SA94

Are the machines ready to take over? Can artificial intelligence replace a human reviewer for literature screening and selection for systematic literature reviews?

Petra Nass[1], Dominika Rekowska[2], Emanuele Arcà[2], Artur Nowak[3], Ewelina Sadowska[3], Ewa Borowiack[3], Nick Halfpenny[2]

[1]OPEN Health HEOR & Market Access, Berlin, Germany; [2]OPEN Health HEOR & Market Access, Rotterdam, The Netherlands; [3]Evidence Prime, Kraków, Poland

OPEN HEALTH

INTRODUCTION

- Systematic literature reviews (SLRs) form the foundation of health technology assessment (HTA) submissions.
- SLRs are time-consuming and labor-intensive .
- The exponential increase in clinical research literature and aggressive timelines for HTA submissions, especially with the upcoming EU HTA process, are making SLRs even more resource-intensive and costly.
- Recent studies have shown that artificial intelligence (AI) could potentially accelerate SLR preparation including through replacing the second reviewer during title/abstract (TI/AB) screening.[1]
- Evidence Prime's Laser AI is a platform offering AI-supported SLR screening and data extraction, with human reviewers retaining full control of the SLR process.

OBJECTIVES

To assess accuracy and workload savings achieved with Evidence Prime's Laser AI in the context of a case study of a comprehensive SLR with eight updates.

METHODS

Systematic Literature Review

- The study assessed the utility of AI for replacing a human reviewer by comparing AI decisions with human-human decisions.
- A completed comprehensive clinical SLR, by human reviewers, involving eight updates, of biologic treatments for Crohn's disease (CD) was the basis for this comparison (**Figure 1**).
- For the human-human comparison, two human reviewers performed literature screening/selection, with conflicts resolved by a third human reviewer.
- The original search resulted in 7,272 records, with 176 records selected for full-text review; after full-text review 63 records remained.

Figure 1. PICOS

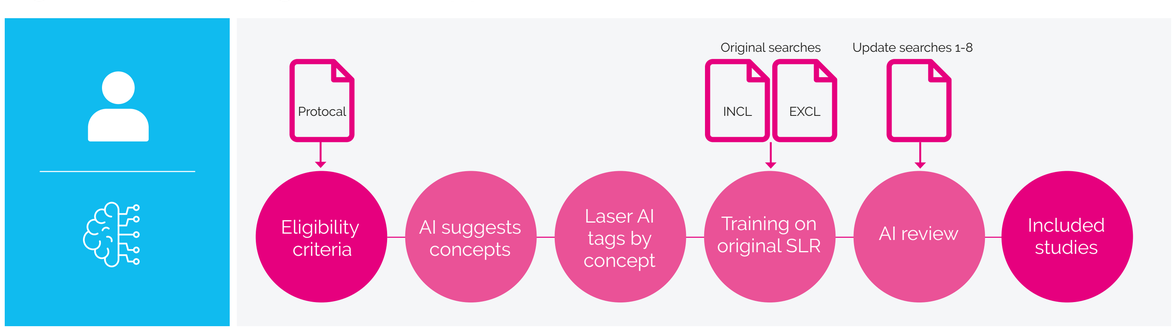
P	I	C	O	S
• Adult patients with moderately to severely active CD	• Biologics and biosimilars used in CD treatment	• Placebo or best supportive care • Any interventions of interest	• Clinical and endoscopic response and remission • CDAI score, surgery, relapse, IBD-Q, FACIT-Fatigue score, IBD-DI, PGIC, safety	• RCTs • English language

Abbreviations: CD: Crohn's disease; CDAI: Crohn's Disease Activity Index; FACIT: Functional Assessment of Chronic Illness Therapy; IBD-DI: Inflammatory Bowel Disease–Disability Index; IBD-Q: Inflammatory Bowel Disease Questionnaire; PGIC: Patient Global Impression of Change; PICOS: population, intervention, comparator, outcome, study design; RCT: randomized controlled trial.

Training the AI Platform

- Eligibility criteria were derived from the SLR protocol.
- Based on these, **Laser AI** suggested concepts for the PICOS domains.
- Then **Laser AI** identified concepts on the abstract level and developed Laser AI tags.
- The original SLR inclusions and exclusions were used to calibrate the concept-based classifier Laser AI tags.
- Subsequently, **Laser AI** replicated the screening for all eight updates, and the results were compared against the human-human literature screening/selection.
- For each of the eight update searches, a list of included studies was generated and compared with the human-generated list.

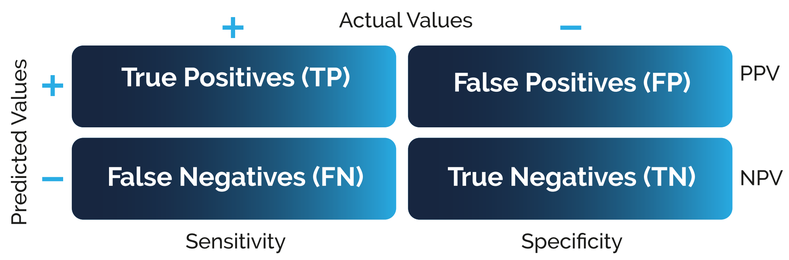
Figure 2. AI Training and AI-Supported SLR Process



Outcomes

- The main outcomes were sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV), as well as estimated workload savings. **Figure 3** provides the calculations for the accuracy outcomes, and definitions for the accuracy inputs are provided below. Workflow calculations are provided in the results section.
- **True Positives:** The number of studies correctly included as evidence by the AI and previously selected by human reviewers for final inclusion in the SLR.
- **False Positives:** The number of studies included after TI/AB screening that were not included as evidence in the final SLR.
- **False Negatives:** The number of True Positives that the AI excluded during TI/AB screening.
- **True Negatives:** The number or studies correctly excluded during TI/AB and FT review.

Figure 3. AI Training and AI-Supported SLR Process



Sensitivity = TP/(TP + FN)
Specificity = TN/(TN + FP)
PPV = TP/(TP + FP)
NPV = TN/(FN + TN)

RESULTS

Records Reviewed and Workload Savings

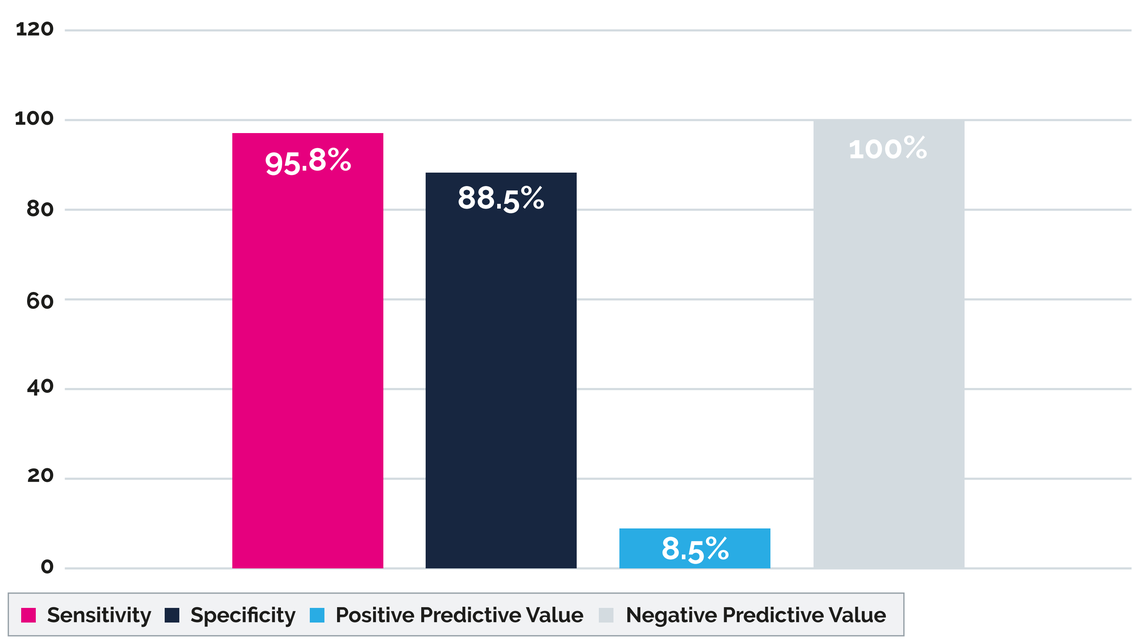
- The number of abstracts screened by both human reviewers and the AI ranged from 321 to 517 (3257 total) across the eight updates.
- The human reviewers included 7 to 35 (165 total) records for full-text review across updates, while the AI included 28 to 93 records (total 466).
- For workload calculations, we assumed that dual screening by two human reviewers would result in 6,514 (2 x 3,257) records for the TI/AB phase.
- Taking into consideration that in the AI-human scenario, 301 additional full-text articles would have to be reviewed by a human reviewer, use of **Laser AI** reduced the overall human screening effort from 6514 to 3558 (3257 + 301) records, resulting in average workload saving of 45.4% (**Figure 4**).

Figure 4. Work Effort and Savings Human-Human Versus AI-Hu

Accuracy

- Across all eight updates, sensitivity was 95.8%, specificity was 88.5%, PPV was 8.5%, and NPV was 100% (**Figure 5**).
- In seven of eight updates, the AI identified all studies that had been included by the human reviewers in the final SLR, which corresponds to 100% sensitivity in these updates.
- In update 6, one of three studies included by the human reviewers after full-text review was missed by the AI during TI/AB screening, resulting in 66.7% sensitivity for this update.
- However, the study that was missed by the AI in update 6 was correctly identified and included in update 7 by the AI.

Figure 5. Accuracy of AI for Literature Screening



What influences accuracy?

Training set

- Increasing the size of a training set generally improves sensitivity and specificity.

Inclusion threshold

- Lowering the inclusion threshold generally improves sensitivity but lowers specificity.

Study design/disease area

- AI tools tend to be more accurate when faced with study designs with consistent terminology such as randomized controlled trials.

CONCLUSIONS

- The results of this case study support the use of AI as a second reviewer for TI/AB screening during update searches.
- AI tools can reduce the time and effort required for the screening phase of SLRs.
- While achieving high sensitivity (capturing all relevant studies) comes at the cost of more inclusions at the full text screening stage, there would still be considerable workload savings when an AI replaces a human screener.
- Note that additional workload savings and accuracy might have been achieved if we had retrained the AI after each update.

REFERENCES

1. Rekowska D, Halfpenny N, Gulser S, Santpurkar N, Fox G, Nass P. Artificial Intelligence for Literature Screening and Selection: Does the Evidence Support Its Use in Systematic Literature Reviews? Oral presentation, HTAi 2024 Annual Meeting; June 15-19, 2024.

DISCLOSURES

Evidence Prime provided the Laser AI platform at no cost to OPEN Health and performed the AI training and screening of updates.

LASERAI

Presented at: ISPOR Europe 2024