MSR103

Human-in-the-loop AI-assisted full-text review for safety evidence identification

Artur Nowak, Ewa Borowiack, Ewelina Sadowska, Monika Opalek
Evidence Prime, Krakow, Poland.

Key message

- AI-assisted full-text screening achieved 98.4% agreement with reviewer decisions across 500 safety-related publications.
- Complementing previous benchmark-based evaluations, this study provides evidence of the practical feasibility of integrating AI-generated screening recommendations into a structured human-in-the-loop workflow.

Background

AI-assisted screening may reduce the review burden, but its role becomes more complex at the full-text stage, where eligibility decisions may depend on information distributed across the publication and require contextual interpretation.

Recent ISPOR Good Practices Task Force findings on generative AI in systematic literature reviews emphasize that GenAI should augment rather than replace human expertise, with appropriate human oversight, accountability, and transparent reporting. [1]

Accordingly, evaluation should extend beyond model performance against a static reference standard to examine how AI recommendations can be transparently verified and incorporated into real reviewer workflows.

Our broader research program evaluates AI-assisted screening across successive stages of evidence review, including title/abstract and full-text screening. Earlier studies focused on model performance against established reference standards. [2]

The present study extends this work to a human-in-the-loop full-text workflow, in which reviewers interact with AI-generated screening recommendations, reasoning, and linked source evidence while retaining control over the final eligibility decision.

Objective

To evaluate agreement between AI decision hints and final reviewer decisions when full-text screening was conducted using an integrated human-in-the-loop workflow across five clinically diverse datasets.

Methods

Approximately 100 randomly selected full-text publications for each of five drugs were included. The drugs were selected to provide a heterogeneous evaluation dataset, spanning older and newer medicinal products, different therapeutic areas and clinical contexts and different publication periods.

This was intended to test the workflow across a range of evidence profiles rather than within a single therapeutic area or publication type. [Table 1]

Before formal evaluation, the Laser AI eligibility-based decision logic was iteratively refined through expert-guided testing. An experienced reviewer then completed a training phase to learn the Laser AI workflow and how to interpret AI-generated screening outputs.

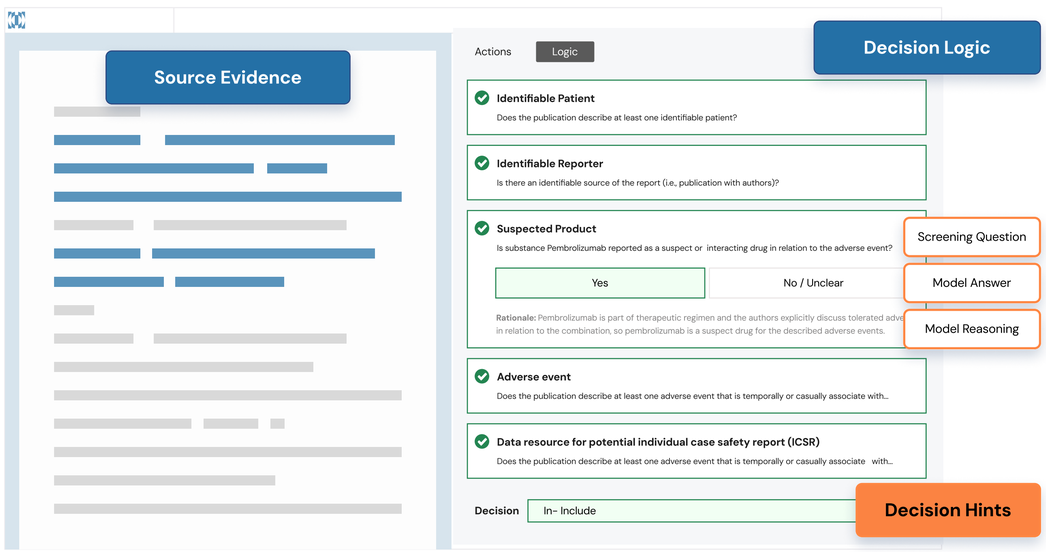
Overview of the evaluation dataset Table 1

Drug	Therapeutic area	Drug context	Publication time span	No. of PDFs evaluated
Metronidazole	Infectious diseases / gynecology	Older; combination regimens	1976–2026	99
Semaglutide	Diabetes / obesity	Innovative; combination regimens	2017–2026	100
Amiodarone	Cardiology	Older; primarily monotherapy	1984–2025	98
Isotretinoin	Dermatology	Older; primarily monotherapy	1993–2026	103
Pembrolizumab	Oncology / immunotherapy	Innovative; combination regimens	2016–2026	100
			Total	500

The decision logic executor assessed each full text using predefined screening questions. For each question, the system generated: an AI decision hint, a rationale, highlighted supporting evidence and a direct link to the source text.

Question-level answers were subsequently aggregated using the predefined final decision logic to produce an overall eligibility recommendation (decision hint). [Figure 1]

HITL full-text screening interface in Laser AI Figure 1

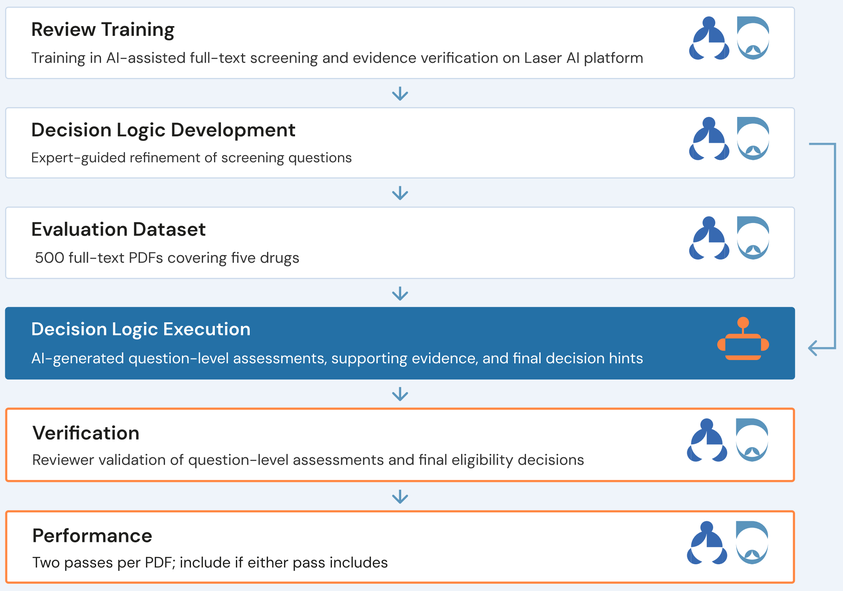


The reviewer evaluated AI outputs at two levels:

- Question-level: accepted or rejected each AI-generated screening answer.
- Final decision-level: reviewed the aggregated eligibility recommendation and made the final include/exclude decision.

Reviewer decisions served as the reference for the HITL evaluation. Outcomes included sensitivity, specificity, precision, and accuracy at the final decision-level, as well as AI–reviewer disagreement rates for individual screening questions. [Figure 2]

HITL full-text screening evaluation workflow Figure 2



Results

Final-decision agreement

Across 500 full-text publications, AI-generated eligibility recommendations agreed with final reviewer decisions in 492 cases (98.4%). Eight disagreements were identified, comprising two false-negative and six false-positive recommendations.

Performance was consistently high across the five drug-specific datasets, with sensitivity ranging from 95.24% to 100% and specificity from 97.26% to 98.73%. [Table 2]

Screening question-level agreement

Agreement was also assessed at the level of individual eligibility questions, enabling a more granular evaluation of AI-generated screening assessments.

AI–reviewer disagreement rates ranged from 0.0% to 2.0% across the five screening questions, indicating consistently high agreement at the individual-question level. [Table 2]

Decision-Level Performance and Screening Question-level Discrepancies Table 2

Drug	Decision-level									Screening Question-level				
	FP	FN	TP	TN	Sensitivity	Specificity	Precision	Accuracy	F1-score	Q1	Q2	Q3	Q4	Q5
Metronidazole	1	0	20	78	100.00%	98.73%	95.24%	98.99%	97.56%	0	0	0	0	1
Semaglutide	1	1	20	78	95.24%	98.73%	95.24%	98.00%	95.24%	2	0	6	4	2
Amiodarone	1	0	32	65	100.00%	98.48%	96.97%	98.98%	98.46%	2	0	0	1	1
Isotretinoin	2	0	30	71	100.00%	97.26%	93.75%	98.06%	96.77%	1	0	1	3	2
Pembrolizumab	1	1	58	40	98.31%	97.56%	98.31%	98.00%	98.31%	0	0	1	2	2
Total	6	2	160	332	98.77%	98.22%	96.39%	98.40%	97.56%	1.00%	0.00%	1.60%	2.00%	1.60%

Q1 - Identifiable patient, Q2 - Identifiable reporter, Q3 - Suspected product, Q4 -Adverse event, Q5 – ICSR data source.

Limitations

- The evaluation involved a single experienced reviewer.
- Reviewer decisions were made within the AI-assisted workflow, rather than through independent, blinded assessment.
- Review time and efficiency gains compared with conventional screening were not directly measured.

Conclusions

AI-assisted full-text screening demonstrated high agreement with reviewer decisions across five clinically diverse safety evidence datasets, with few disagreements at both final-decision and individual question-levels.

The structured human-in-the-loop approach combined AI-decision hints, supporting reasoning, and directly linked source evidence, enabling reviewer verification while maintaining human control over final decisions.

These findings support the feasibility of integrating AI assistance into full-text evidence review. Further evaluation using multiple reviewers and direct measures of review time is needed to establish its impact on screening efficiency and reliability.

References

1. Fleurence RL, Qureshi R, Aggarwal R, Bian J, Dawoud D, Delnoij D, Faden RR, Klijn S, Sonabend R, Chhatwal J; GenAI for Systematic Literature Reviews ISPOR Task Force. Generative Artificial Intelligence for Systematic Literature Reviews: A Good Practices Report of an ISPOR Special Task Force. Value Health. 2026 Sep 19:S1098-3015(26)02642-2. doi: 10.1016/j.jval.2026.08.007. Epub ahead of print. PMID: 42763092.
2. Nowak A., Sadowska E., Borowiack E., Opalek M. Evaluating fully automated Full-text screening for HTA systematic review. 2026-11, ISPOR Europe 2026, Vienna, Austria. Value in Health, Volume 29, Issue 12S

This research was co-funded by the European Union – European Regional Development Fund (Programme: European Funds for a Modern Economy 2021-2027, grant no. FENG.01.01-IP.02-4479/23).

European Funds · Republic of Poland · Co-funded by the European Union · Poland. Business Forward

Presented at ISPOR Europe 2026, Vienna.

Contact us at hello@laser.ai

More about this study