Breaking Through Limitations: Enhanced Systematic Literature Reviews With Large Language Models

Reason, Tim, MSc, [1] Langham, Julia, Ph.D,[1] Gimblett, Andy, Ph.D,[1] Malcolm, Bill, MSc, [2] Klijn, Sven, MSc. [3]
[1] Estima Scientific, London, United Kingdom; [2] Bristol Myers Squibb, Uxbridge, UK; [3] Bristol Myers Squibb, Princeton, NJ, USA

Introduction

Assimilating and synthesising high volumes of existing and emerging evidence using traditional systematic literature review (SLR) methods is challenging and resource-intensive, with results often being out of date by the time they are published.[1-3]
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.[4-7]
The advancement of foundation models, including large language models (LLMs) like OpenAI's Generative Pre-Trained Transformer-4 (GPT-4), offers new opportunities for automating SLRs. Their functionality has huge potential for automating tasks such as data extraction (e.g., study characteristics and data) and text classification (e.g., categorizing article abstracts, full texts, references, etc.) with great accuracy.

Aim

Using a recently published SLR and network meta-analysis (NMA), the aim was to assess the accuracy of GPT-4 compared with traditional methods of double screening by human reviewers to identify eligible studies from title and abstract screening and through full-text review.

Key messages

What is already known on this topic

- High-quality SLRs play a critical role in evidence-based decision-making.
- AI can assist in automating some of the more labour-intensive manual tasks of SLRs, such as data extraction.
- The advancement of foundation models, including LLMs like GPT-4, offers new opportunities for automating SLRs.

What this study adds

- In this case study, we compared GPT-4 with double-screening by two human reviewers for title and abstract screening and for full-text review.
- GPT-4 performed with high sensitivity and specificity to identify the relevant eligible studies included in the case study.

How this study might affect research, practice or policy

- Our case study demonstrates that GPT-4 has the potential to quickly and efficiently assist in title and abstract screening to full-text review which can assist in producing up-to-date syntheses of all available evidence.

Methods

Case study: A published SLR and network meta-analysis (NMA) assessing the safety and efficacy of Nivolumab for advanced non-small cell lung cancer[8] was used as a case study. In the case study, screening was conducted by two independent human reviewers with a third reviewer for arbitration.
Titles and abstracts were screened after the removal of duplicates and incorrect publication types. Full-text screening was performed on the subset of publications that were identified by both human reviewers and GPT-4 (Figure 1).
GPT-4: A Python application programming interface (API) was used to send "prompts" and text (titles and abstracts, and text extracted from PDFs of full publications) to GPT-4 with instructions to summarise text and to answer questions regarding eligibility.

Metrics: GPT-4 screening results were compared with the final results from human reviewers, the 73 studies identified as eligible for inclusion, using the following metrics:

- Sensitivity: (the ability of GPT-4 to correctly identify eligible citations) TP / (TP + FN)
- Specificity: (the ability of GPT-4 to correctly exclude citations) TN / (TN + FP)
- Accuracy (percentage of correct classifications): (TP + TN) / (TP + TN + FP + FN)
- Precision (positive predictive value) : TP / (TP + FP)

Key: True positives (TP); True Negative (TN); False negative (FN); False positive(FP)

Results

The Prisma flow diagram for the title and abstract screening and full-text review compared between human and GPT-4 is shown in Figure 1. It shows that reasons for exclusion were similar for both human reviewers and GPT-4. The sensitivity and specificity of GPT-4 title and abstract screening, compared to the final set of eligible studies (Table 1), were 95.9% and 86.7% respectively (accuracy 87.0%, precision 22.7%).

Table 1 GPT-4 Title & abstract screening results compared to human final decision.

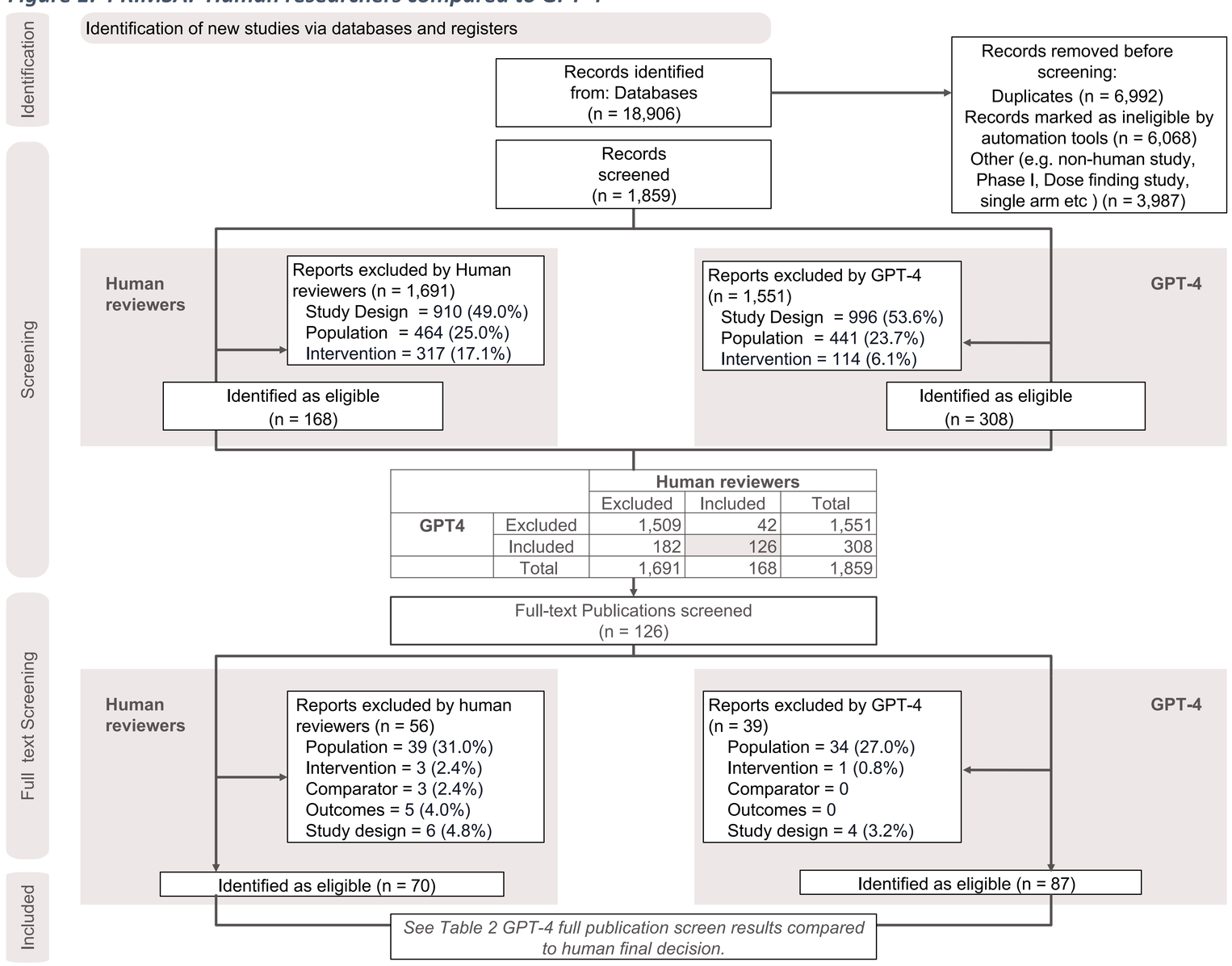
		Human researchers' final included studies		
		Exclude	Include	Total
GPT-4 title & abstract screen	Exclude	1548	3	1551
	Include	238	**70**	308
	Total	1786	73	1859

Table 2 GPT-4 full publication screen results compared to human final decision.

		Human researchers' final included studies		
		Exclude	Include	Total
GPT-4 full publication screen	Exclude	37	2	39
	Include	19	68	87
	Total	56	70	**126**

After full text review, sensitivity and specificity was 97.1% and 66.1% respectively (accuracy 83.3%, precision 78.2%) shown in Table 2.
The approximate time required for GPT-4 to process the information was one hour for 500 titles and abstracts screened, and one hour for 25 full text publications screened.

Figure 1: PRIMSA: Human researchers compared to GPT-4



Conclusions

- We found that GPT-4 can accurately summarise relevant study characteristics and determine eligibility both from titles and abstracts (sensitivity 95.9%, specificity 86.7%), and full-text screening (sensitivity 97.1%, specificity 66.1%). GPT-4 successfully identified the same set of studies that humans identified and that were included in the NMA of the case study.
- GPT-4 data extraction tables that were part of the output generated facilitate PRISMA-compliant tracking and quality assessment.
- There are clear time-savings as screening was accomplished in a fraction of the time it takes humans without compromising the quality of the NMA.
- 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. Further testing on new samples to improve prompting and to demonstrate generalisability are required.

References

1. Siddaway, A. P. . *et al*, How to do a Systematic Review: A Best Practice Guide for Conducting and Reporting Narrative Reviews, Meta-Analyses, and Meta-Syntheses. *Annu Rev Psychol* **70**, 747–770 (2019).
2. Rohit Borah . *et al*, Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. *BMJ Open* **7**, e012545 (2017).
3. Shojania, K. G. *et al.* How Quickly Do Systematic Reviews Go Out of Date? *Ann. Intern. Med.* **147**, 224–233 (2007).
4. Wagner, G., . *et al*. Artificial intelligence and the conduct of literature reviews. *J. Inf. Technol.* **37**, 209–226 (2021).
5. Blaizot, A. *et al.* Using AI methods for systematic review in health sciences. *Res Synth Methods* **13**, 353–362 (2022).
6. Tsafnat, G. *et al.* Systematic review automation technologies. *Syst Rev* **3**, 74 (2014).
7. O'Mara-Eves, A., Thomas, J., McNaught, J., Miwa, M. & Ananiadou, S. Using text mining for study identification in systematic reviews: a systematic review of current approaches. *Syst. Rev.* **4**, 5 (2015).
8. Cope, S. PCN29 COMPARATIVE EFFICACY OF NIVOLUMAB VERSUS RELEVANT TREATMENTS IN PRETREATED ADVANCED NON-SMALL CELL LUNG CANCER (NSCLC): An SLR and ITC of RCTS. *Value Health* **22**, s440 (2019).

Presented at ISPOR Europe 2023; November 11-15,; Copenhagen

Email: tim.reason@estima-sci.com

This study was supported by Bristol Myers Squibb