MSR18

Validation of an Agentic Large Language Model (LLM) System in the Review Stage of a Real-Time AI-assisted Living Systematic Literature Review (REAL-SLR): a Solution to Instant and Easy access to Clinical Trial Data (CTD)

Rozee Liu[1], Rhiannon Campden[1], Eddie Xiaole Liu[2], Oscar Correa[3], Anna Forsythe[1]

[1]Oncoscope-AI, Miami, FL, USA; [2]Independent, Toronto, ON, Canada; [3]Eviviz Inc., Vancouver, BC, Canada

BACKGROUND

- Health economics and outcomes research (HEOR) professionals often struggle to stay updated on the latest published clinical trial data
- Traditionally, building de novo SLRs requires extensive time and manual effort

OBJECTIVES

- → To assemble a REAL-SLR of clinical trial data using an agentic LLM system to generate review annotations, following Cochrane's initiative of Adherence to the Responsible AI use in evidence SynthEsis (RAISE)[1]
- → To evaluate the system's accuracy and associated time savings

METHODS

- An agentic AI system was developed using two OpenAI LLMs (GPT-5, GPT4.1), Gemini 2.5Pro, and Claude Sonnet 4.5 with ethical, transparent, and fit-for-purpose AI development
- It was designed to emulate expert-led Cochrane-compliant[2] SLR processes (transparency, reproducibility, and research integrity), subdividing complex processes into smaller subtasks, and documenting its reasoning for traceable results
- Accuracy of review was evaluated on publications in five cancers: non-small cell lung cancer (NSCLC), prostate cancer (PC), breast cancer (BC), bladder cancer (BldC) and multiple myeloma (MM) compared to human results
- The system follows an annotation manual, constructed by human scientists manually annotating 61,069 publications (17,085 NSCLC, 15,114 PC, 21,904 BC, 9,719 BldC, 6,966 multiple myeloma) (**Figure 1**)
- Each study was annotated with 4 review variables (population, intervention/comparator, reported outcomes, study design) to facilitate clear audit trails of decisions and to determine the inclusion into the database that is the basis of the living DST. Annotations for each review variable was entered independently, while the overall include/exclusion decision was based on one or more review variables being marked as an exclude
- Overall accuracy, sensitivity (ability of the model to accurately predict excluded abstracts), and specificity (ability of the model to accurately predict included abstracts) were calculated, fitting the Cochrane Evaluation of (Semi-) Automated Review Methods (CESAR) study[3] for SLR AI tool selection criteria (**Table 1**)

Figure 1. Distribution of training data across five cancers

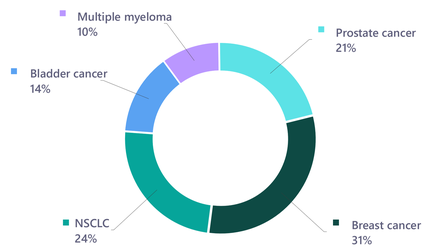


Table 1. Cochrane Evaluation of (Semi-) Automated Review Methods (CESAR) study[3] for SLR AI tool selection criteria

1. Tool usability and alignment with Cochrane expectations
2. Alignment with RAISE (Responsible AI use in evidence SynthEsis)
3. Evaluation standards and validation approaches
4. Capacity building and infrastructure

Performance Metrics	CESAR study Futility Boundaries (Point Estimate)
Sensitivity	80%
Specificity	50%

Figure 2. Accuracy, sensitivity, specificity results across PICOS variables by five cancers

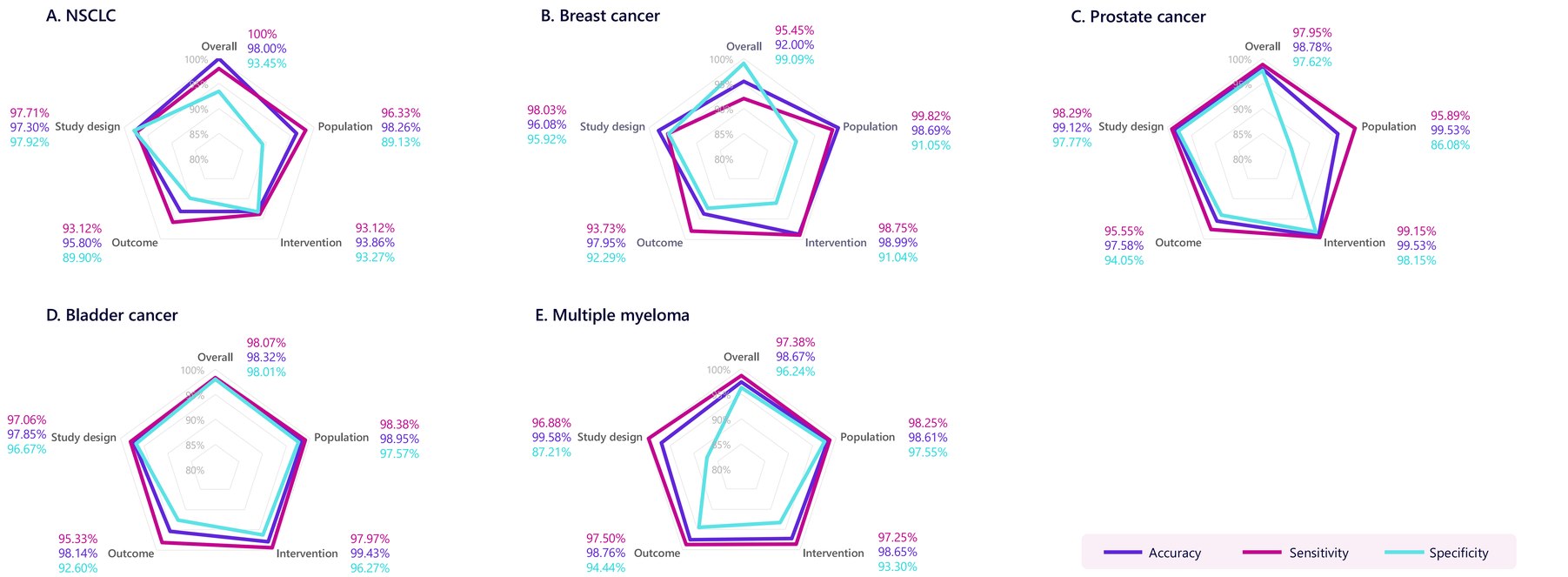
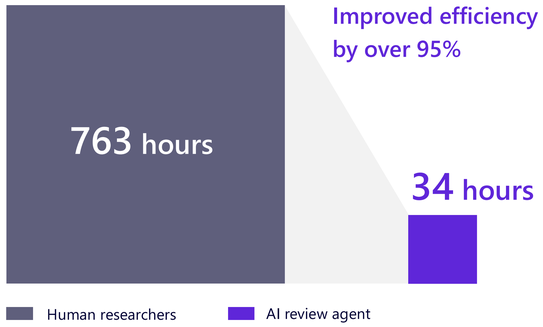


Table 2. Variation in false negative rates across five cancers by PICOS variables

PIC/OS variables	NSCLC	Breast	Prostate	Bladder	MM
Population	1.38%	1.00%	0.34%	0.61%	0.38%
Intervention	3.21%	0.33%	0.24%	0.41%	1.00%
Outcome	2.29%	3.67%	1.03%	0.91%	0.88%
Study design	0.92%	0.00%	1.03%	0.71%	0.75%
Overall Accept/Reject	0.36%	0.33%	0.24%	0.30%	0.63%

Figure 3. Time savings: agentic AI systems vs. human researchers in the SLR review stage



RESULTS

- Accuracy ranged from 93.73% to 99.82% (**Figure 2**)
- The sensitivity and specificity ranged from 93.86% to 99.58%, and 86.08% to 98.15%, respectively, far exceeding CESAR study futility boundaries (**Figure 2**)
- The false negative rates were 0.34%, 0.33%, 0.88%, 0.00% for the 4 PICOS variables, with a 0.30% cumulative rate (**Table 2**)
- Our system completed review in 33.93 hours, compared to an estimated 763.36 hours by trained human researchers, resulting in 95.56% time savings (**Figure 3**)

CONCLUSIONS

- → Our agentic LLM system can accurately review publications with performance superior to human experts
- → This level of accuracy highlights our system's potential to deliver real-time clinical data, empowering HEOR professionals with expedited evidence generation, with the hopes of ultimately improving patient access

REFERENCES

1. Thomas J, Hair K, Noel-Storr, A. et al. Responsible use of AI in evidence SynthEsis (RAISE): recommendations for practice (version 3; updated 13 March 2026). In: Open Science Framework [https://osf.io/], Washington DC: Center for Open Science. DOI 10.17605/OSF.IO/FWAUD (accessed April 24 2026 via https://osf.io/cqa82)
2. Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (editors). Cochrane Handbook for Systematic Reviews of Interventions version 6.5 (updated August 2024). Cochrane, 2024. Available from www.cochrane.org/handbook
3. Gartlehner G, Banda S, Callaghan M, et al. Cochrane Evaluation of (Semi-) Automated Review Methods (CESAR): protocol for an adaptive platform study within review. *medRxiv*. Published April 13, 2026. doi:10.64898/2026.04.13.26350802

ONCOSCOPE AI
REAL-TIME ONCOLOGY LIBRARY

Presented at ISPOR 2026 | May 17-20, 2026 | Philadelphia, PA, USA | Poster Session 1

See more Oncoscope AI publications