Assessing AI and ML Tool Performance in SLRs:

A Targeted Literature Review and Performance Benchmark Framework

MSR88

Raju Gautam, PhD[1], Saeed Anwar, MSc[2], Ratna Pandey, MSc[2], Khushbu Baranwal, MSc[2], Tushar Srivastava, MSc[1]

[1]ConnectHEOR, London, UK; [2]ConnectHEOR, Delhi, India. I Email: raju.gautam@connectheor.com

connectHEOR

KEY MESSAGE

Current evidence suggests that well-configured AI/ML tools can achieve performance comparable to human reviewers in SLRs, and this study presents the first evidence informed, HTA-aligned benchmark framework to guide evaluation of AI-assisted workflows. Based on 25 studies, title/abstract screening sensitivity ranged 14%–≥99% and specificity 19%–99%, while full-text screening showed consistently high sensitivity (76–99%). Establishing evidence-based performance benchmarks is essential for the adoption of AI-driven approaches with confidence by HTA agencies.

BACKGROUND

- ✓ Systematic literature reviews (SLRs) are a cornerstone of evidence-based decision-making in healthcare, particularly for health technology assessment (HTA) agencies. However, traditional SLR processes are time-consuming and resource-intensive, often requiring extensive manual screening and data extraction.[1]
- ✓ With the growing volume of published research, AI/ML tools have emerged as promising solutions to improve efficiency and reduce workload. These tools can assist in key stages such as title/abstract screening, full-text review, and data extraction.[2]
- ✓ Despite increasing adoption, there remains uncertainty about their reliability and consistency compared to human reviewers. HTA agencies currently lack standardized benchmarks to evaluate AI performance in SLR workflows.[2]

OBJECTIVES

- Systematically summarize reported performance metrics of AI-based SLR tools across different stages of the review process
- Propose evidence-informed, practical benchmarks to guide the evaluation and implementation of AI in SLRs
- Address the gap in standardized performance criteria for HTA agencies evaluating AI-assisted SLR workflows

METHODS

Study Design
A targeted literature review was conducted to identify and evaluate studies assessing the performance of AI-based tools in SLRs.

Search Strategy
Key databases: PubMed, Google Scholar, and ISPOR database. Studies published within the last 5 years were included

Eligibility Criteria
Studies evaluating AI-driven SLR tools reporting ≥1 quantitative performance metric (accuracy, sensitivity, or specificity).

SLR stages assessed
Title/abstract screening, full-text screening, and data extraction were performed using predefined criteria to ensure consistency

Analysis
Extracted performance metrics were synthesized using a descriptive approach to identify patterns and inform evidence-based benchmarks.

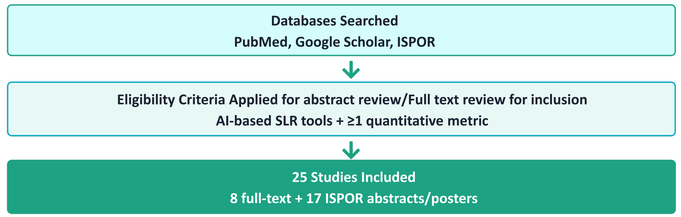
➢ In the SLRs that are aimed to support HTA submissions, sensitivity is of utmost importance to minimize missed evidence, while maintaining acceptable specificity to reduce screening burden.

Performance metrics:

- ✓ **Accuracy:** Agreement between AI and human reviewer decisions, reflecting overall classification performance
- ✓ **Sensitivity:** Ability of AI to correctly identify relevant studies, critical for minimizing missed evidence
- ✓ **Specificity:** Ability of AI to correctly exclude irrelevant studies, important for reducing screening workload

RESULTS

Figure 1: Study Selection Flow



- ✓ Studies were identified through database searches and screened based on predefined eligibility criteria, resulting in 25 included studies for final analysis.
- ✓ Most of included evidence was derived from ISPOR abstracts/posters, reflecting the emerging and evolving nature of AI applications in SLRs.

25	8	17
Studies Included	Full-Text Publications	ISPOR Abstracts

Table 1: HTA aligned PROPOSED BENCHMARKS

This table presents the proposed evidence-informed, HTA-aligned performance benchmarks for AI-assisted SLR workflows across key stages, with emphasis on high sensitivity to minimize missed evidence while maintaining acceptable specificity and accuracy

Title/Abstract Screening		
≥95% Sensitivity	≥80% Specificity	≥85% Accuracy
Full-Text Screening		
≥90% Sensitivity	≥70% Specificity	≥80% Accuracy
Data Extraction		
— Sensitivity	— Specificity	≥85% Accuracy

Table 2: Performance Metrics by SLR Stage

- ✓ It summarizes the wide variability in AI performance across SLR stages, with consistently higher sensitivity in full-text screening and greater variability in specificity and accuracy across all stages.

SLR Stage	Accuracy	Sensitivity	Specificity
Title/Abstract Screening	10%–100%	14%–≥99%	19%–99%
Full-Text Screening	40%–98.5%	76%–99%	19%–77%
Data Extraction	40%–100%	—	—

Note: Only 5 studies reported data extraction accuracy

RESULTS (CONTINUED)

Figure 2: Performance range in full-text screening

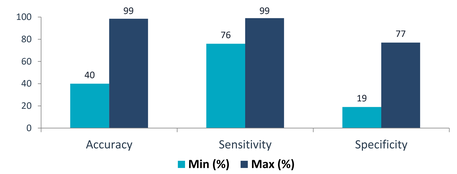
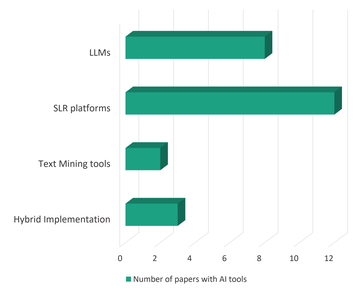


Table 3: AI Tools and Technologies Landscape



AI Tool Categories	Number of papers
Large Language Models (LLMs)	• GPT-4, GPT-3.5 Turbo, GPT-4 Turbo • Claude 3.5 (Anthropic) • Gemini 1.5 • DeepSeek-V3 • ChatGPT-based tools
AI-assisted SLR Platforms	• Rayyan AI • EasySLR • Nested Knowledge • DistillerSR • Auto SLR • Laser AI
Text Mining/ Traditional ML Tools	• Abstrackr • Algorithm-based screening tools • Neural networks, SVMs, NLP models
Custom/ Hybrid Implementations	• Flask-based applications (GPT-4o-mini API) • Proprietary AI models • RAG-based workflows (e.g. Rayyan AI)

➢ Across **25 studies**, diverse AI tools were evaluated spanning multiple technologies and platforms. These included **LLMs, AI-assisted SLR platforms, traditional ML models** and **hybrid implementations**

➢ The increasing adoption of LLM-driven and AI-assisted platforms reflects the rapid evolution of AI in SLR workflows

CONCLUSIONS

- AI is transforming SLRs by combining efficiency with near human-level performance.
- This study demonstrates that well-configured and validated AI-based tools can achieve comparable sensitivity and specificity to human reviewers, with particularly high sensitivity observed in full-text screening.
- For HTA decision-making, prioritizing sensitivity remains critical to minimize the risk of missed evidence while maintaining acceptable specificity to reduce workload. The variability observed across studies underscores the need for the standardized evaluation of AI tools.
- This study provides the first evidence-informed, HTA-aligned benchmark framework to support consistent validation and confident adoption of AI-assisted SLR workflows within transparent, human-in-the-loop approach.

REFERENCES

1. Li Y, et al. Enhancing systematic literature reviews with generative artificial intelligence: development, applications, and performance evaluation. J Am Med Inform Assoc. 2025 Apr 1;32(4):616-625.
2. Atkinson, C. F. (2023). Cheap, quick, and rigorous: Artificial Intelligence and the Systematic Literature Review. *Social Science Computer Review*, *42*(2), 376–393.
3. Carey N, et al. A text-mining tool generated title-abstract screening workload savings: performance evaluation versus single-human screening. Journal of Clinical Epidemiology, 2022; 149, 53-59.

Financial Disclosure: Authors are employees of ConnectHEOR Limited. No external funding received. No conflict of interest to declare.

Poster presented at ISPOR, Pennsylvania, Philadelphia, USA, May 17–20, 2026