

An Artificial Intelligence (AI)-assisted Systematic Literature Review (SLR) of the Economic Burden in Metastatic Pancreatic Adenocarcinoma: A Proof-of-Concept Study

PT3

Carolina Casañas i Comabella¹, Mansee Jajoo², Jing Wang-Silvanto², He Guo², Allie Cichewicz³, Rishi Ohri²

¹Thermo Fisher Scientific, London, UK, ²Astellas Pharma Global Development, Inc., ³Thermo Fisher Scientific, Waltham, MA, USA

Background

- Systematic literature reviews (SLRs) are crucial in health economics and outcomes research (HEOR) and are the preferred methodology by health technology assessment (HTA) bodies worldwide to identify and synthesize data.
- Traditional methods¹ used in SLRs are time-consuming, labor-intensive, and costly. Advances in artificial intelligence (AI) have been increasingly explored in HEOR to accelerate SLRs, particularly for screening and data extraction tasks. The AI position statement by the UK's National Institute for Health and Care Excellence expressed caution regarding the use of AI and emphasized the need for transparency, rigor, and human supervision.² No other HTA bodies have issued detailed guidance on AI use in SLRs.
- Given the rapid evolution of AI technologies and the lack of HTA guidance, it is necessary to assess the performance and potential time and cost efficiencies of using AI in SLRs.

Objectives

- To explore the performance and potential efficiencies of leveraging AI to assist with the screening and data extraction tasks in SLRs.
- To inform future use cases for AI in SLRs in terms of performance and potential time and cost savings.


Methods

- A traditional SLR was replicated by deploying Nested Knowledge's (NK) AI platform for screening and data extraction. Both SLRs aimed to synthesize evidence on the economic burden (healthcare resource utilization and costs) of metastatic pancreatic adenocarcinoma.
- Titles and abstracts were screened by a human and an AI machine-learning (ML) model. The ML model was trained using bibliographic data, abstract content, and citation counts from an initial set of 50 publications screened by a human, and was retrained automatically after every 10 additional records were screened.
- Each full-text paper included during title/abstract screening was screened by a human and a large-language model (LLM) that used prompts based on each Population, Intervention, Comparator, Outcome, Study design (PICOS) eligibility criterion. Records were excluded if the LLM did not identify information for all PICOS criteria.
- For data extraction, the prompts used for full-text screening were expanded to include more detail. Data extracted by AI were categorized as correct, incomplete, missing, incorrect, or requiring human checks or interpretation. Means and minimum/maximum values were calculated overall and by type of variable.

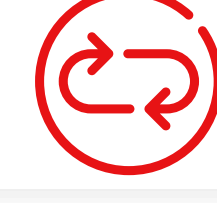
Results

- The Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) diagrams in **Figure 1** show the differences between the decisions made by AI in the AI-assisted SLR versus those made by humans in the traditional SLR.
- After screening about 25% of titles/abstracts (n=326), the ML model achieved 87% accuracy, 82% recall, and 40% precision and was deemed appropriately trained (**Figure 2**). Low precision means that AI used a safer approach, i.e., included more records than humans (251 vs. 85).

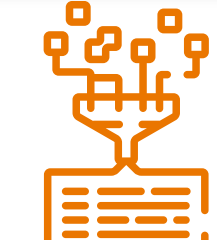
Accuracy

Indicates how often the model is correct in classifying abstracts as included or excluded records. The higher the accuracy the better.

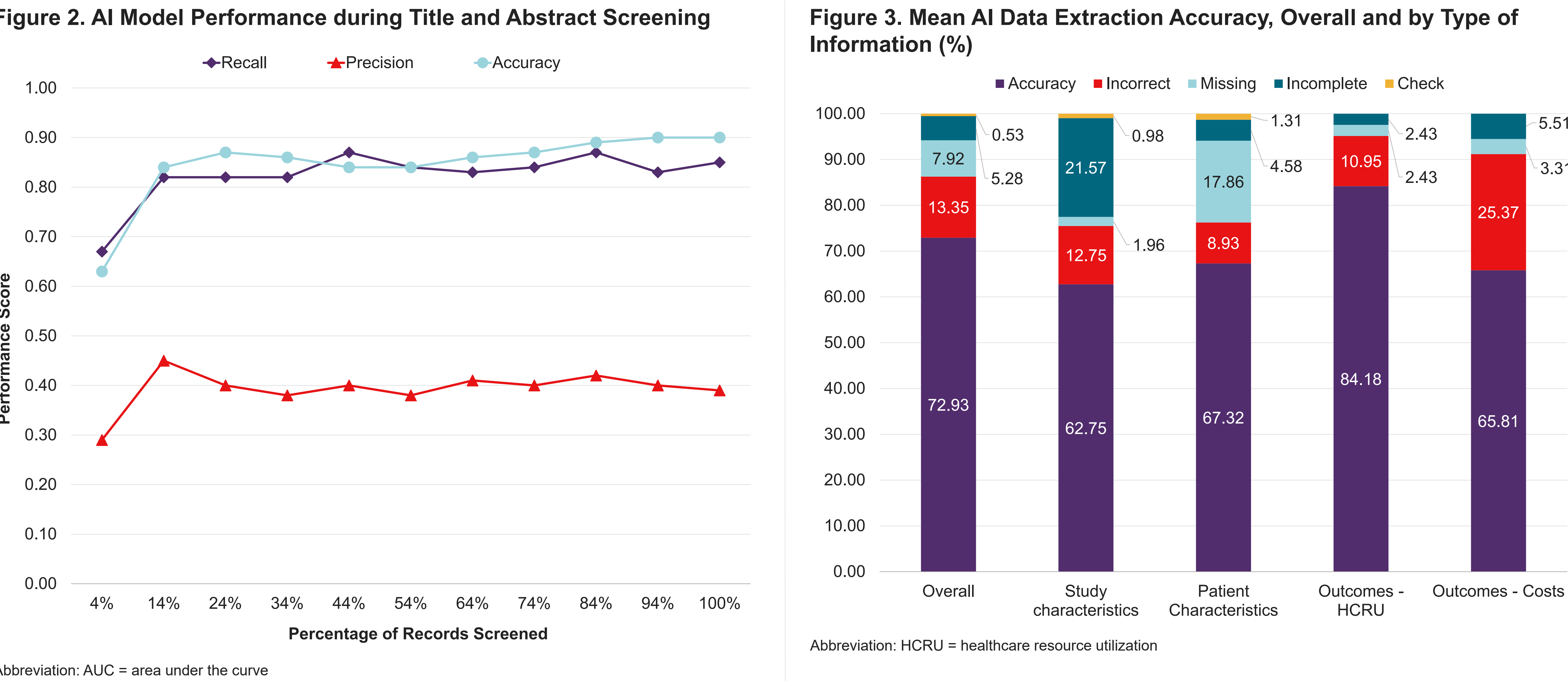
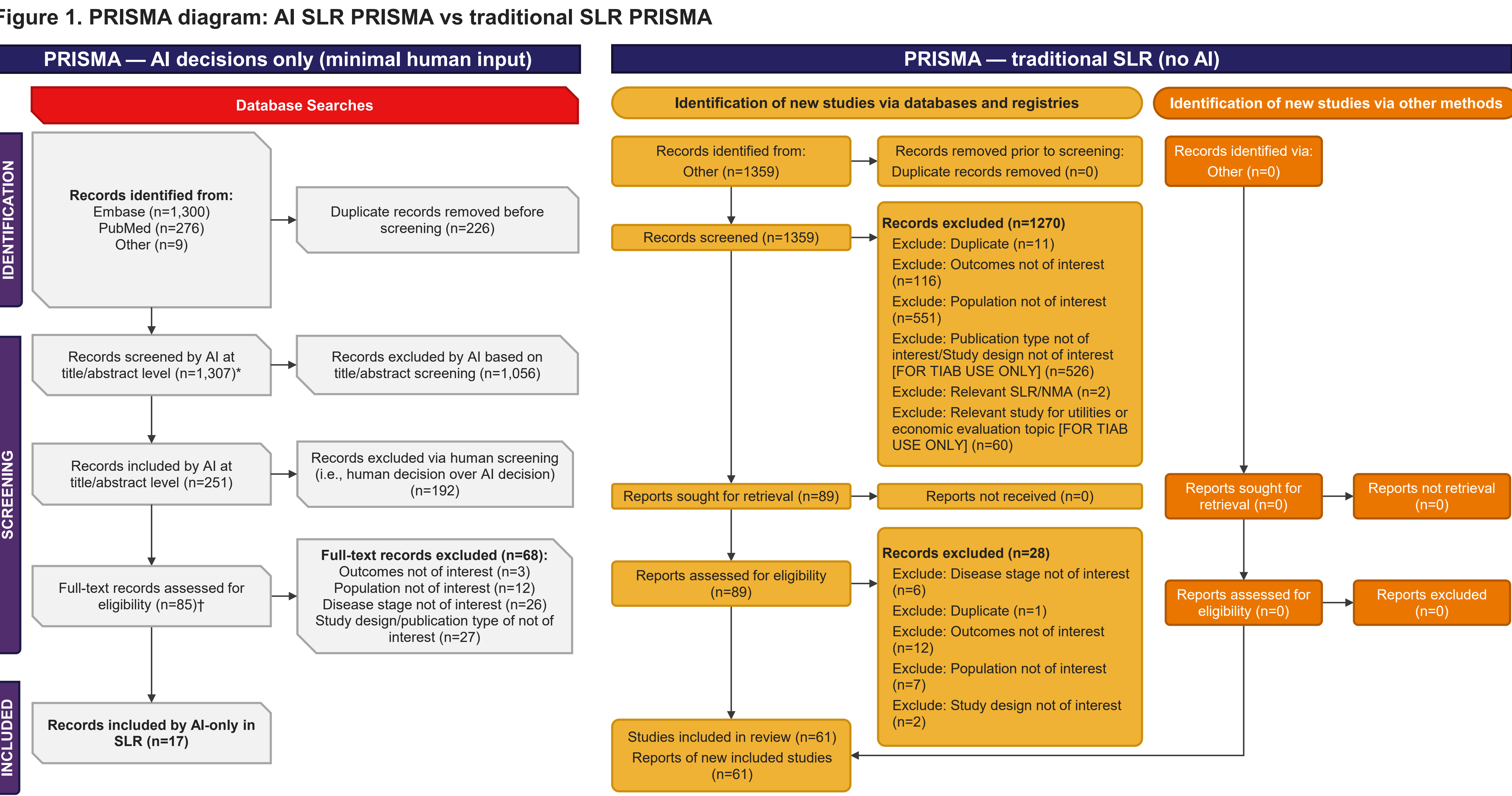
Recall

Reflects the tool's ability to identify relevant studies. The AI model aims to achieve high recall (>70%), which indicates that the model is less likely to exclude relevant records.

Precision

Measures the proportion of studies flagged by the tool that are relevant, i.e., the proportion of relevant abstracts retrieved among all abstracts retrieved. Low precision (<50%) indicates the model is more likely to include irrelevant records for full-text review compared with human reviewers.

Results (cont.)



- During full-text screening, the LLM responses based on PICOS-based prompts included 27.7% (17/61) of records compared with the traditional SLR. This was primarily driven by the LLM not being able to identify information related to publication/study type (81.4%, 22/27 in disagreement) and disease stages (69.2%, 18/27 in disagreement), resulting in AI excluding relevant records.
- For LLM-extracted data, mean accuracy was 72.93%. Mean incorrect, missing, and incomplete extractions were 13.35%, 7.92%, and 5.28% respectively (**Figure 3**).

Table 1. Time and Cost Implications (Base Case and Range)

Task	Hours Required by Task		Time Difference (Hours)	Saving %	Assumptions
	Traditional SLR (2 human reviewers)	AI-assisted SLR (1 human reviewer & 1 AI reviewer)			
Title/abstract screening	52	26	26	50%	<ul style="list-style-type: none">Traditional SLR: two humans screened 1,307 records each (double screening).AI SLR screened 1,257 records (i.e., full set minus 50 abstracts screened by humans to train the AI model).Time required for prompt engineering is not included.
Full-text screening	63	31	31	50%	<ul style="list-style-type: none">Traditional SLR: two humans screened 251 full-text records each (double screening).AI SLR screened 251 records.Time required for prompt engineering is not included.
Data extraction	29	12	17	59%	<ul style="list-style-type: none">Traditional SLR: one human extracted 17 records, one human validated each data entry.AI SLR: AI extracted 17 records, one human validated each data entry.Time required for prompt engineering and formatting Microsoft Excel® output is not included.
Total	144	69	75	52%	
Range	102–219	69–100	33–119	32%–61%	<ul style="list-style-type: none">Lower range is based on AI having identified more records to screen at full text than humans (85 vs. 251); therefore, based on the AI alone, there would be an increased volume of full texts to screen.Higher range is based on AI having included fewer publications in the SLR than the humans (17 vs. 61); therefore, based on AI alone, there would be higher savings because the AI included fewer records—this would incur data loss of >70%, because AI excluded 44 relevant studies.

Abbreviations: AI = artificial intelligence; SLR = systematic literature review

Discussions

- NK's ML model performed well at title and abstract screening, but the low precision and high recall resulted in a larger volume of records to screen at full text compared with human screening.
- The mean accuracy achieved during data extraction (72.9%) may be acceptable for rapid, targeted reviews but may be inadequate for SLRs intended for submission to HTA bodies.
- Unlike other LLMs, NK did not hallucinate content (i.e., generate data that does not exist), because it is trained to only pull direct text excerpts. This minimized risk of error and time wasted by human validators.

Limitations

- The SLR evaluated costs and healthcare resource utilization from observational studies. Findings may not be generalized to other topics, study designs, or outcomes. Testing AI use in larger SLRs (i.e., larger volume of publications to screen and extract) might yield higher savings.
- This proof-of-concept study assessed the NK platform. It should be noted that NK's LLM is not designed to screen full texts; therefore, an adapted method was used for this task.
- Other AI tools may use different approaches and produce different results. Hence, our findings may not be generalizable to other AI models.

Conclusions

- Human input is necessary when deploying AI to screen studies and extract data to ensure accuracy. AI-literate reviewers are required to ensure efficient prompt engineering.
- AI support may be suitable for rapid, targeted reviews, where methods are less stringent, or scoping exercises, where a certain degree of error may be acceptable. However, the use of AI in SLRs to support HTA submissions should be cautious, follow guidelines for transparency and replicability, and careful performance monitoring by humans.
- Given the rapid development of AI models, it is reasonable to expect that future improvements in accuracy will provide better to support literature reviews. Guidance by HTA bodies is needed to establish the acceptability and use of AI in SLRs.

References

1. Higgins JPT, Thomas J, Chandler J, et al. Cochrane handbook for systematic reviews of interventions version 6.4. Cochrane. Updated August 2023. Accessed October 25, 2023. <https://training.cochrane.org/handbook>

2. National Institute for Health and Care Excellence. Use of AI in evidence generation: NICE position statement. Accessed March, 2025.