Roberta on the Job

A proof-of-concept study of a customised GPT tool for risk-of-bias assessment of randomised clinical trials

MSR 183

Eric Julian Manalastas || Juliette C Thompson || Aditi Hombali || David A Scott
Eric.Manalastas@VisibleAnalytics.co.uk // Juliette.Thompson@VisibleAnalytics.co.uk // Aditi.Hombali@VisibleAnalytics.co.uk // David.Scott@VisibleAnalytics.co.uk

INTRODUCTION

Risk-of-bias (ROB) assessment is a key element of systematic reviews.[1] However, procedures using instruments such as the Cochrane Risk of Bias Tool for Randomised Trials (ROB-2),[2,3] are known to be:

- challenging to use;[4–6]
- highly time-consuming;[4,5,7]
- marred by low inter-rater consistency.[7–9]

Various attempts have been made to use general-purpose large language models (LLMs) such as ChatGPT (Generative Pre-trained Transformer) for ROB-2 assessment.[10–13] However, the success of these attempts have been *mixed,* with fair to moderate agreement with human reviewers, hallucinations, and unknown consistency over time.

OUR KEY IDEA

A customised GPT-based tool may have the potential to assist in ROB-2 assessments with greater efficiency and consistency.

Customisation refers to the process of adapting a generic LLM to become more domain-specific, targeting particular operations and using specialised knowledge.[14,15]

AIM

To develop and assess a simple, customised GPT-based tool for ROB-2 assessment

METHODS

Study design

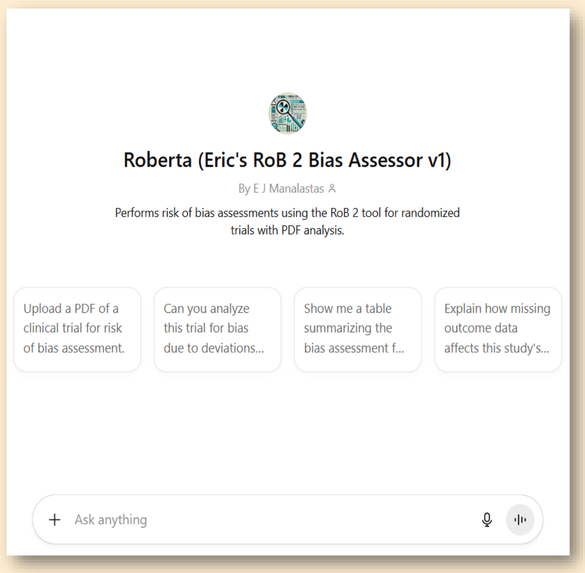
Proof-of-concept study.

Tool development

A customised GPT for ROB-2 assessment of randomised trials, aligned with Cochrane standards (nicknamed: Roberta)

- **Model:** OpenAI's GPT-4o architecture
- **Customisation:** Retrieval-augmented generation using published guidance documents for the ROB-2 tool.

Roberta's interface (screenshot)



Testing & evaluation

We tested Roberta on three evaluation criteria:

- **Speed** (time to generate a ROB-2 assessment)
- **Accuracy** (performance against ratings by trained human assessors)
- **Test–retest reliability** (consistency of assessments across time points)

RESULTS

Speed — Evaluation for speed indicated extremely rapid performance.

In tests using 20 clinical trial publications, average time per ROB-2 assessment was less than half a minute (M: 25.2 seconds, SD: 4.7).

Accuracy — Evaluation for accuracy showed variation by level of ROB.

For domains rated by human assessors as low-risk and 'some concerns', high agreement was observed. However, considerable disagreement was observed in domains rated as 'high-risk', with Roberta ratings being more lenient than human assessors.

Reliability — Evaluation for test–retest reliability indicated acceptable reliability across one week (Cohen's kappa = 1.00; perfect consistency) and two weeks (Cohen's kappa = 0.54; moderate consistency).

VERY HIGH SPEED

25.2 seconds on average to draft a ROB-2 assessment

vs. human standard of 28 minutes per assessment[5]

FAIR ACCURACY

96% agreement for low-risk domains

83% agreement for domains with 'some concerns'

ACCEPTABLE RELIABILITY

κ=1.00 perfect consistency across one week

κ=0.54 moderate consistency across two weeks

CONCLUSIONS

Preliminary proof-of-concept evaluations suggest a customised GPT tool such as Roberta can support rapid ROB-2 assessments with a fair degree of accuracy and acceptable test–retest reliability, complementing – but not replacing – human review.

Further testing will enhance the utility of LLM-based (including GPT) tools, towards realising the potential of AI to facilitate systematic reviews and evidence synthesis.[17,18]

Specifically, future tests should:

- Compare the customised GPT versus its base version (i.e. uncustomised GPT)
- Evaluate the customised GPT using the latest models which are rapidly evolving
- Explore the utility of such tools to ROB assessments of non-randomised trials

REFERENCES

1. Viswanathan M, Patnode CD, Berkman ND, et al. Recommendations for assessing the risk of bias in systematic reviews of health-care interventions. J Clin Epidemiol. 2018;97:26-34. doi:10.1016/j.jclinepi.2017.12.004
2. Higgins J P T, Altman D G, GÃ¸tzsche P C, JÃ¼ni P, Moher D, Oxman A D et al. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials BMJ 2011; 343 :d5928 doi:10.1136/bmj.d5928
3. Sterne JAC, Savović J, Page MJ, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ. 2019;366:l4898. Published 2019 Aug 28. doi:10.1136/bmj.l4898
4. Crocker TF, Lam N, Jordão M, et al. Risk-of-bias assessment using Cochrane's revised tool for randomized trials (RoB 2) was useful but challenging and resource-intensive: observations from a systematic review. J Clin Epidemiol. 2023;161:39-45. doi:10.1016/j.jclinepi.2023.06.015
5. Minozzi S, Cinquini M, Gianola S, Gonzalez-Lorenzo M, Banzi R. The revised Cochrane risk of bias tool for randomized trials (RoB 2) showed low interrater reliability and challenges in its application. J Clin Epidemiol. 2020;126:37-44. doi:10.1016/j.jclinepi.2020.06.015
6. Savović J, Weeks L, Sterne JA, et al. Evaluation of the Cochrane Collaboration's tool for assessing the risk of bias in randomized trials: focus groups, online survey, proposed recommendations and their implementation. Syst Rev. 2014;3:37. Published 2014 Apr 15. doi:10.1186/2046-4053-3-37
7. Minozzi S, Dwan K, Borrelli F, Filippini G. Reliability of the revised Cochrane risk-of-bias tool for randomised trials (RoB2) improved with the use of implementation instruction. J Clin Epidemiol. 2022;141:99-105. doi:10.1016/j.jclinepi.2021.09.021
8. Armijo-Olivo S, Ospina M, da Costa BR, et al. Poor reliability between Cochrane reviewers and blinded external reviewers when applying the Cochrane risk of bias tool in physical therapy trials. PLoS One. 2014;9(5):e96920. Published 2014 May 13. doi:10.1371/journal.pone.0096920
9. Hartling L, Hamm MP, Milne A, et al. Testing the risk of bias tool showed low reliability between individual reviewers and across consensus assessments of reviewer pairs. J Clin Epidemiol. 2013;66(9):973-981. doi:10.1016/j.jclinepi.2012.07.005
10. Šuster S, Baldwin T, Verspoor K. Zero- and few-shot prompting of generative large language models provides weak assessment of risk of bias in clinical trials. Res Synth Methods.
11. Pitre T, Tanvir Jassal, Jhalok Ronjan Talukdar, Mahnoor Shahab, Michael Ling, Dena Zeraatkar. ChatGPT for assessing risk of bias of randomized trials using the RoB 2.0 tool: A methods study. medRxiv 2023.11.19.23298727; doi: https://doi.org/10.1101/2023.11.19.23298727
12. Lai H, Ge L, Sun M, et al. Assessing the Risk of Bias in Randomized Clinical Trials With Large Language Models. JAMA Netw Open. 2024;7(5):e2412687. Published 2024 May 1. doi:10.1001/jamanetworkopen.2024.12687
13. Taneri PE. Human Versus Artificial Intelligence: Comparing Cochrane Authors' and ChatGPT's Risk of Bias Assessments. Cochrane Evid Synth Methods. 2025;3(5):e70044. Published 2025 Aug 31. doi:10.1002/cesm.70044
14. Jiang Z, Liu A, Zhang D, Xiwei Xu, Dai Y. Customization and personalization of large language models for engineering design. CIRP Ann Manufact Technol. 2025;74(1):191-195 https://doi.org/10.1016/j.cirp.2025.03.001
15. Chen, J., Liu, Z., Huang, X. et al. When large language models meet personalization: perspectives of challenges and opportunities. World Wide Web 27, 42 (2024). https://doi.org/10.1007/s11280-024-01276-1
16. Huang J, Lai H, Zhao W, et al. Large Language Model–Assisted Risk-of-Bias Assessment in Randomized Controlled Trials Using the Revised Risk-of-Bias Tool: Evaluation Study. J Med Internet Res. 2025;27:e70450. Published 2025 Jun 24. doi:10.2196/70450
17. Rose CJ, Bidonde J, Ringsten M, et al. Using a large language model (ChatGPT) to assess risk of bias in randomized controlled trials of medical interventions: protocol for a pilot study of interrater agreement with human reviewers. BMC Med Res Methodol. 2025;25(1):182. Published 2025 Jul 31. doi:10.1186/s12874-025-02631-0
18. Rose, C. J., Bidonde, J., Ringsten, M., Glanville, J., Potrebny, T., Cooper, C., Muller, A. E., Bergsund, H. B., Meneses-Echavez, J. F., & Berg, R. C. (2025). Using a Large Language Model (ChatGPT-4o) to Assess the Risk of Bias in Randomized Controlled Trials of Medical Interventions: Interrater Agreement With Human Reviewers. Cochrane evidence synthesis and methods, 3(5), e70048. https://doi.org/10.1002/cesm.70048

Visible Analytics
from data to decisions™

ISPOR Europe | Glasgow | 09-12 November 2025