, the workload of performing SLRs is increasing as the number of articles that requires screening grows larger. The average time to complete an SLR is over 15 months, while the proportion of truly relevant articles for data extraction may be as low as 1% of the total search results [2]. The high workload may reduce researchers' willingness to conduct an SLR or may lead to search strategies that are too narrow when prioritising time constraints over review quality. Moreover, it may render an SLR outdated by the time it is published.

2 Artificial Intelligence as a New Research Ally

Artificial intelligence (AI) tools to support title and abstract screening have been developed to reduce screening time. A comprehensive list of released AI tools is continuously updated at the Systematic Review Toolbox hosted by the NIHR Innovation Observatory [3]. Over the past 10 years, initial study results show that when AI is appropriately used [4], it may provide substantial time savings by only requiring manual screening of a relevant subset of papers [5–10]. Recent research also shows that such semi-autonomous screening processes may be more reliable than fully autonomous screening [11]. Semi-autonomous screening requires repeated input from reviewers to confirm the articles' relevance. Additionally, reviewers decide whether to stop screening based on predefined stopping rules. Determining the optimal point to stop screening can be challenging and affects the screening process' transparency and workload. Simultaneously, guidance to using AI assistance in the screening process is increasing [12]. Practical guidance details how to prepare for AI-assisted screening including search strategy, database selection and retrieval of records [13], how to build an initial training set and to proceed screening [13], and how to select stopping rules [14, 15].

However, adoption barriers exist due to a lack of trust in software [16]. Studies that validate the use of AI tools for SLRs in health economics are limited. To contribute to validation efforts, this research letter describes the performance of AI-assisted screening in a post-hoc analysis of our

✉ Sopany Saing
s.saing@utwente.nl

1 Department of Radiology, Leiden University Medical Centre, Leiden, The Netherlands

2 Department of Biomedical Photonic Imaging, University of Twente, Enschede, The Netherlands

3 Department of Health Technology and Services Research, University of Twente, Enschede, The Netherlands

4 Centre for Health Economics Research and Evaluation, University of Technology Sydney, Sydney, NSW, Australia

5 Department of Radiation Science and Technology, Delft University of Technology, Delft, The Netherlands

6 Department of Medical Imaging, Radboud University Medical Centre, Nijmegen, The Netherlands

Published online: 10 April 2025

recently published SLR [17]. Based on our findings, next steps towards implementation and acceptation of AI-assisted SLRs in health economics are presented.

3 Post-Hoc Analysis with Open-Source ASReview Software

For the original full manual screening, a sample of 2398 articles was identified from Scopus, PubMed, iHTA and NHS-EED databases through systematic searches in February 2023, resulting in 57 articles included for a full-text review and ten articles included for data extraction [17]. For this post-hoc analysis, the following data were extracted from each article: title, abstract, and inclusion for full text screening and data extraction. The extracted information was collected and used as inputs for AI-assisted screening with open-source ASReview Software [18].

The number of relevant articles found through screening with ASReview was compared to the number of relevant articles found through a full manual screening and this proportion was defined as screening accuracy. The performance of ASReview was assessed through simulation with Makita (an extension of the ASReview Software [18]). The R Statistical Software codes used for data analysis and visualisation are published, [19] and available at https://zenodo.org/record/8217881. All simulation settings were kept at default (Naïve Bayes classifier, term frequency-inverse document frequency feature extraction, double balance strategy, maximum query strategy). Simulations were iterated 1000 times alternating the order of title and abstract screening. The position where each relevant article was found during the iterative screening was extracted and stored per iteration. Simulation results were extracted and the accuracy per number of articles screened was calculated for each iteration (Fig. 1).

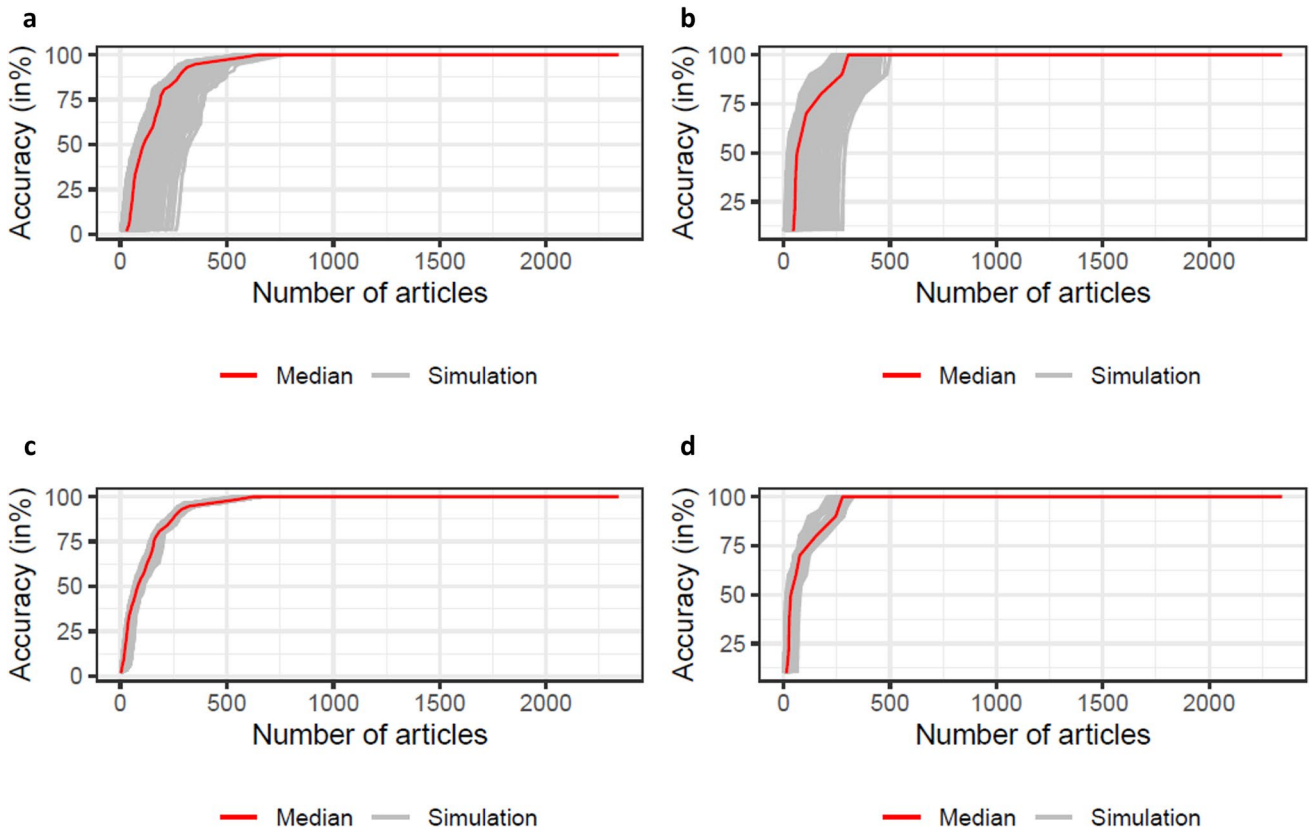

Fig. 1 Accuracy per iteration (*grey lines*) and as a median value for all iterations (*red line*). In panels **a** and **b**, ASReview Software has to start screening random articles to find relevant articles, thus no prior knowledge. In panels **c** and **d**, ASReview Software was informed with one relevant and one irrelevant article, resulting in lower variation. Median accuracy levels are, however, similar. In panels **a** and panel **c**, accuracy for articles originally included for full-text review is depicted. In panels **b** and **d**, accuracy for articles originally included for data extraction is depicted. Generally, articles included for data extraction in the original full manual screening were found early in the artificial intelligence-assisted screening process

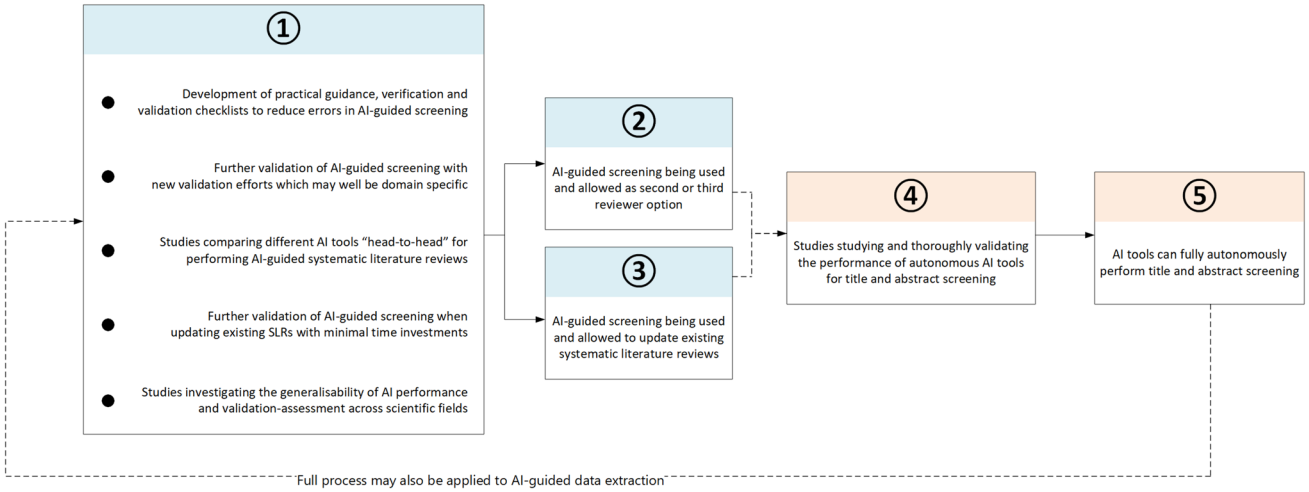

Fig. 2 Overview of the opportunities and next steps before widespread implementation of artificial intelligence (AI) tools for semi-autonomous and fully autonomous title and abstract screening. Practical guidance, verification and validation checklists should be further developed. Simultaneously, further validation efforts and head-to-head comparisons of AI tools should be performed. Thereafter, AI-guided screening might be used and allowed as a second or third reviewer option, and to update existing reviews with minimal time investments. The latter should be thoroughly validated. In parallel, similar evaluation and validation efforts may be performed focusing on AI-guided data extraction. *SLRs* systematic literature reviews

4 Discussion and Conclusions

Artificial intelligence tools have been introduced with varying levels of success in supporting title and abstract screening [20], but the question remains: what evidence is needed for wide use and acceptance? A traditional SLR is time consuming, transparent and has a low risk of biases. Awareness exists that AI tools allow for extensive database searches with more generic search terms, without increasing the workload for reviewers [2]. However, AI tools are currently not considered a valid reviewer option nor to update existing SLRs in PRISMA or Cochrane statements [21, 22], which shows that widespread acceptance and trust in AI tools require more validation efforts. Validation may well be domain specific [19], therefore, this post-hoc analysis contributes to domain-specific validation of AI tools for SLRs in health economics.

Future validation studies should compare the accuracy and expected time savings of different AI tools to demonstrate which tools may be preferred for SLRs in health economics. Different AI tools available in the literature can be compared using the same database search results. Cross-validity testing can be used to show the variability between outcomes of AI tools, with the extent to which a tool's predictions are accurate, transparent and time saving being the most important aspects to evaluate. Additionally, practical guidance is needed to support (technical) documentation, including version control, settings and stopping rules, to improve transparency and reproducibility.

The potential of AI for screening is broad. Figure 2 lists opportunities and next steps before widespread implementation. Given the current level of evidence, AI tools will likely be used as a complement to full manual screening as a second or third reviewer, or to update existing SLRs with minimal time investments [10]. With further acceptance, validation and guidance on good reporting of use, AI may completely substitute manual screening in the future.

Declarations

Funding The authors declare that no funds, grants or other support were received during the preparation of this research letter.

Conflicts of Interest Sietse van Mossel, Martijn Johan Oude-Wolcherink, Rafael Emilio de Feria Cardet, Lioe-Fee de Geus-Oei, Dennis Vriens, Hendrik Koffijberg and Sopany Saing have no conflicts of interest that are directly relevant to the content of this research letter.

Ethics Approval Not applicable.

Consent to Participate Not applicable.

Consent for Publication Not applicable.

Availability of Data and Material The software codes and datasets generated/analysed on which the reported results in this research letter rely are publicly available using the following Zenodo link: https://zenodo.org/record/8217881. The originally published search strategies and title/abstract screening criteria are publicly available using the following https://doi.org/10.1007/s40273-024-01447-y.

Code Availability The software codes on which the reported results in this research letter rely are publicly available using the following Zenodo link: https://zenodo.org/record/8217881.

Authors' Contributions All authors contributed to the study's conception and design. Material preparation, data collection and analysis were performed by SvM and MOW. The first draft of the manuscript was written by SvM and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

References

1. Cosby K, Yang D, Fineberg HV. Assessing diagnostic performance. NEJM Evid. 2024. https://doi.org/10.1056/EVIDra2300232.
2. Blaizot A, Veettil SK, Saidoung P, Moreno-Garcia CF, Wiratunga N, Aceves-Martins M, et al. Using artificial intelligence methods for systematic review in health sciences: a systematic review. Res Synth Methods. 2022;13:353–62.
3. Johnson EE, O'Keefe H, Sutton A, Marshall C. The systematic review toolbox: keeping up to date with tools to support evidence synthesis. Syst Rev. 2022. https://doi.org/10.1186/s13643-022-02122-z.
4. Van Dijk SHB, Brusse-Keizer MGJ, Bucsán CC, Van Der Palen J, Doggen CJM, Lenferink A. Artificial intelligence in systematic reviews: promising when appropriately used. BMJ Open. 2023;13:e072254.
5. Hamel C, Kelly SE, Thavorn K, Rice DB, Wells GA, Hutton B. An evaluation of DistillerSR's machine learning-based prioritization tool for title/abstract screening: impact on reviewer-relevant outcomes. BMC Med Res Methodol. 2020. https://doi.org/10.1186/s12874-020-01129-1.
6. Tsou AY, Treadwell JR, Erinoff E, Schoelles K. Machine learning for screening prioritization in systematic reviews: comparative performance of Abstrackr and EPPI-Reviewer. Syst Rev. 2020. https://doi.org/10.1186/s13643-020-01324-7.
7. Howard BE, Phillips J, Tandon A, Maharana A, Elmore R, Mav D, et al. SWIFT-active screener: accelerated document screening through active learning and integrated recall estimation. Environ Int. 2020;138:105623.
8. Gates A, Gates M, Sebastianski M, Guitard S, Elliott SA, Hartling L. The semi-automation of title and abstract screening: a retrospective exploration of ways to leverage Abstrackr's relevance predictions in systematic and rapid reviews. BMC Med Res Methodol. 2020. https://doi.org/10.1186/s12874-020-01031-w.
9. Ferdinands G, Schram R, de Bruin J, Bagheri A, Oberski DL, Tummers L, et al. Performance of active learning models for screening prioritization in systematic reviews: a simulation study into the average time to discover relevant records. Syst Rev. 2023. https://doi.org/10.1186/s13643-023-02257-7.
10. Reddy SM, Patel S, Weyrich M, Fenton J, Viswanathan M. Comparison of a traditional systematic review approach with review-of-reviews and semi-automation as strategies to update the evidence. Syst Rev. 2020. https://doi.org/10.1186/s13643-020-01450-2.
11. Gates A, Guitard S, Pillay J, Elliott SA, Dyson MP, Newton AS, et al. Performance and usability of machine learning for screening in systematic reviews: a comparative evaluation of three tools. Syst Rev. 2019. https://doi.org/10.1186/s13643-019-1222-2.
12. Marshall IJ, Wallace BC. Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. Syst Rev. 2019. https://doi.org/10.1186/s13643-019-1074-9.
13. Hamel C, Hersi M, Kelly SE, Tricco AC, Straus S, Wells G, et al. Guidance for using artificial intelligence for title and abstract screening while conducting knowledge syntheses. BMC Med Res Methodol. 2021. https://doi.org/10.1186/s12874-021-01451-2.
14. Boetje J, van de Schoot R. The SAFE procedure: a practical stopping heuristic for active learning-based screening in systematic reviews and meta-analyses. Syst Rev. 2024;13:81.
15. Callaghan MW, Müller-Hansen F. Statistical stopping criteria for automated screening in systematic reviews. Syst Rev. 2020. https://doi.org/10.1186/s13643-020-01521-4.
16. O'Connor AM, Tsafnat G, Thomas J, Glasziou P, Gilbert SB, Hutton B. A question of trust: can we build an evidence base to gain trust in systematic review automation technologies? Syst Rev. 2019. https://doi.org/10.1186/s13643-019-1062-0.
17. Van Mossel S, De Feria CR, De Geus-Oei L-F, Vriens D, Koffijberg H, Saing S. A systematic literature review of modelling approaches to evaluate the cost effectiveness of PET/CT for therapy response monitoring in oncology. Pharmacoeconomics. 2025;43:133–51.
18. van de Schoot R, de Bruin J, Schram R, Zahedi P, de Boer J, Weijdema F, et al. An open source machine learning framework for efficient and transparent systematic reviews. Nat Mach Intell. 2021;3:125–33.
19. Oude Wolcherink MJ, Pouwels XGLV, van Dijk SHB, Doggen CJM, Koffijberg H. Can artificial intelligence separate the wheat from the chaff in systematic reviews of health economic articles? Expert Rev Pharmacoecon Outcomes Res. 2023;23:1049–56.
20. Harrison H, Griffin SJ, Kuhn I, Usher-Smith JA. Software tools to support title and abstract screening for systematic reviews in healthcare: an evaluation. BMC Med Res Methodol. 2020. https://doi.org/10.1186/s12874-020-0897-3.
21. Higgins JPT, Lasserson T, Thomas J, Flemyng E, Churchill R. Methodological expectations of Cochrane intervention reviews. London: Cochrane; 2023.
22. Page MJ, Moher D, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. PRISMA explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. BMJ. 2020;2021:372.