

THE ECOSYSTEM OF EVIDENCE

Navigating the future
of the ecosystem of evidence

Artificial intelligence and digital transformation
in research and healthcare

11th EBHC International Conference
Taormina, 21st – 24th October 2026

#EBHC2026

Artificial Intelligence in Evidence Synthesis

Insights from Studies Within a Review (SWAR)

KM Saif-Ur-Rahman
Evidence Synthesis Ireland and Cochrane Ireland
Ireland GRADE Network
University of Galway, Ireland

30th 1996-2026 GIMBE EVIDENCE FOR HEALTH



Background

- Artificial intelligence (AI) tools are increasingly used in systematic reviews (SRs) to improve efficiency in tasks such as searching, screening, and data extraction.
- However, empirical evidence on their performance remains limited.
- Studies Within A Review (SWARs) offer a methodological framework for evaluating alternative approaches within SRs, but how AI has been used and assessed within SWARs remains unclear.

Aims

To identify and describe SWARs evaluating AI tools for stages of systematic reviews, to summarise their methodological characteristics, and to determine the extent of completed and published evidence.

Received: 18 September 2022 | Accepted: 29 November 2022

DOI: 10.1111/jebm.12505

PERSPECTIVE

WILEY

Study within a review (SWAR)

Declan Devane[1,2,3] | Nikita N. Burke[1,2] | Shaun Treweek[4] | Mike Clarke[5] | James Thomas[6] | Andrew Booth[7] | Andrea C. Tricco[8,9,10] | K. M. Saif-Ur-Rahman[1,2]

[1]Evidence Synthesis Ireland and Cochrane Ireland, University of Galway, Galway, Ireland

[2]School of Nursing and Midwifery, University of Galway, Galway, Ireland

[3]HRB-Trials Methodology Research Network, University of Galway, Galway, Ireland

[4]Health Services Research Unit, University of Aberdeen, Aberdeen, UK

[5]Northern Ireland Methodology Hub, Queen's University Belfast, Belfast, UK

[6]EPPI-Centre, UCL Social Research Institute, University College London, London, UK

[7]School of Health and Related Research (ScHARR), University of Sheffield, Sheffield, UK

[8]Li Ka Shing Knowledge Institute, St. Michael's Hospital, Unity Health Toronto, Toronto, Ontario

[9]Epidemiology Division and Institute of Health Policy, Management, and Evaluation, Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario

[10]Queen's Collaboration for Health Care Quality: A JBI Centre of Excellence, Toronto, Ontario

Methods

- We searched the SWAR Repository to identify all registered SWARs involving AI or semi-automated tools.
- We extracted data on the SR stage assessed, the type of AI tool used, planned outcomes, year of registration, and publication status.
- We conducted a narrative synthesis, with particular attention to the published SWARs.

Results

25

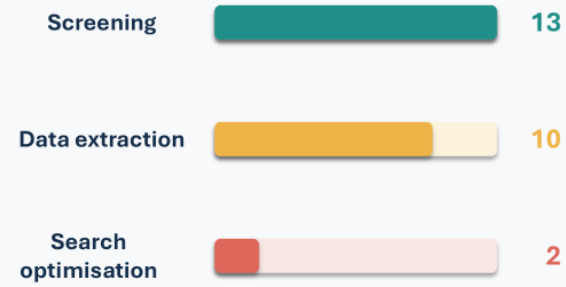
SWARs

involving AI or automation

2023–2026

registered SWARs

WHERE AI WAS APPLIED



REVIEW TYPES

Systematic reviews	Most
Rapid review	1
Diagnostic test accuracy	1
Scoping review	1
Embedded in realist reviews	2

Limited published results

2 **living SWARs**

registered as living

3 **published results**

so far



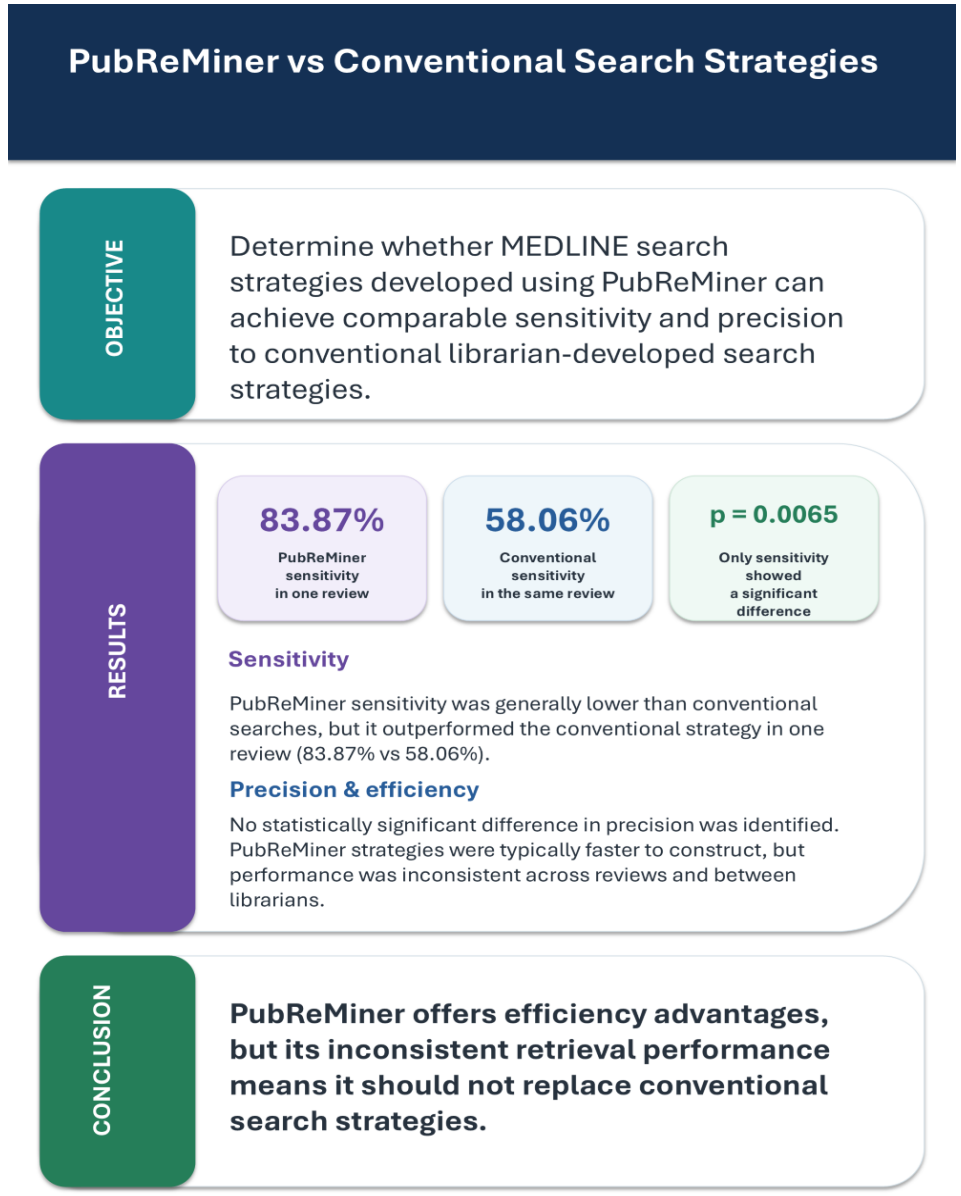
Only 3 of 25 registered SWARs had published results.

TAKEAWAY • AI is already being tested through SWAR, but published evidence remains limited.

Results

SWAR 26: Sensitivity and completeness of search strategies built using a text mining word frequency tool (PubReMiner) compared to current best practice for building a search strategy

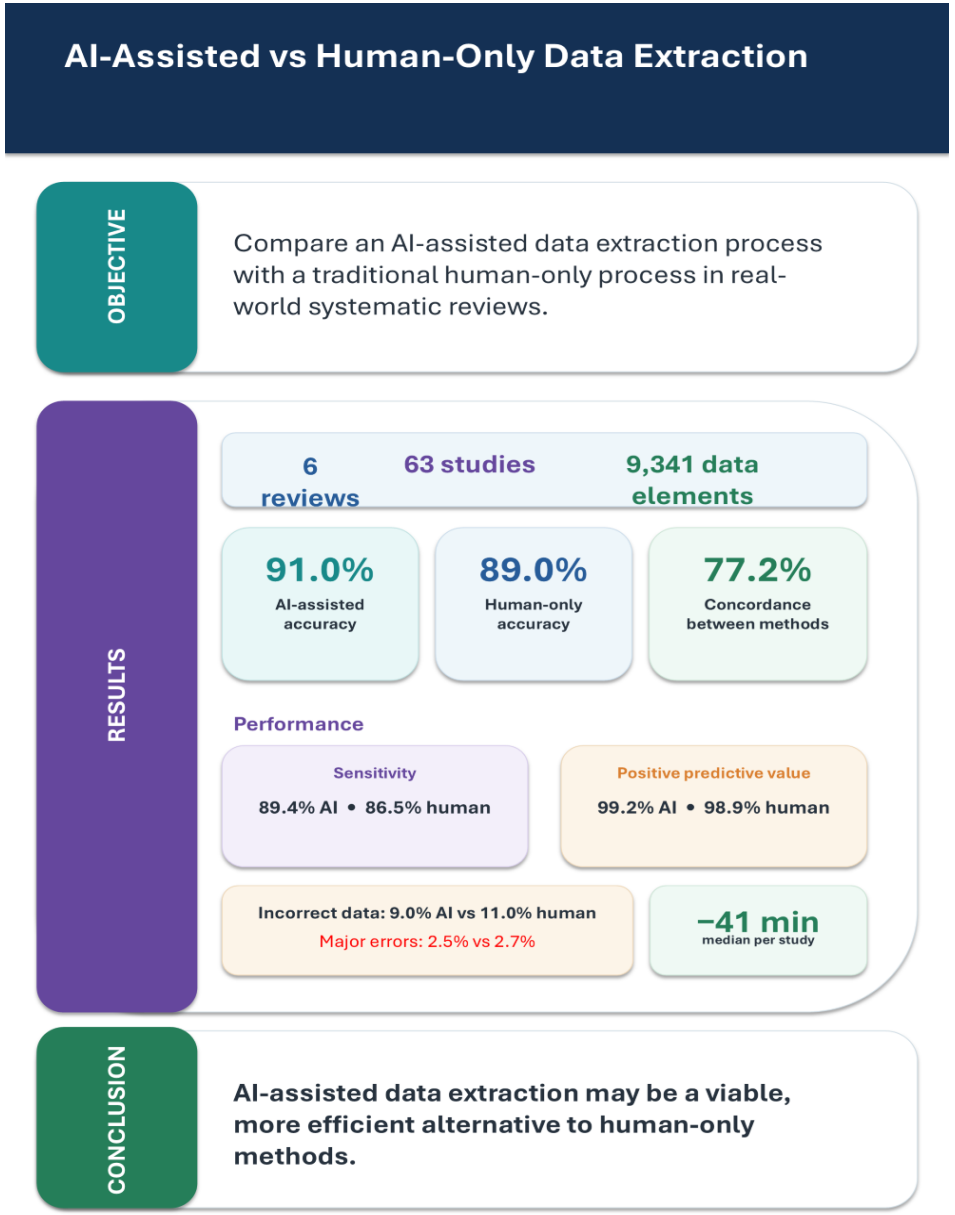
- Sensitivity of PubReMiner strategies was generally lower than that of conventional strategies.
- PubReMiner strategies were typically faster to construct but yielded inconsistent performance across reviews and between librarians.
- While PubReMiner offers efficiency advantages, however, it should not replace conventional search strategies.



Results

SWAR 28: Semi- automated data extraction for evidence syntheses using Claude 2 and Claude 3

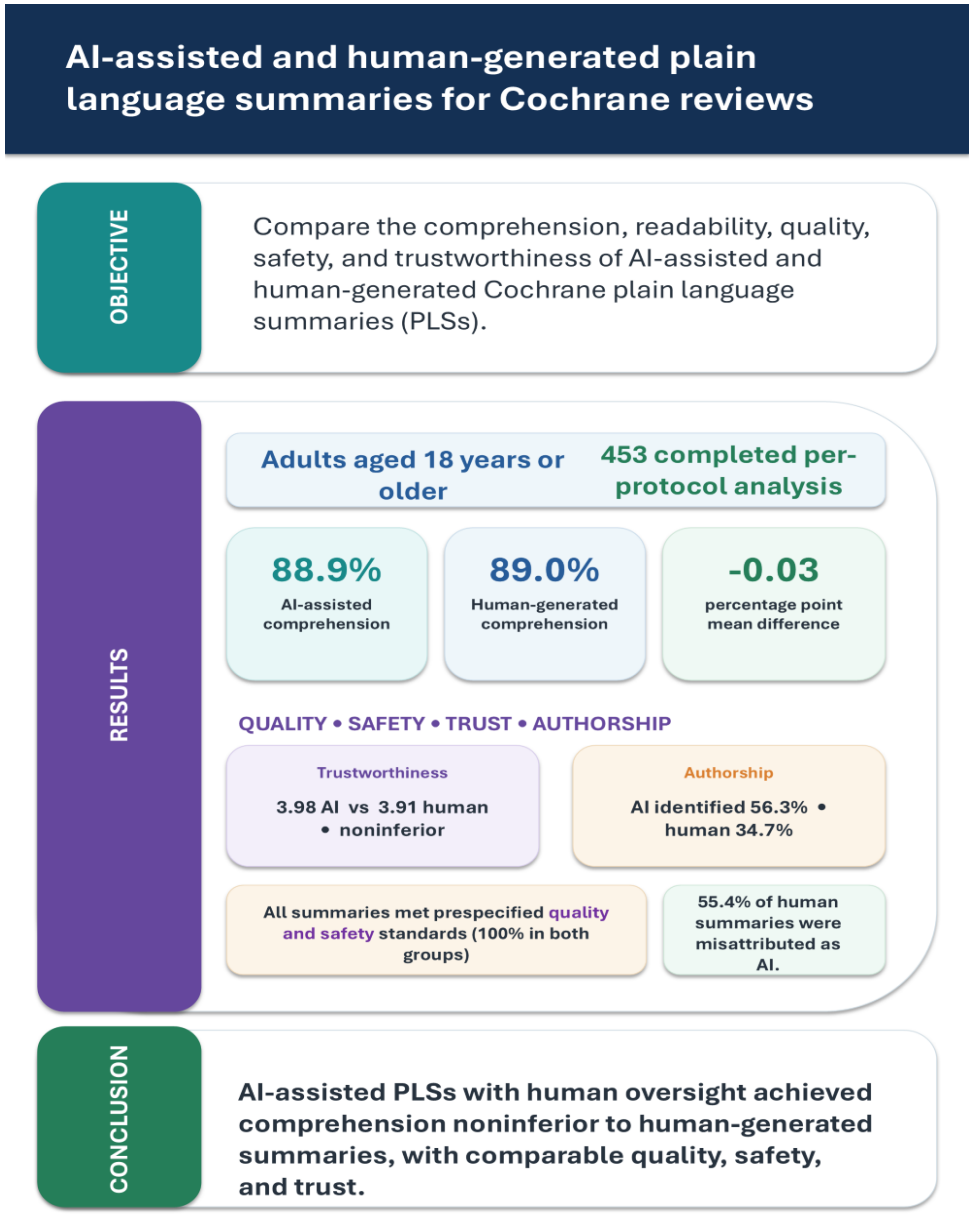
- Automation reduced reviewer workload while maintaining accuracy (91%).
- The semi-automated method achieved a sensitivity of 89.4% and a positive predictive value of 99.2%, compared with 86.5% and 98.9%, respectively, for manual extraction.
- Incorrect data were extracted in 9.0% of AI-assisted cases and 11.0% of human-only cases, with corresponding proportions of major errors of 2.5% versus 2.7%.



Results

SWAR 72: Comparison of AI-assisted and human-generated plain language summaries for Cochrane reviews: a randomised non-inferiority trial (HIET-1)

- Mean comprehension scores were 88.9% (n = 228) in the AI-assisted group and 89.0% (n = 225) in the human-generated group (mean difference −0.03 percentage points).
- All summaries met prespecified quality and safety standards (100% in both groups).
- Trustworthiness scores were comparable (3.98 vs 3.91).
- Participants demonstrated limited ability to distinguish between authorship.



Limits

- Findings are constrained by the small number of completed SWARs, limiting insight into the practical benefits and risks of AI in SR workflows.
- Most registered SWARs have no published results.
- Repository entries varied in completeness, and we relied solely on publicly available information with no access to protocols or internal documentation.

Conclusions

- AI tools are beginning to be evaluated using the SWAR method across screening, searching, data extraction, and plain-language summaries.
- Considerable gaps persist in validation, outcome reporting, and assessment across the full SR workflow.
- As AI adoption increases, rigorously designed and transparently reported SWARs are needed to establish reliable evidence on the effectiveness and efficiency of AI-assisted systematic review