Large Language Models for Risk of Bias Assessment of Randomized Controlled Trials: A Case Study

Edwards M[1], Ferrante di Ruffano L[1]

[1] York Health Economics Consortium, University of York, York, YO10 5NQ, UK

INTRODUCTION

Risk of bias (RoB) assessment of primary studies is a key part of any systematic review. Despite the highly structured nature of RoB assessment tools, the process is open to subjectivity, and achieving consistency in judgments across a review requires expertise and can be time consuming.

As a repetitive and structured task, risk of bias assessment would initially appear to be well suited to automation or AI support, and the involvement of a non-human tool may aid consistency in decision making and achieve time efficiencies.

We assessed the chat interface to a large language model (LLM), Claude 3 Opus[1], for accuracy, consistency, presentation of data, and time savings in the context of risk of bias assessment of reports of randomized controlled trials (RCTs) for a systematic review. A "zero-shot" approach was used, i.e. the model was used to complete a task for which no specific prior examples or training were offered by the user.

METHODS

Six RCTs were selected from three reviews conducted by our consultancy over the past five years. Three methods were used to conduct RoB assessment using the Cochrane RoB 1 tool[2] (summarised in Figure 1).

Figure 1: Cochrane RoB 1 tool

Cochrane Collaboration's tool for assessing risk of bias[2]

Reviewer authors should assess each domain as low, unclear or high risk of bias

Random sequence generation: Describe the method used to generate the allocation sequence in sufficient detail to allow an assessment of whether it should produce comparable groups.
Allocation concealment: Describe the method used to conceal the allocation sequence in sufficient detail to determine whether intervention allocations could have been foreseen before or during enrolment.
Blinding of participants and personnel: Describe all measures used, if any, to blind trial participants and researchers from knowledge of which intervention a participant received. Provide any information relating to whether the intended blinding was effective.
Blinding of outcome assessment: Describe all measures used, if any, to blind outcome assessment from knowledge of which intervention a participant received. Provide any information relating to whether the intended blinding was effective.
Incomplete outcome data: Describe the completeness of outcome data for each main outcome, including attrition and exclusions from the analysis.
Selective reporting: State how selective outcome reporting was examined and what was found.
Other bias: State any important concerns about bias not covered in the other domains in the tool.

Method 1: Conduct fully automated assessment of each paper using the LLM, with the model providing both a judgement for each question, and a descriptive rationale for that judgement.

Method 2: The LLM supplied information relating to each question to facilitate a human judgement. Human judgements were checked by a second independent reviewer.

Method 3: Fully human assessment, in which a single human review provided a judgement on each question and a descriptive rationale; these were both checked by a second independent reviewer.

The prompts used for Methods 1 and 2 were developed following an initial prompt engineering phase using a report of a seventh RCT.

We recorded the results of the three methods in an Excel spreadsheet and compared the agreement between methods for each question. Following a review of the LLM generated portion of the results by a second independent reviewer, we also noted whether the LLM had identified information not recorded by the human reviewer.

RESULTS

Method 1 resulted in very brief answers, with little supporting information provided by the model. Method 2, asking for supporting information only, resulted in the LLM extracting better quality and more complete data. Table 1 illustrates agreement across the six studies assessed, for each domain. Green indicates 100% agreement across studies, yellow more than 50% agreement, and red 50% or less.

Table 1: Agreement between methods across the 6 RCTs assessed

Risk of bias domain	Percentage agreement		
	Human v. LLM (method 3 v. method 1)	Human v. LLM & human (method 3 v. method 2)	LLM v. LLM & human (method 1 v. method 2)
1. Was the allocation sequence adequately generated?	100%	100%	100%
2. Was the concealment of treatment allocation adequate?	67%	67%	67%
3. Was knowledge of the allocated interventions adequately prevented from participants and personnel?	67%	83%	83%
4. Was knowledge of the allocated interventions adequately prevented from outcome assessors?	50%	50%	50%
5. Were incomplete outcome data adequately addressed?	100%	100%	100%
6. Are reports of the study free of suggestion of selective outcome reporting?	50%	33%	67%
7. Was the study apparently free of other problems that could put it at a high risk of bias?			

The model and the human reviewers made consistent judgements across allocation generation and selective outcome reporting domains, although we note that this may have been influenced by the fact that all six RCTs assessed provided adequate sequence generation and full outcome reporting.

The model tended to be poor at identifying data relating to handling of incomplete data, and the rationale data provided by the LLM suggested that it did not correctly comprehend the question asked.

The final question ("other problems") is arguably the most open to subjectivity when completed by human reviewers. This domain saw heterogeneity across the six studies in the judgements arrived at using all three methods, with the lowest levels of agreement between methods 3 and 2, i.e. fully human v. human judgments based on data identified by the LLM. Disagreements related to whether a short study duration and the lack of an active comparator arm contributed to risk of bias. To achieve better consistency, a model could be pre-trained with information on what issues should and should not be considered as contributing to risk of bias. However, such an inflexible approach does not allow for subtle differences in study design, purpose, and context.

For domain 3, the LLM identified some information not picked up by the human reviewer. This may have been due to blinding information being reported in an unexpected location within the published paper(s).

CONCLUSION

Despite the widespread availability of online data relating to use of the Cochrane RoB 1 tool for assessing RCTs, the zero-shot use of LLMs for fully automated risk of bias assessment is not currently recommended over two experienced human reviewers. LLMs can misinterpret questions and provide limited or incorrect justification for judgments, and this problem is likely to be more pronounced for less widely used tools / study designs, such as economic evaluations.

Even when used as support by a single inexperienced reviewer, the use of a LLM for RoB assessment comes with the risk of failing to identify and critically engage with the methodological problems of a given study. However, with suitable prompt engineering, and training of models using existing data, the opportunities offered by LLMs for assisting in conducting RoB assessment are likely to improve with time.

REFERENCES

1. https://www.anthropic.com/claude 2. Higgins et al. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. BMJ. 2011.

CONTACT US

mary.edwards@york.ac.uk
+44 1904 323437
www.yhec.co.uk
York Health Economics Consortium

Providing Consultancy & Research in Health Economics

UNIVERSITY of York

INVESTORS IN PEOPLE® We invest in people Gold

YHEC
York Health Economics Consortium