A Minimum Performance-Reporting Standard for AI in Systematic Literature Review in Abstracts Presented in ISPOR

MSR231

Shilpi Swami, Hanan Irfan, Raju Gautam, Tushar Srivastava

ConnectHEOR, London, UK

Email: shilpi.swami@connectheor.com

connectHEOR

> **Of 86 AI-SLR abstracts at ISPOR Philadelphia 2026, 56% claimed time savings, only 30% reported recall and none reported work saved over sampling; we propose six minimum reporting items so that screening claims can be audited.**

1 BACKGROUND

Systematic literature reviews (SLRs) underpin health technology assessment (HTA), and study selection is among their most resource-intensive steps.[1] Artificial intelligence (AI) screening, increasingly using large language models (LLMs), is offered as the remedy, however:

- The real question is whether an AI-assisted screen is **safe enough to rely on**, because a missed eligible study could change a recommendation.
- HTA bodies require justified AI use, human oversight and validated outputs [2]; for instance, ELEVATE-GenAI sets reporting domains and PRISMA-trAIce proposes a 14-item extension.[3,4]

The gap: guidance has converged on principles, not numbers; none states the minimum metrics a primary screening study must report.

2 OBJECTIVES

We analysed recent ISPOR AI abstracts to quantify the reporting gap, asking three questions:

1. How is AI-assisted SLR screening performance reported, in metrics, values and reference standards, across a large, recent conference corpus?
2. How much incomparability does that reporting produce for a reviewer weighing one claim against another?
3. What minimum performance information would make AI screening claims auditable and decision-grade for HTA?

3 METHODS

We conducted a structured content analysis of 86 evidence-synthesis and SLR-automation abstracts from the ISPOR Philadelphia 2026 AI-poster corpus: the largest AI theme within 314 AI-related abstracts.

1 Sample → 86 of 314 AI-related abstracts

2 Code → Sub-task, method, metrics, reference standard, error audit

3 Assess → Screening statistics [5] and HTA expectations [2,4]

4 Derive → Six-item minimum reporting standard

- Coding recorded what was automated, by what system, measured how, against what ground truth, and with what check on errors.
- **Anchor 1:** eligible records are rare, so recall, precision and prevalence are needed before any accuracy figure can be read.[5]
- **Anchor 2:** emerging HTA expectation of disclosed, justified AI use, human accountability and quantified workload reduction.[2,4]
- Gaps common to both anchors became a checklist of metrics, not a tool, a model or a pass–fail threshold.

4 RESULTS

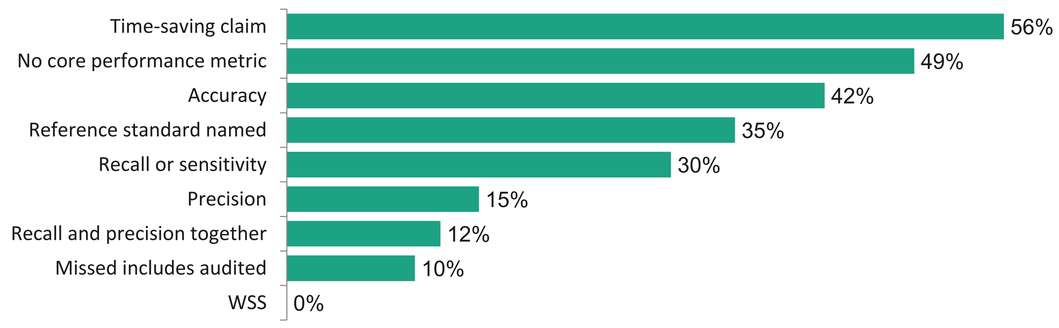
Reporting was led by **efficiency, not safety**. Accuracy was reported most often, ahead of recall or sensitivity and precision; only 12% reported recall and precision together, the minimum pair needed to characterise a screening trade-off.

56% claimed time savings (48/86)

49% reported no core performance metric (42/86)

None reported Work Saved over Sampling (WSS)

Share of the 86 abstracts reporting each item (%)



Efficiency claims led and safety metrics trailed. **Time savings of about 40% to 500% had no common baseline**, and sensitivity (10% to 100%) and accuracy (40% to 100%) ranged too widely to compare.

Each gap maps to one item of the derived minimum standard

Standard item	Gap observed in the corpus
1. Recall at a pre-specified threshold	Recall or sensitivity reported by only 30%
2. WSS at that recall	WSS absent; time savings had no common baseline
3. Precision and missed includes	Precision 15%; missed includes audited by 10%
4. Reference standard	Named by 35%, often on incompatible bases
5. Dataset prevalence	Not assessed in this corpus; needed to interpret any accuracy figure
6. Screening stage and task	Performance does not transfer across them

One validation reported sensitivity 0.99 with accuracy 0.34 and precision 0.21: **accuracy alone misleads under class imbalance**, which is why prevalence and precision are required.

5 DISCUSSION

The problem is not that AI screening performs badly; performance is **reported in ways that cannot be reconciled**, and the metric protecting against missed evidence is the least reported.

- Without the threshold, prevalence and ground truth, a screening claim is a description rather than a measurement.
- Accuracy dominates while WSS is absent; WSS obliges authors to declare a recall target and stand behind it.
- Like CONSORT, PRISMA and TRIPOD+AI, the standard mandates no tool or threshold but **makes findings appraisable.**[6–8]
- **Limitations:** one corpus and year, brief abstracts and coding judgement; the standard needs formal consensus and HTA endorsement.

CONSORT, Consolidated Standards of Reporting Trials; NICE, National Institute for Health and Care Excellence; PRISMA, Preferred Reporting Items for Systematic Reviews and Meta-Analyses; WSS, Work Saved over Sampling.

6 CONCLUSION

No abstract reported WSS and half reported no core metric, so performance is reported **too inconsistently for HTA bodies to compare tools**. A proposed six-item minimum reporting standard, modelled on the logic of CONSORT and PRISMA, would make efficiency claims auditable.

- **Authors should** report all six items, led by recall at a set threshold
- **HTA bodies and reviewers should** treat unreported recall as unverified
- **Conferences could ask** for the six items in AI-SLR abstracts
- **Next:** formal consensus and HTA endorsement of the standard

Reporting makes human accountability for AI screening checkable.

REFERENCES

1. Nussbaumer-Streit B, et al. J Clin Epidemiol. 2021;139.
2. NICE. Position statement ECD11. 2024.
3. Fleurence RL, et al. Value Health. 2025;28.
4. Holst D, et al. JMIR AI. 2025;4.
5. Cohen AM, et al. J Am Med Inform Assoc. 2006;13.
6. Hopewell S, et al. BMJ. 2025;389.
7. Page MJ, et al. BMJ. 2021;372.
8. Collins GS, et al. BMJ. 2024;385.

Financial disclosure: The authors are employees of ConnectHEOR Limited. No external funding was received. No conflict of interest to declare.

Poster presented at ISPOR Europe 2026 in Vienna, Austria, 8-11 November 2026