View all Parexel's posters at ISPOR 2025

Exploring Artificial Intelligence's (AI's) Role in Literature Screening: A Comparative Analysis Paving the Way for More Efficient Evidence Synthesis

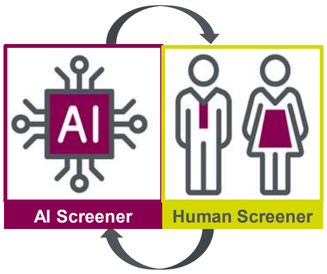
#MSR102

S. Budhia[1], G.S. Mangat[2], A. Jain[2], and S. Sharma[3]
Parexel International, 1London, United Kingdom; 2Mohali, India; 3Chandigarh, India

Background

- Despite a decade of Artificial Intelligence (AI) applications in citation screening and recent advances in natural language processing, quantitative comparative evidence on AI reliability across diverse review contexts remains critically insufficient.
- Literature screening for evidence synthesis traditionally demands significant human resources, creating bottlenecks when processing thousands of citations. While AI shows promise in streamlining this process, its performance may vary depending on factors such as review complexity, study quality, methodological diversity, and disease context.
  - The optimal solution likely lies in underexplored hybrid models where human expertise and AI capabilities could work together complementarily (Figure 1). These collaborative systems might combine human contextual understanding with AI's consistency and pattern recognition.

Figure 1: The Hybrid Screening Model: Balancing Human Judgment and AI Efficiency



- Health Technology Assessment (HTA) bodies have also begun providing guidance on AI-assisted evidence synthesis, but these recommendations remain limited. This gap necessitates additional comparative effectiveness studies to evaluate key performance metrics across diverse review contexts. Such empirical evidence is crucial for developing robust methodological frameworks for both HTA bodies and sponsor companies.

Objectives

- To comprehensively evaluate the effectiveness of human-only screening versus AI-assisted screening approaches in literature review processes by:
  - Comparing and quantifying key performance metrics (accuracy, recall, precision, and F1 scores) across three distinct oncology review types: treatment patterns, clinical, and epidemiology.
  - Assessing the reliability and efficiency of AI-assisted screening when trained on a limited subset (30%) of human-reviewed citations.
  - Identifying specific review contexts where AI demonstrates optimal performance and those requiring enhanced human oversight.
  - Determining the practical implications of AI implementation for resource allocation and timeline management in systematic review processes.

Methodology

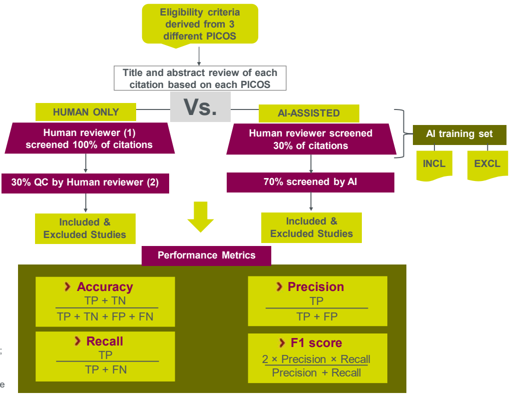
- This study assessed the utility of AI assisting a human reviewer by comparing Human-AI decisions with human-human decisions.
- Three comprehensive literature reviews, completely conducted by humans alone, were the basis for this comparison. Multiple disease areas with varying levels of complexity were deliberately selected to evaluate the AI's adaptability across different clinical contexts and to test its potential effectiveness under diverse screening conditions (Figure 2).

Figure 2: PICOS

Review	P	I	C	O	S
Treatment patterns	Adult patients with locally advanced gastric cancer	Systemic treatments	No restriction	Distribution, adherence, persistence	Observational
Epidemiology		Not applicable	Not applicable	Incidence, prevalence, mortality	
Clinical	Adult patients with pheochromocytoma and paraganglioma	Systemic treatments, radiotherapy	No restriction	Efficacy and safety (survival, response rates, adverse events)	RCTs, nRCTs, single-arm trials

Abbreviations - PICOS: Patient population, intervention, comparator, outcomes, study design; RCT: Randomized controlled trial; nRCT: Non-randomized controlled trial

Figure 3: Literature review process (Human vs. Human-AI)


Eligibility criteria derived from 3 different PICOS
Title and abstract review of each citation based on each PICOS
Vs.
HUMAN ONLY
Human reviewer (1) screened 100% of citations
30% QC by Human reviewer (2)
Included & Excluded Studies
AI-ASSISTED
Human reviewer screened 30% of citations
AI training set
INCL
EXCL
70% screened by AI
Included & Excluded Studies
Performance Metrics
Accuracy
TP + TN / TP + TN + FP + FN
Precision
TP / TP + FP
Recall
TP / TP + FN
F1 score
2 × Precision × Recall / Precision + Recall

Abbreviations - EXCL: Excluded; INCL: Included; FN: False negative; FP: False positive; TN: True negative; TP: True positive

- We compared two screening approaches across three distinct oncology reviews. The conventional human-only method utilized a senior reviewer for complete title/abstract screening, with 30% undergoing quality verification. For the AI-assisted approach, 30% of citations were human-reviewed to train the AI system, which then independently screened the remaining 70%. We comprehensively evaluated performance using accuracy, recall, precision, and F1 scores to assess the comparative effectiveness of both approaches (Figure 3).

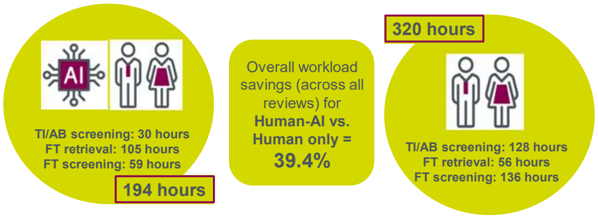
Methodology….

- **TP (True Positives)**: Studies correctly identified as relevant (by AI-assisted screening) that were also included by human reviewers.
- **TN (True Negatives)**: Studies correctly identified as irrelevant (by AI-assisted screening) that were excluded by human reviewers.
- **FP (False Positives)**: Studies incorrectly identified as relevant (recommended for inclusion by AI-assisted screening) when they should have been excluded according to the eligibility criteria.
- **FN (False Negatives)**: Studies incorrectly identified as irrelevant (recommended for exclusion by AI-assisted screening) when they should have been included according to the eligibility criteria; these represent missed relevant studies.

Results

- The investigation evaluated three distinct types of literature reviews: treatment patterns (n=1,959 citations), clinical (n=1,331 citations), and epidemiology (n=4,124 citations). During title and abstract screening, human reviewers alone identified 213 relevant citations in the treatment patterns review, 274 in the clinical review, and 558 in the epidemiology review, establishing the reference standard for comparison with the AI-assisted approach.
  - The AI-assisted approach flagged 582 citations in the treatment patterns review, 553 in the clinical review, and 840 in the epidemiology review as potentially relevant for further consideration.
- The AI-assisted approach reduced the total workload by ~39% (126 hours) across the three reviews. While requiring 88% more time during full-text retrieval (due to low precision), the AI-assisted approach delivered substantial efficiencies in title/abstract screening (77% reduction) and full-text screening (57% reduction), demonstrating that AI-assistance creates significant overall time savings despite varying phase-specific impacts (Figure 3).

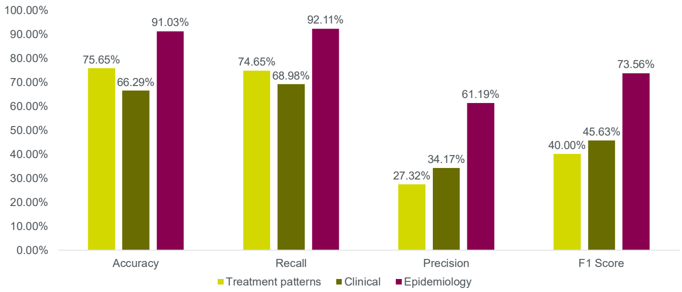
Figure 3: Work effort and savings (AI-assistance vs. Human only)


AI
TI/AB screening: 30 hours
FT retrieval: 105 hours
FT screening: 59 hours
194 hours
Overall workload savings (across all reviews) for Human-AI vs. Human only = 39.4%
320 hours
TI/AB screening: 128 hours
FT retrieval: 56 hours
FT screening: 136 hours

AB: Abstracts; FT: Full-texts; TI: Titles

- In contrast to purely human output, the AI-assisted methodology showed considerable accuracy, achieving scores of 75.65% for treatment patterns, 66.29% for clinical reviews, and 91.03% for epidemiological studies (Figure 4).
- The AI-assisted approach also demonstrated strong sensitivity, with recall rates reaching 74.65%, 68.98%, and 92.11% respectively, across these review categories, indicating the AI's capability to identify at least two-thirds of relevant citations across different review scenarios. Nevertheless, precision remained consistently suboptimal across all reviews [27.32% (treatment patterns), 34.17% (clinical), and 61.19% (epidemiology)], indicating the AI's tendency to include irrelevant citations (Figure 4).

Figure 4: Performance metrics of AI-assisted human decisions (vs. human only) for literature screening (titles/abstracts)



- The resulting low F1 scores of 40.00% (treatment patterns) and 45.63% (clinical) highlighted the disparity between precision and recall, with only the epidemiology review (73.56%) showing relatively balanced performance metrics (Figure 4).

Conclusion

- The AI-assisted screening methodology exhibits strong accuracy and recall capabilities, revealing significant potential despite its tendency toward over-inclusion. Although precision remains an area for enhancement, the AI's primary strength is its thorough identification of relevant research. Considering the swift evolution of AI technologies, future versions will likely deliver improved performance metrics.
- We highly recommend the implementation of AI for epidemiology and treatment pattern analyses.
- For clinical reviews that quantify treatment benefits in comparative effectiveness assessments critical to reimbursement decisions, AI requires significant improvements and substantial human oversight due to current precision limitations.
- HTA bodies should establish formal guidance for evaluating AI-assisted evidence synthesis, while sponsor companies should adopt hybrid workflows that leverage AI strengths without compromising decision quality.

www.parexel.com