EPH28

Revolutionizing Systematic Reviews: The Precision of LLMs in Screening Observational Studies

Julia Langham,[1] Tim Reason,[1] Andy Gimblett,[1] Bill Malcolm,[2] Nathan Hill[3]

[1]Estima Scientific, London, United Kingdom; [2] Bristol Myers-Squibb, Uxbridge, UK; [3] Bristol Myers-Squibb, USA

Introduction

- Performing a high-quality Systematic Literature Review (SLR) can be costly and time-consuming, requiring a large number of skilled person-hours. Consequently, using Artificial Intelligence (AI) tools to assist with study selection (e.g. screening of titles and abstracts, and full-text review with data extraction), has seen rapid growth.[1]
- We recently reported highly accurate results of using GPT-4 for screening and risk of bias assessment in clinical trials. Compared to the final set of eligible studies, the sensitivity and specificity were 95.9% (95% confidence interval [CI]: 88.5 to 99.1) and 86.7% (95% CI: 85.0 to 88.2), respectively.[2,3]
- Some of the success of accurate screening of clinical trials may rely on the adherence to Consolidated Standards of Reporting Trials (CONSORT).[4] SLRs including observational studies such as Non-randomized Studies of Interventions, are less straightforward.
- Non-randomized, observational studies are more difficult to categorise. This may be due to lack of adherence to reporting standards or reporting insufficient information to allow them to be classified (or for inclusion-exclusion criteria to be applied) particularly at the abstract screening stage. Therefore, to identify studies, less specific searches and screening terms are required to ensure unbiased search strategies. This leads to a far higher volume of studies to search and screen.

Aim

The aim of this study was to compare an LLM (GPT-4) to human reviewers for the identification of non-randomized, observational studies based on title and abstract screening and full-text review in two case studies.

Methods

GPT-4

- A Python application programming interface (API) was used to send "prompts" and text (titles and abstracts, text extracted from PDFs of full publications) to GPT-4 (model: gpt-4-0613) to summarise text against pre-specified criteria, and to GPT-4o-mini (model: gpt-4o-mini-2024-07-18) to assess eligibility based on study design, population, treatment and outcomes.
- This poster refers to GPT-4 in the tables to although a combination of models was used.

Primary screening: Titles and abstract

- GPT-4 was instructed to assess eligibility of studies included in a list of title and abstracts and for full text publications
- Lists of citations were prepared from two previously conducted SLRs of observational studies which used gold standard methods of study selection (i.e. two human reviewers screened titles and abstracts, and subsequent full-text publications), (described below).
- The citation lists included 2% (of n=993) and 4% (of n= 756) eligible studies (i.e. citations relevant to the research question) for GPT-4 to correctly identify.

Secondary screening: Full-text screening

- Full-text screening was performed on a subset of the publications that were both screened in the original review (for which there was a documented reason for exclusion) and were open access (freely available). The full-text dataset included 13% (out of a total n=162) and 22% (n=61) eligible studies for GPT-4 to correctly identify.
- *"Open access articles were used for this research*"

Data Analysis

- The sensitivity, specificity, accuracy, and precision of GPT-4 screening decisions were calculated for both the title and abstract and full-text screening. This was achieved by comparing GPT-4 decisions with the final screening decisions made in each case study (i.e. final decisions, defined as the number of citations deemed relevant by the experienced reviewers after both screening and full-text review).
- Title and abstracts were included if GPT-4 classified as correct; (1) study design, (2) population, and (3) treatments. Full publications were included if GPT-4 classified as correct; (1) study design, (2) population, (3)treatments, and (4) outcomes.
- **Table 1 presents the sensitivity and specificity results of title/abstract and full-text screening.** The definitions of the metrics used in this analysis are shown in Table 1

Results

Screening	TP n (%)	TN n (%)	FP n (%)	FN n (%)	Sensitivity % (95% CI)	Specificity % (95% CI)	PPV (%)	NPV (%)	Accuracy (%)
Title and abstract									
Case study 1 (IV vs SC)	13 (92.9%)	643 (86.7%)	149 (15.7%)	1 (7.1%)	92.9 (79.4 to 106.4)	86.7 (84.2 to 89.1)	99.8	11.6	86.8
Case study 2 (CLL)	42 (97.7%)	759 (79.9%)	191 (20.1%)	1 (2.3%)	97.67 (93.2 to 100)	79.9 (77.4 to 82.4)	99.9	18.0	80.7
Full publication									
Case study 1 (IV vs SC)	8 (100%)	44 (83.0%)	9 (17.0%)	0 (0.0%)	100	83.0 (72.9 to 93.1)	100	47.1	85.2
Case study 2 (CLL)	34 (97.1%)	95 (74.8%)	32 (25.2%)	1 (2.9%)	97.14 (91.62 to 100)	74.80 (67.3 to 82.4)	99.0	51.5	79.6

Metrics (Outcome definitions)

- True positives (TP): number of references GPT-4 correctly identified as eligible for inclusion.
- True Negative (TN): number of references GPT-4 correctly excluded and also excluded by humans.
- False positive(FP): number of references GPT-4 incorrectly identified as eligible that were excluded by humans.
- False negative (FN): number of references that GPT-4 incorrectly excluded that were included by human screeners.
- Sensitivity: proportion of true positive identifications of all references that should have been included. [TP percentage = TP / (TP + FN).]
- Specificity: proportion of true negative identifications of all references that should have been excluded. [TN, percentage = TN / (FP + TN).]
- Precision, PPV: (positive predictive value) the proportion of relevant references correctly identified by GPT-4, TP / (TP + FP)
- NPV: Negative Predictive Value= the proportion of irrelevant references correctly identified by GPT-4 i.e TN / (TN + FN).
- Accuracy: the proportion of references GPT-4 correctly classified as either relevant or irrelevant i.e (TP + TN) / (TP + TN + FP + FN)

Table 1: Sensitivity and specificity metrics for the primary and secondary screening

- The approximate time required for GPT-4 to screen 500 titles and abstracts was 1 hour
- Full-text screening and data extraction for each batch of 50 PDFs took the LLM (GPT-4) 3 minutes
- **Tables 1** shows the sensitivity and specificity of GPT-4 correctly identifying studies. A summary of the results are:
  - **Title and abstract review (**compared to the final set included in the study):
    - Case study 1 (IV vs SC): 92.9% and 86.7% respectively, (n= 756)
    - Case study 2 (CLL): 95.5% and 79.9% respectively, (n=993)
  - **Full publication** review (compared to the final set included in the study)
    - Case study 1 (IV vs SC): 100%, and 83.0% respectively, (n =61)
    - Case study 2 (CLL): 94.4% and 74.6%, (n=162)
- **Table 2** presents the number of studies included and excluded, based on title/abstract and full-text screening, conducted by the LLM against those included or excluded by the human reviewer for each case study

Primary screening			**Human researchers**		
Case study 1 (IV vs SC)			Exclude	Include	Total
	GPT-4	Exclude	643	1	644
		Include	99	13	112
		Total	742	14	756
Case study 2 (CLL)			Exclude	Include	Total
	GPT-4	Exclude	734	1	735
		Include	215	43	258
		Total	949	44	993
Secondary screening			**Human researchers**		
Case study 1 (IV vs SC)			Exclude	Include	Total
	GPT-4	Exclude	44	0	44
		Include	9	8	17
		Total	53	8	61
Case study 2 (CLL)			Exclude	Include	Total
	GPT-4	Exclude	95	1	96
		Include	32	34	66
		Total	127	35	162

Table 2: GPT-4 compared to the original final eligibility decision for the primary screening and secondary screening

Conclusion

- This study reports the results of two case studies measuring the accuracy of title and abstract and full text screening and selection with GPT-4 against the gold standard (the conventional double screening by humans) method for SLRs of observational studies.
- GPT-4 quickly and accurately summarised relevant study characteristics from the title and abstract and full text review to correctly determine eligibility against pre-specified inclusion and exclusion criteria in two diverse SLRs of observational studies.
- Screening was accomplished in a fraction of the time it takes humans without compromising the quality of the SLR.
- Detailed prompts were required to ensure GPT-4 was able to undertake this task. Further prompt refinement and fine-tuning with GPT-4 would increase the accuracy, particularly for the more complex decisions. The advancement of large language models offers new opportunities for automating even complex decisions and labour-intensive manual tasks of SLRs. This assistance is particularly useful for SLRs of observational studies, which, compared to RCTs, are more varied due to a a variety of possible study types and lack of adherence to reporting guidelines.
- To benchmark accuracy of different LLMs over time, having a variety of benchmarking datasets is essential. Different types of SLRs may require different prompts or approaches, and benchmarking datasets would make it possible to measure performance across various types of studies and review tasks to ensure and demonstrate accuracy that is repeatable and transparent.

Case studies

- Case study 1 (IV vs SC): SLR to compare the outcomes for oncology therapies that have both an intravenously (IV) administered vs subcutaneous (SC) formulation from clinical trials and observational studies. Comparison of administration of oncology therapies
- Case study 2 (CLL): SLR of the efficacy and safety of current therapies in patients with relapsed or refractory chronic lymphocytic leukaemia or small lymphocytic lymphoma (R/R CLL/SLL)

References

1. Wagner G, Lukyanenko R, Paré G. Artificial intelligence and the conduct of literature reviews. *J Inf Technol*. 2021;37(2):209-226. doi:10.1177/02683962211048201
2. Reason T, Langham J, Malcolm B, Klijn S, Gimblett A. MSR46 Breaking Through Limitations: Enhanced Systematic Literature Reviews With Large Language Models. *Value Health*. 2023;26(12):S402. doi:10.1016/j.jval.2023.09.2105
3. Langham J, Reason T, Malcolm B, Klijn S, Gimblett A. MSR80 AI-Enabled Risk of Bias Assessment of RCTs in Systematic Reviews: A Case Study. *Value Health*. 2023;26(12):S408. doi:10.1016/j.jval.2023.09.2139
4. Schulz KF, Altman DG, Moher D, for the CONSORT Group. CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. *BMJ*. 2010;340(mar23 1):c332-c332. doi:10.1136/bmj.c332

Acknowledgments

- This study was supported by Bristol Myers Squibb

Presented at ISPOR Europe 2024; November 17-20, 2024; Barcelona, Spain

Email: julia.langham@estima-sci.com