AI-Enabled Risk of Bias Assessment of RCTs in Systematic Reviews

Langham, Julia, Ph.D,[1] Reason, Tim, MSc,[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

A critical component of high-quality systematic literature reviews (SLRs) is the risk of bias (RoB) estimate. It requires a sophisticated understanding of the design and conduct of RCTs and requires considerable human time, expertise and training to ensure accuracy and agreement between human reviewers.[1]

This complexity and level of subjectivity poses a challenge for automation and such projects have had limited success.[2-5]

The advancement of foundation models, including large language models (LLMs) like OpenAI's Generative Pre-Trained Transformer-4 (GPT-4), offers new opportunities for overcoming these challenges.

LLMs have greater versatility than other purpose-built pre-trained AI tools, they have been trained on massive datasets, adapt to new and unseen tasks with minimal fine-tuning, and have a level of interactivity which allows them to be modified and queried easily.

The revised Cochrane Risk-of-Bias Tool for Randomized Trials (RoB 2)[6] uses a method of assigning bias by answering specific (signalling) questions which, potentially, makes the tool less subjective. The ability of LLMs to answer these questions has not been reported.

Aim

To investigate the feasibility of leveraging LLMs, specifically GPT-4, for automating RoB assessment in a case study of 12 RCTs using Cochrane RoB 2.

Key messages

What is already known on this topic

- High-quality SLRs require a RoB estimate for included studies, which requires considerable human time and expertise.
- LLMs have huge potential for automating tasks such as data extraction and text classification with great accuracy, but their potential for estimating RoB of RCTs has not been tested

What this study adds

- In this case study we show that GPT-4 can accurately estimate the RoB utilizing the Cochrane RoB 2 tool in a sample of RCTs.

How this study might affect research, practice or policy

- The results provide preliminary evidence supporting the utility of LLMs to assist, or (partially) replace, human reviewers in the assessment of RoB of RCTs enhancing the ability to complete SLRs more rapidly.

Box 1: Metrics for performance assessment

- Sensitivity: proportion of domains that GPT-4 correctly identified as 'Low Risk'.
- Specificity: proportion of domains that GPT-4 correctly identified as 'High Risk' or 'Some Concerns'.
- Positive Predictive Value (PPV): proportion of 'Low Risk' identifications by GPT-4 that were correct.
- Negative Predictive Value (NPV): proportion of 'High Risk' or 'Some Concerns' identifications by GPT-4 that were correct.
- Observed Agreement: proportion of agreement between GPT-4 and the human reviewer. (1) dichotomised answers "low risk" vs "some concerns – high risk"; (2) direct agreement with low, high risk or some concerns
- Cohen's Kappa: a statistical measure of agreement between GPT-4 and the human reviewer assessments (correct and incorrect) , accounting for the possibility of random agreement
  - Correct: True Positive : Both Human and GPT4 = "Low" ; True Negative : Both Human and GPT4 = "Some concerns" or "High"
  - Incorrect: False Positive : Human ="Some concerns" or "High" ; GPT-4 = "Low"; False Negative : Human ="Low" ; GPT-4 = "Some concerns" or "High"

Methods

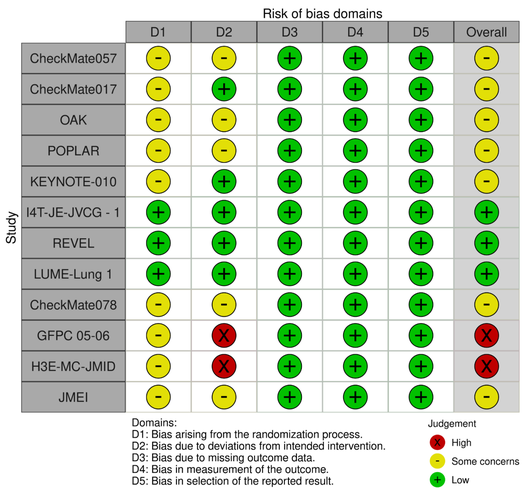
Case study: A published SLR and network meta-analysis (NMA) assessing the safety and efficacy of Nivolumab for advanced non-small cell lung cancer was used as a case study.[7] 12 RCTs were assessed for RoB using RoB 2 by answering 11 signalling questions and then using the published algorithm to estimate the risk of bias in each of the five domains and overall.

RoB was assessed by one independent human reviewer and by GPT-4 in the same set of studies.

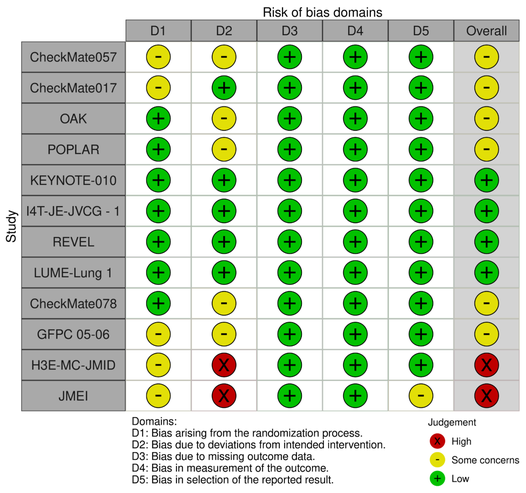
GPT-4 was given prompts (instructions) sent via a Python Application Programming Interface (API). Text (RCT publication and protocol) was sent in "chunks" to comply with the token limit. GPT-4 was instructed to summarize text and answer each of the signalling questions with 'Yes,' 'No,' or 'NI' (No Information) and extract information that supported the answer, equivalent to the spreadsheet filled out by the human. RoB in each domain was classified as 'Low Risk,' 'Some Concerns,' or 'High Risk' by applying the algorithm to signalling question answers. Estimates of RoB from the human reviewer and GPT-4 were compared to using the metrics shown in Box 1.

Figure 1. Risk of bias estimates: Human and GPT-4 [8]

Human reviewer RoB 2 Visualization for the 12 studies



GPT-4 RoB 2 Visualization for the 12 studies



Results

Table 1. Sensitivity and specificity metrics: Human vs GPT-4

Domain	TP n (%)	TN n (%)	FP n (%)	FN n (%)	Sensitivity % (95% CI)	Specificity % (95% CI)	PPV (%)	NPV (%)	Agreement of low and not low (%)	Agreement of low, high and some concerns (%)	Cohen's Kappa (95% CI)
DOMAIN 1: Bias Arising from the Randomization Process	3 (25.0%)	5 (41.7%)	4 (33.3%)	0	100%	55.6% (23.09 to 88.0)	43.0%	100%	67.0%	66.7%	0.38 (-0.11 to 0.88)
DOMAIN 2: Bias Due to Deviations from Intended Interventions	5 (41.7%)	7 (58.3%)	0	0	100%	100%	100%	100%	100%	83.3%	0.73 (0.40 to 1.07)
DOMAIN 3 : Bias Due to Missing Outcome Data	12 (100%)	0	0	0	100%	na	100%	0%	100%	100%	na
DOMAIN 4 : Bias in Measurement of the Outcome	12 (100%)	0	0	0	100%	na	100%	0%	100%	100%	na
DOMAIN 5: Risk of bias in selection of the reported result	11 (91.7%)	0	0	1 (8.3%)	91.7% (76.03 to 100)	na	100%	0%	92.0%	91.7%	0.00 (-1.88 to 1.88)
Overall bias assessment	3 (25.0%)	8 (66.7%)	1 (8.3%)	0	100%	88.9% (68.36 to 100)	75.0%	100%	92.0%	75.0%	0.58 (0.17 to 0.99)

CI: confidence interval; na: not applicable/calculable ; FN: False Negative; FP: False Positive; TN: True Negative; TP: True Positive
Calculations: specificity = TN / (TN + FP) if (TN + FP) ; sensitivity = TP / (TP + FN) if (TP + FN); Negative Predictive Value (NPV) = TN / (TN + FN) if (TN + FN); Positive Predictive Value (PPV) = TP / (TP + FP) if (TP + FP); observed agreement = (TP + TN)

GPT-4 successfully summarised provided text, extracted relevant text and answered the signalling questions in less than an hour in total for all 12 studies. RoB was then estimated via the Cochrane algorithm (Figure 1).

There was high agreement between the human reviewer and GPT-4 estimates of RoB both overall and across domains (Table 1).

- Complete agreement for 3 domains (Missing Outcome Information, Measurement of Outcome and Selection of Reported Result).
- Good agreement for Domain 2 (Bias Due to Deviations from Intended Interventions) Kappa value 0.73 (95% CI 0.40 to 1.07).
- Fair agreement for Domain 1: (Bias Arising from the Randomization Process) 0.38 (95% CI -0.11 to 0.88) and for the overall assessment: 0.58 (0.17 to 0.99).

Box 2. Included RCTs

POPLAR, Fehrenbacher 2016: https://linkinghub.elsevier.com/retrieve/pii/S0140673616005870
CheckMate017, Brahmer 2015 : http://www.nejm.org/doi/10.1056/NEJMoa1504627
CheckMate057, Borghaei 2015 : http://www.nejm.org/doi/10.1056/NEJMoa1507643
KEYNOTE-010, Herbst 2016 : https://linkinghub.elsevier.com/retrieve/pii/S0140673615012817
I4T-JE-JVCG – 1, Yoh 2016: https://linkinghub.elsevier.com/retrieve/pii/S0169500216304172
REVEL, Garon 2014 : https://linkinghub.elsevier.com/retrieve/pii/S014067361460845X
LUME-Lung 1, Reck 2014 : https://linkinghub.elsevier.com/retrieve/pii/S1470204513705862
GFPC 05-06, Vergnenegre 2011 : https://linkinghub.elsevier.com/retrieve/pii/S1556086415319110
H3E-MC-JMID, Sun 2013 : https://linkinghub.elsevier.com/retrieve/pii/S0169500212006162
JMEI, Hanna 2004 : https://doi.org/10.1200/JCO.2004.08.163
OAK, Rittmeyer 2017 : https://linkinghub.elsevier.com/retrieve/pii/S014067361632517X
CheckMate 078, Wu 2019 : https://linkinghub.elsevier.com/retrieve/pii/S1556086419300206

Conclusions

- GPT-4 accurately answered RoB signalling questions to calculate RoB, closely agreeing with the human estimate and producing data extraction tables to facilitate validation and quality control.
- There were clear time-savings, and the output (data extraction tables) improved decision transparency and feasibility of quality assessment.
- Development of detailed prompts was required to obtain these results and the generalisability will need to be tested on further samples. Further prompt refinement, particularly for more complex decisions will improve accuracy.
- As we refine prompting and employ fine-tuning methods (training GPT-4 on a greater number of examples) the quality and generalisability of automated RoB assessments are expected to further improve. As LLMs evolve, the usability and accuracy of using LLMs in complex tasks will improve.

References

1. Minozzi, S., *et al.* Reliability of the revised Cochrane risk-of-bias tool for randomised trials (RoB2) *J. Clin. Epidemiol.* **141**, 99–105 (2022).
2. Marshall, I ., *et al.*. RobotReviewer: evaluation of a system for automatically assessing bias in clinical trials. *J. Am. Med. Inform. Assoc.* **23**, 193–201 (2016).
3. Higgins, J. P. T. *et al.* The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *BMJ* **343**, d5928–d5928 (2011).
4. Armijo-Olivo, S ., *et al.*. Comparing machine and human reviewers to evaluate the risk of bias in randomized controlled trials. *Res Synth Methods* **11**, 484–493 (2020).
5. marshall, C. & sutton, A. (2022). The Systematic Review Toolbox. Available from: http://www.systematicreviewtools.com/}. (2022).
6. Sterne, J. A. C. *et al.* RoB 2: a revised tool for assessing risk of bias in randomised trials. *BMJ* **366**, l4898 (2019).
7. Cope, S. PCN29 COMPARATIVE EFFICACY OF NIVOLUMAB VERSUS RELEVANT TREATMENTS. *Value Health* **22**, s440 (2019).
8. McGuinness, L. A. & Higgins, J. P. T. Risk-of-bias VISualization (robvis):. *Res Synth Methods* **12**, 55–61 (2021).

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

Email: julia.langham@estima-sci.com

This study was supported by Bristol Myers Squibb