Original Paper

Large Language Model–Assisted Risk-of-Bias Assessment in Randomized Controlled Trials Using the Revised Risk-of-Bias Tool: Evaluation Study

Jiajie Huang[1,2], MSN; Honghao Lai[1,2], MM; Weilong Zhao[1,2], MM; Danni Xia[1,2], MM; Chunyang Bai[3], MM; Mingyao Sun[4], MSN; Jianing Liu[5], MSN; Jiayi Liu[1,2], MM; Bei Pan[6,7], MM; Jinhui Tian[6,7], MD; Long Ge[1,2,6], MD

[1]Department of Health Policy and Management, School of Public Health, Lanzhou University, Lanzhou, China
[2]Evidence-Based Social Science Research Center, School of Public Health, Lanzhou University, Lanzhou, China
[3]School of Nursing, Southern Medical University, Guangzhou, China
[4]School of Nursing, Peking University, Beijing, China
[5]College of Nursing, Gansu University of Traditional Chinese Medicine, Lanzhou, China
[6]Evidence-Based Medicine Center, School of Basic Medical Sciences, Lanzhou University, Lanzhou, China
[7]Key Laboratory of Evidence Based Medicine of Gansu Province, Lanzhou, China

Corresponding Author:
Long Ge, MD
Department of Health Policy and Management
School of Public Health
Lanzhou University
No. 222 South Tianshui Road
Lanzhou, 730000
China
Phone: 86 13893192463
Email: gelong2009@163.com

Related Article:
This is a corrected version. See correction statement in: https://www.jmir.org/2025/1/e80519

Abstract

Background: The revised Risk-of-Bias tool (RoB2) overcomes the limitations of its predecessor but introduces new implementation challenges. Studies demonstrate low interrater reliability and substantial time requirements for RoB2 implementation. Large language models (LLMs) may assist in RoB2 implementation, although their effectiveness remains uncertain.

Objective: This study aims to evaluate the accuracy of LLMs in RoB2 assessments to explore their potential as research assistants for bias evaluation.

Methods: We systematically searched the Cochrane Library (through October 2023) for reviews using RoB2, categorized by interest in adhering or assignment. From 86 eligible reviews of randomized controlled trials (covering 1399 RCTs), we randomly selected 46 RCTs (23 per category). In addition, 3 experienced reviewers independently assessed all 46 RCTs using RoB2, recording assessment time for each trial. Reviewer judgments were reconciled through consensus. Furthermore, 6 RCTs (3 from each category) were randomly selected for prompt development and optimization. The remaining 40 trials established the internal validation standard, while Cochrane Reviews judgments served as external validation. Primary outcomes were extracted as reported in corresponding Cochrane Reviews. We calculated accuracy rates, Cohen κ, and time differentials.

Results: We identified significant differences between Cochrane and reviewer judgments, particularly in domains 1, 4, and 5, likely due to different standards in assessing randomization and blinding. Among the 20 articles focusing on adhering, 18 Cochrane Reviews and 19 reviewer judgments classified them as "High risk," while assignment-focused RCTs showed more heterogeneous risk distribution. Compared with Cochrane Reviews, LLMs demonstrated accuracy rates of 57.5% and 70% for overall (assignment) and overall (adhering), respectively. When compared with reviewer judgments, LLMs' accuracy rates were 65% and 70% for these domains. The average accuracy rates for the remaining 6 domains were 65.2% (95% CI 57.6-72.7) against Cochrane Reviews and 74.2% (95% CI 64.7-83.9) against reviewers. At the signaling question level, LLMs achieved 83.2% average accuracy (95%

CI 77.5-88.9), with accuracy exceeding 70% for most questions except 2.4 (assignment), 2.5 (assignment), 3.3, and 3.4. When domain judgments were derived from LLM-generated signaling questions using the RoB2 algorithm rather than direct LLM domain judgments, accuracy improved substantially for Domain 2 (adhering; 55-95) and overall (adhering; 70-90). LLMs demonstrated high consistency between iterations (average 85.2%, 95% CI 85.15-88.79) and completed assessments in 1.9 minutes versus 31.5 minutes for human reviewers (mean difference 29.6, 95% CI 25.6-33.6 minutes).

Conclusions: LLMs achieved commendable accuracy when guided by structured prompts, particularly through processing methodological details through structured reasoning. While not replacing human assessment, LLMs demonstrate strong potential for assisting RoB2 evaluations. Larger studies with improved prompting could enhance performance.

(J Med Internet Res 2025;27:e70450) doi: 10.2196/70450

KEYWORDS

systematic review; large language models; risk of bias 2; artificial intelligence; efficiency

Introduction

Systematic reviews (SRs) are important for summarizing all relevant published evidence on a specific research question. They play a key role in developing guidelines and making health care decisions [1]. During the conduct of a systematic review, assessing the risk of bias (RoB) in the included primary studies is essential, as it allows the identification of potential flaws and assessment of the internal validity of the review's results [2]. The Cochrane RoB tool is the most widely used for assessing RoB in randomized controlled trials (RCTs) in both Cochrane and non-Cochrane SRs [3]. In 2019, the revised version of this tool RoB2 was released to address limitations of the previous version, such as inconsistent domain use and the lack of an overall judgment domain. RoB2 evaluates potential biases from the randomization process, deviations from intended interventions, missing data, measurement of the outcomes, selection of the reported results, and overall bias. For each domain, RoB2 guides the reviewer in formulating a judgment on the RoB, which can be expressed as "low," "some concern," or "high."

However, the study by Minozzi et al [4] demonstrated low interrater reliability (IRR) of RoB2, indicating significant challenges in its application. Further experiments were subsequently conducted to evaluate the reliability of RoB2, with the findings suggesting that the observed discrepancies in assessment might be attributed to a lack of sufficient subject matter expertise. This knowledge gap is particularly critical for accurately addressing the "deviations from intended interventions" and "measurement of outcomes" domains [5]. A multitude of questions and high professional knowledge requirements make RoB2 challenging to apply and increase the time required for systematic reviews. Studies have shown that the time needed to assess one outcome using RoB2 can range from approximately 28 to 40 minutes [4,5]. This is an important reason why only 69.3% of SRs use RoB2 or even 28.8% of SRs to evaluate multiple outcomes [6].

Large language models (LLMs) have gradually been embraced by the medical field due to their excellent text comprehension, information extraction, and language processing capabilities, and are regarded as key technologies that could revolutionize medical practice and research [7,8]. LLMs demonstrate unique advantages when dealing with the RoB2 assessment framework, they can not only identify logical connections between different parts of RCT reports through attention mechanisms but also precisely capture details in complex trial methodologies, which is particularly important for areas requiring comprehensive judgment. The chain-of-thought reasoning ability of LLMs enables them to simulate the step-by-step decision-making process required by professional reviewers performing RoB2 evaluations, thereby reducing inconsistencies caused by subjective human judgment. By maintaining standardized assessment criteria across different studies, LLMs have the potential to address the IRR issues frequently observed in previous research. Existing studies have confirmed that with specially designed prompting techniques, LLMs can successfully perform a modified version of the Cochrane ROB tool assessment, significantly reducing the human resources and time costs associated with traditional assessment methods [9,10]. However, given the multilevel complexity of the RoB2 structure and the vast amount of information that needs to be integrated during the assessment process, there is still insufficient evidence supporting the effectiveness of LLMs in performing complete RoB2 assessments in systematic reviews.

This research aims to explore prompt engineering methods based on the Claude (Anthropic) model to address the efficiency and consistency challenges currently faced in RoB2 assessment practice. Our main purpose is to comprehensively evaluate whether LLMs can skillfully apply the RoB2 tool to professionally assess RCTs and to compare its assessment results with those of human experts using strict noninferiority standards. Through this study, we hope to provide empirical foundations for establishing a reliable LLM-assisted systematic review methodology, thereby improving the efficiency and quality of evidence synthesis in medical research.

Methods

Study Design

This feasibility study aims to investigate the capability of LLMs in assessing the RoB of RCTs with RoB2.

Data Source

We searched the Cochrane library using the terms: (Revised Cochrane RoB) OR RoB2 OR RoB-2 from inception through October 8, 2023, to identify systematic reviews that assess the RoB in RCTs with RoB2. We categorized the Cochrane Reviews based on the domain of deviations from intended interventions (assignment or adhering) of interest in Cochrane Reviews. We extracted all RCTs included in Cochrane Reviews, and the extracted results were entered into Microsoft Excel, with each RCT assigned a unique identifier. Subsequently, we generated 23 random numbers for both categories using a computer and matched these random numbers with the identifiers of the RCTs to determine which RCTs would be assessed. For each RCT included, we extracted the primary outcome as reported in the corresponding Cochrane review.

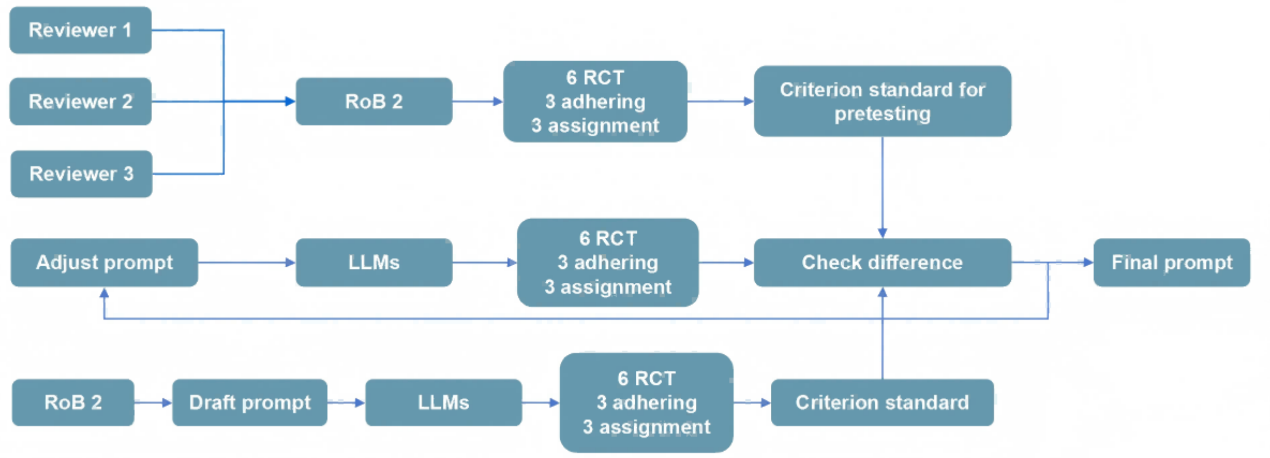
Establishment of the Criterion Standard

We selected 3 experienced reviewers to assess the RoB of the 46 RCTs included. None of the reviewers had previous exposure to the selected RCTs. All reviewers completed a week-long training on systematic reviews to ensure a consistent understanding of RoB2. Three reviewers (WZ, DX, and CB) initially conducted independent judgments of 46 RCTs using standardized criteria, recording the time taken for the judgments. All results were resolved through consensus. We randomly selected 3 evaluation results from each category to construct the prompt, which were used as benchmarks to assess the accuracy of the answers generated by the LLMs.

Prompt Construction

We designed a structured draft prompt based on the documents publicly available on the RoB2 website. The prompt was intended to guide LLMs in completing the following tasks: identify and extract key information relevant to each signaling question within RoB2, respond to the signaling questions, make judgments on each domain, and provide the basis for the judgments. We used the draft prompt to assess 6 RCTs [11-16], compared the outcomes with the criterion standard, and iteratively refined the prompt based on the responses. This process was repeated until satisfactory performance was attained. Figure 1 shows the main study process. The final prompt can be found in Multimedia Appendix 1.

Figure 1. Flow diagram of the study. LLMs: large language models, RCTs: randomized controlled trials, RoB2: risk-of-bias tool.


Reviewer 1
Reviewer 2
Reviewer 3
RoB 2
6 RCT 3 adhering 3 assignment
Criterion standard for pretesting
Adjust prompt
LLMs
6 RCT 3 adhering 3 assignment
Check difference
Final prompt
RoB 2
Draft prompt
LLMs
6 RCT 3 adhering 3 assignment
Criterion standard

LLMs Assessment

The final prompt was used to guide Claude 3.5 Sonnet in conducting assessments from October 10, 2023, to August 22, 2024, with no continued training or fine-tuning. The outputs were accurately transcribed into Multimedia Appendix 1. Any assessment interrupted by technical issues was excluded and promptly redone. Each RCT was assessed twice with Claude 3.5, using the same prompt and ensuring consistent model versions. Throughout the process, strict protocol adherence was maintained to ensure assessment quality.

Statistical Analysis

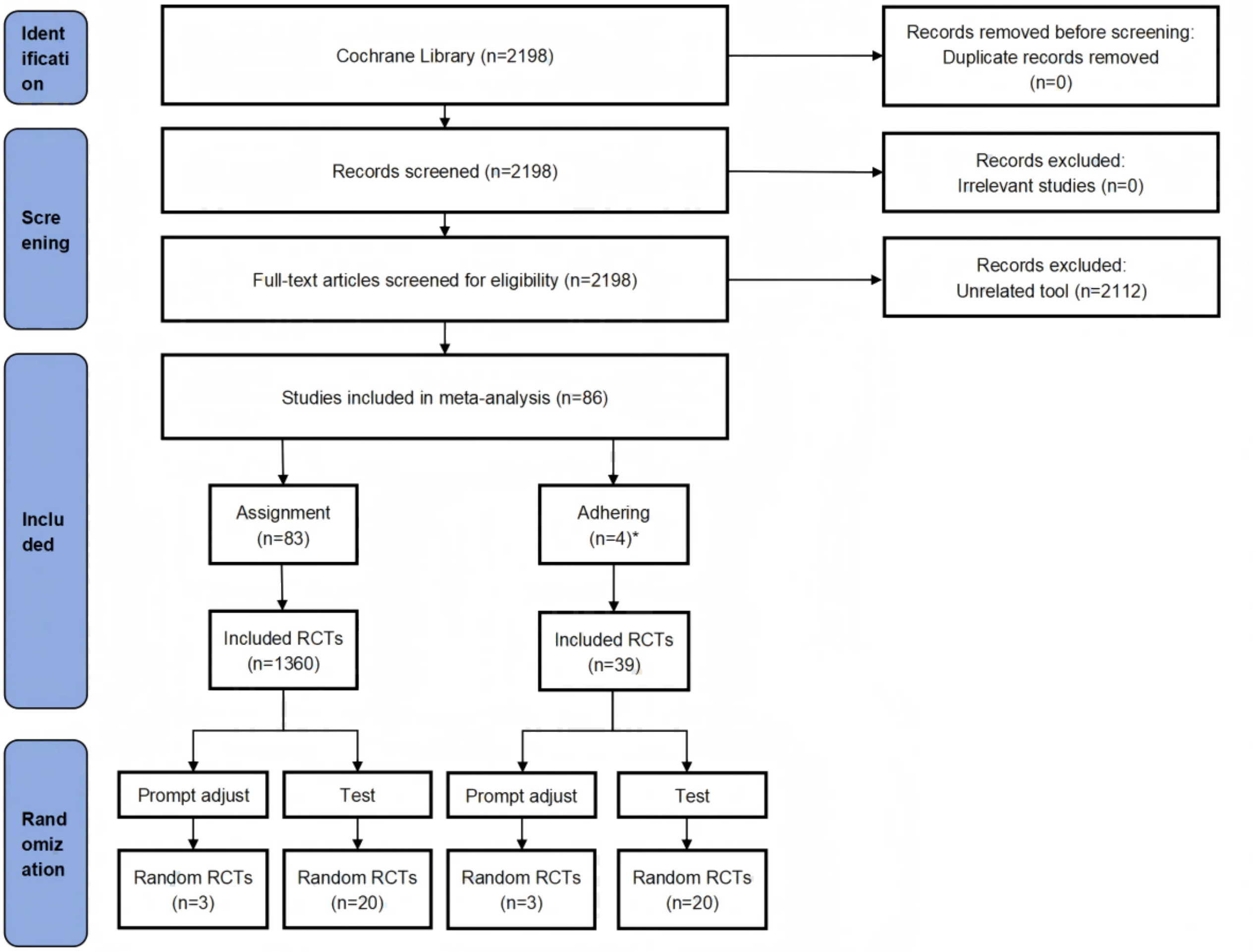
We performed a descriptive analysis to evaluate the accuracy of the LLM-generated judgments in comparison with the judgments of reviewers and the judgments of Cochrane systematic reviews. This analysis focused on the accuracy of domain judgment and signaling question judgment. During prompt development and optimization, we observed that the inclusion of an “NA” option led to its overuse by the LLMs, potentially interfering with domain judgment. Consequently, the final prompt omitted instructions for generating “NA” responses. In our comparative analysis, signaling questions judged as “NA” by reviewers were excluded from accuracy calculations, as LLM-generated results for these questions would not influence the outcome. Due to the logical issues previously identified in the LLMs during previous research [9], we further investigated the discrepancies between the domain judgments generated by the LLMs and the judgments formed by the LLM-generated signaling questions according to the RoB2 algorithm. To measure the stability of LLMs, we calculated the consistent assessment rates and Cohen κ in domains and signaling questions [5]. To measure the IRR, we calculated the Fleiss’ κ for reviews for individual domains. We compared the time spent by the LLMs and the reviewer group to assess the efficiency gains obtained through using LLMs.

Results

Overview

Our initial search identified 2198 records. We subsequently reviewed the full texts of 2198 potentially eligible Cochrane Reviews. Finally, 86 Cochrane Reviews including 1399 RCTs were determined to meet the inclusion criteria (Figure 2). Details of included studies can be found in Table S1 in Multimedia Appendix 1.

Figure 2. Literature screening and randomization flow diagram. RCT: randomized controlled trial.


Identification
Screening
Included
Randomization
Cochrane Library (n=2198)
Records removed before screening: Duplicate records removed (n=0)
Records screened (n=2198)
Records excluded: Irrelevant studies (n=0)
Full-text articles screened for eligibility (n=2198)
Records excluded: Unrelated tool (n=2112)
Studies included in meta-analysis (n=86)
Assignment (n=83)
Adhering (n=4)*
Included RCTs (n=1360)
Included RCTs (n=39)
Prompt adjust
Test
Prompt adjust
Test
Random RCTs (n=3)
Random RCTs (n=20)
Random RCTs (n=3)
Random RCTs (n=20)

Characteristics of Included Cochrane Reviews and RCTs

Among 86 Cochrane Reviews, 83 were interested in the effect of assignment to the domain of bias due to deviations from intended interventions, and 4 were interested in adhering. Among these, 1 review reported on the assessment of both assignment and adhering. The 20 RCTs [17-36] interested in assignment were drawn from 19 Cochrane reviews, 2 of these studies focused on COVID-19, while the other studies focused on adolescents, patients who initiated in vitro fertilization, and others. The interventions included medications, exercise, supplements, meditation, and others. The 20 RCTs [37-56] interested in adhering were drawn from 4 Cochrane Reviews. The primary population of interest for the RCTs was chronic obstructive pulmonary disease (COPD), followed by asthma. Interventions included in the RCTs involved drugs, exercise, supplements, and others (Table S2 in Multimedia Appendix 1).

RoB2 Assessment With Reviewers, the Cochrane Reviewers, and LLMs

Among the 20 articles focusing on adhering, 18 Cochrane Reviews and 19 reviewer judgments resulted in “High risk,” primarily due to low reporting rates of adhering-related information, which led to all articles examining adhering being classified as “High risk.” In contrast, the distribution of RoB of RCTs focusing on allocation was more heterogeneous. The majority of these RCTs were categorized as having “Some concerns,” with fewer RCTs categorized into “Low risk” or “High risk.” The judgments from both Cochrane Reviews and reviewers for the randomly selected RCTs and the consistency rate among them are presented in Table 1. The IRR of reviewers is 0.57, with detailed results available in the Multimedia Appendix 1.

Table 1. Risk-of-bias tool judgments.

RoB2[a]	Cochrane Reviews			Reviewers			IRR[b]	Consistency rate (%)
	Low risk	Some concern	High risk	Low risk	Some concern	High risk		
Domain 1	26	12	2	22	17	1	0.67	80
Domain 2 (assignment)	10	8	2	7	11	2	0.27	85
Domain 2 (adhering)	1	1	18	1	0	19	0.54	95
Domain 3	24	3	13	28	1	11	0.87	85
Domain 4	32	4	4	20	17	3	0.61	67.5
Domain 5	18	21	1	22	18	0	0.58	77.5
Overall (assignment)	4	12	4	2	15	3	0.57	85
Overall (adhering)	0	0	20	0	0	20	0.89	100

[a] RoB2: Risk-of-Bias tool.

[b] IRR: interrater reliability.

Accuracy of LLMs Compared With Reviewers and the Cochrane Reviewers in Domain

The complete judgments are summarized in Multimedia Appendix 1. Compared with Cochrane Reviews, the LLMs' judgments demonstrated accuracy rates of 57.5% and 70% for overall (assignment) and overall (adhering), respectively. When compared with reviewer judgments, the LLMs' accuracy rates for overall (assignment) and overall (adhering) were 65% and 70%, respectively. The average accuracy rates for the remaining 6 domains were 65.2% (95% CI 58.9-71.4) and 74.2% (95% CI 65.3-83.1) when compared with Cochrane Reviews and reviewers, respectively. In summary, except for Domain 2 (both assignment and adhering), the LLMs exhibited relatively good accuracy rates in comparison to both Cochrane Reviews and reviewer judgments (Table 2).

Table 2. Accuracy of large language models compared with reviewers and the Cochrane reviewers in domain.

RoB2[a]	Cochrane Reviews (%)	Reviewers (%)
Domain 1	75	85
Domain 2 (assignment)	57.5	65
Domain 2 (adhering)	50	55
Domain 3	70	83.8
Domain 4	67.5	76.3
Domain 5	73.8	93.8
Overall (assignment)	57.5	65
Overall (adhering)	70	70

[a] RoB2: Risk-of-bias tool.

Accuracy of LLMs Compared With Reviewers in Signaling Questions

Due to incomplete reporting of signaling question judgments in some Cochrane Reviews, our comparative analysis at the signaling question level was conducted exclusively against reviewer judgments. At the signaling question level, accuracy rates exceeded 70% for all questions except 2.4 (assignment), 2.5 (assignment), 3.3, and 3.4. The signaling questions 2.4 (assignment) and 2.5 (assignment) were excluded when calculating accuracy at the signaling question level, as they only had one available judgment after excluding the NA option. LLMs achieved an average accuracy of 83.2% (95% CI 77.5-88.9), excluding signaling questions 2.4 (assignment) and 2.5 (assignment; Multimedia Appendix 2). Given our previous findings of potential logical inconsistencies in LLM-generated judgments that deviated from prompt-specified methodologies, we conducted a further analysis. This investigation compared LLM-generated domain judgments with those derived from LLM-generated signaling question responses using the RoB2 decision algorithm. This secondary analysis revealed marked improvements in accuracy for Domain 2 (assignment) and overall (adhering) areas where LLM-generated domain judgments had initially demonstrated lower accuracy. Moderate accuracy improvements were also observed for Domain 2 (assignment), Domain 4, and overall (assignment). Other domains showed negligible differences (Figure 3 and Tables 3-4).

Figure 3. Cohen κ value and consistency between 2 large language model outputs.

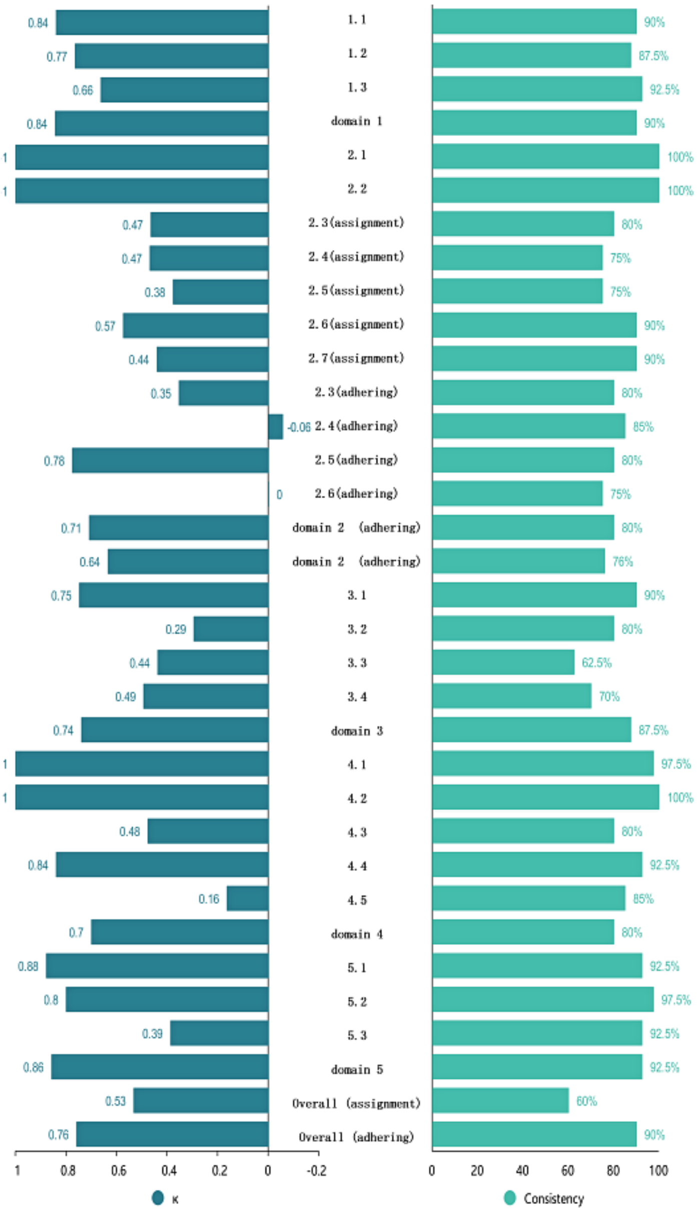


Table 3. Accuracy of large language models compared with reviewers and the Cochrane reviewers in domain.

RoB2[a]	Cochrane Reviews (%)	Reviewers (%)
Domain 1	73.8	85
Domain 2 (assignment)	67.5	67.5
Domain 2 (adhering)	77.5	95
Domain 3	67.5	83.8
Domain 4	67.5	85
Domain 5	72.5	93.6
Overall (assignment)	62.5	65
Overall (adhering)	90	90

[a] RoB2: Risk-of-bias tool.

Table 4. Accuracy of large language model–generated and algorithm-derived judgment compared with Cochrane reviews and reviewers for each domain.

Domain	Cochrane Reviews		Reviewers	
	LLM[a] generation (%)	Algorithm generation (%)	LLM generation (%)	Algorithm generation (%)
Domain 1	75	75	85	85
Domain 2 (assignment)	57.5	67.5	65	67.5
Domain 2 (adhering)	50	77.5	55	95
Domain 3	70	67.5	83.8	83.8
Domain 4	67.5	67.5	76.3	85
Overall (assignment)	65	65	57.5	62.5
Overall (adhering)	70	90	70	90

[a]LLM: large language model.

Stability

Due to the homogeneity of RCTs in some signaling questions (eg, 2.4 adhering), a significant difference between the calculated Cohen κ and the actual consistent assessment rate (Figure 3). Consequently, this study primarily uses the stability of LLM-generated results as the benchmark for assessing the stability of LLM outputs. The LLMs demonstrated significant consistency between its 2 output iterations, with an average consistent assessment rate of 85.2% (95% CI 85.15-88.79). A 100% consistent assessment rate was achieved for signaling questions 2.1, 2.2, and 4.2, while an additional 12 signaling questions or domains exhibited a consistent assessment rate exceeding 90%. However, the consistent assessment rate for signaling question 3.3 and the "Overall (assignment)" domain fell below 70%.

Efficiency

The mean assessment time of human reviewers was 31.5 (SD 12.9) minutes, while the LLMs completed assessments in 1.9 (SD 0.2) minutes. The mean difference in assessment time between human reviewers and the LLMs was 29.6 (95% CI 25.6-33.6) minutes, representing a 93.6% reduction from the average human assessment time.

Discussion

Principal Findings

In this study, we evaluated the accuracy of LLMs in assessing the RoB in RCTs with RoB2. To our knowledge, this is the first study using LLMs to assess RCTs through RoB2. Cochrane Reviews involve multiple research teams, each potentially applying different standards when judging specific signaling questions. This could lead to variations in the accuracy of LLM judgments. Therefore, we compared LLM-generated judgments with results from Cochrane Reviews and from 3 professional reviewers who assessed RCTs according to standardized criteria. This approach more accurately simulates the use of RoB2 and LLMs to assess the RoB of RCTs under consistent protocols.

We found differences between Cochrane systematic review judgments and reviewer judgments in Domains 1, 4, and 5. These differences are likely attributable to Cochrane Reviews assuming randomization and allocation concealment based on the authority of the RCT research team, such as: "States only that this was a randomized trial, without mentioning concealment. However, this was conducted by an experienced investigator and clinical trial network. These details were likely omitted for word‐count purposes." In domain 4, the main disagreement concerned patient-reported subjective outcomes: when outcome assessors were blinded but patients were not, some Cochrane Reviews selected "N" for signaling question 4.4, reasoning that assessors were blinded, whereas our approach dictated selecting "Y." Furthermore, to align with the LLMs' analytical capabilities for existing information, reviewers only assessed content reported in the main text, excluding Multimedia Appendices and protocols. This approach may have resulted in reviewers missing details reported only in the Multimedia Appendices and protocols.

LLMs achieved an average accuracy of 83.2% in signaling questions, excluding signaling questions 2.4 (allocation) and 2.5 (allocation). The primary issue in signaling question 1.1 stems from the LLMs' ability to extract randomization methods but made erroneous judgments. For instance, while correctly extracting the statement "The article does not describe specific methods for generating the random sequence, only stating 'The randomization list was generated by an independent psychologist using the RAND function of Microsoft Excel 2010'," LLMs incorrectly classified it as "NI." Signaling question 2.3 predominantly exhibited insufficient information extraction, rendering LLMs unable to properly evaluate existing evidence. The errors in signaling question 3.1 mainly originated from the misidentification of both the number of participants randomized and those completing the intervention, leading to incorrect calculations of data completeness percentages. This fundamental error subsequently propagated to inaccuracies in questions 3.3 and 3.4. For signaling question 4.3, the LLM frequently neglected the concept of blinding outcome assessors. When articles explicitly mentioned outcome assessors, LLMs misclassified patient-reported outcomes as assessor-evaluated outcomes, and misinterpreted assessor-evaluated outcomes as patient-reported ones, resulting in faulty judgments. The errors of deviating from the predetermined judgment method and insufficient extraction of corresponding information were prevalent across other signaling questions as well. For instance, in signaling question 2.6 (adhering), judgments of “Y” were observed when “N” should have been selected according to the intention-to-treat analysis protocol. We explored potential solutions, such as having LLMs judge only one signaling question per response, which improved accuracy. However, due to Claude access frequency limitations and the tedious nature of decomposing individual signaling questions, we did not apply this approach as it would compromise the goal of rapid RoB2 assessment using LLMs. In addition, importing PDFs through Claude’s built-in plugin, rather than converting them to Word format beforehand, resulted in incomplete conversion of image content to text, leading to some information loss.

Furthermore, errors in judging signaling questions 2.4 (assignment) and 2.5 (assignment) were caused by the incorrect judgment of 2.3 (assignment) as “N”, leading LLMs to generate “NA” and blank judgments for 2.4 and 2.5. This cascading effect of errors in prerequisite questions affecting subsequent signaling questions was also observed in other questions with applicable preconditions.

In internal validation, LLMs’ accuracy in domains was lower than on signaling questions. This is because the domain is influenced by multiple signaling questions, but there is not a strong correlation between accuracy in the domain and signaling questions within the domain. Discrepancies between “NI” and “Y” or “NI” and “N” were inconsistent for signaling questions, but they all indicated the same RoB in a domain. Moreover, LLMs sometimes deviated from the RoB2 decision algorithm when generating domain judgments, instead making independent assessments of overall domain RoB, with a tendency to generate “Some Concerns” judgments. Consequently, we attempted to form domain judgments using Excel based on the RoB2 decision algorithm applied to signaling question judgments, rather than relying on LLM-generated domain judgments. Results showed significant improvements in Domains 2 (adhering), 2 (assignment), and overall (adhering), which were the domains with lower accuracy in LLM-generated judgments.

Although the accuracy of domain judgments generated by LLM is relatively low, the 83.2% accuracy of signaling questions is sufficient to give researchers confidence in LLM’s capabilities. Previous studies and reviews have shown that there are certain differences in the results of human reviewers (IRR=0.16) [4]. LLMs accuracy of 83.2% represents a level of consistency that is comparable to or may exceed human-assessed reliability in certain situations. However, LLMs still carry a certain risk of error, and the current level of LLM performance is not yet sufficient to completely replace independent assessments by researchers. Instead, LLMs are better suited to complement the work of researchers, potentially replacing the need for a second researcher in back-to-back assessments, thereby supplementing the detection of unidentified bias risks. In addition, when using LLM for RoB2 evaluation, researchers should clearly indicate in the literature the specific LLM, its version, and the prompt strategies adopted to ensure reproducibility and transparency.

Limitations

This study has several limitations. First, due to the limited focus on adhering assessment in Cochrane Reviews and considering the number of eligible RCTs, we used a relatively small sample size. This may introduce certain biases, necessitating future large-scale studies to derive more trustworthy conclusions. In addition, due to this limitation, the research domain for adhering studies was restricted to COPD and asthma, which restricts the exploration of the bias of LLM in different medical domains. Second, the RoB of RCTs, which were classified as adhering, were high because of the judgments of Domain 2 (deviations from intended interventions), which precluded a thorough examination of the LLMs’ capability to assess low-risk studies of this domain. Third, current LLMs are unable to process extensive appendices in a single iteration. Consequently, our standardization process with human reviewers did not incorporate supplementary materials such as relevant registration documents and appendices. Our judgments were based solely on the full text of publications, with protocol availability as the only consideration for registration information, without further investigation into the completeness of the information. While this approach more precisely validates the LLMs’ judgment capabilities on available information, it slightly deviates from actual research practices. Fourth, potential biases within the LLM itself must be acknowledged, as these models may perpetuate or amplify biases present in their training data, potentially affecting their judgment of certain types of studies or methodologies.

Future Studies

The application research LLMs for RoB2 assessments began in early 2023, but initial efforts were limited by lower intelligence levels of earlier models and suboptimal prompting strategies, resulting in modest accuracy. With rapid advancements in LLM capabilities, our current study demonstrates significantly improved performance, which suggests that the accuracy rates reported in this study are likely to further increase as LLM technology continues to evolve. Future research should monitor these advancements to track improvements in RoB2 assessment accuracy. Interestingly, during our validation process, we observed that modern LLMs can independently access and apply RoB2 information to evaluate RCTs even without extensive prompting. Our prompt engineering, therefore focused on correcting specific error patterns rather than providing comprehensive guidance. This raises an intriguing research question: how accurately could advanced LLMs perform RoB2 assessments with minimal or no specific prompting? Future studies could explore this “zero-shot” or “few-shot” performance to better understand the evolving capabilities of these models. In addition, due to the limited number of Cochrane Reviews focusing on adhering, particularly in COPD and asthma interventions, the generalizability of our findings for adhering assessment remains constrained. Future research should expand to other medical domains to validate these results across a broader range of conditions and interventions. Furthermore, studies examining how LLMs adapt to evolving medical knowledge and comparing their performance with other automated tools could provide valuable insights into the robustness and practical utility of LLM-assisted RoB2 assessments in dynamic research environments. Longitudinal studies evaluating model performance over time could also help measure consistency and adaptation to evolving medical guidelines and practices.

Conclusions

In this study on applying LLMs to assess RoB2 in RCTs, we found that LLMs achieved commendable accuracy and consistency with a structured prompt, which particularly attributed to their ability to process complex methodological details and simulate assessors' decision-making processes via chain-of-thought reasoning. While the accuracy of the LLMs had not yet reached the level that can completely replace human assessment, their high accuracy demonstrated strong potential in assisting researchers with RoB2 assessments.

Acknowledgments

This study was jointly supported by the Fundamental Research Funds for the Central Universities (lzujbky-2024-oy11) and the Science and Technology Program Funding Projects of Gansu, China (23YFWA0004).

Data Availability

Data are supplied in supporting files available for download along with the published manuscript.

Authors' Contributions

JH was responsible for the original draft and data curation. JH and HL carried out the conceptualization. WZ, DX, CB, JL, JL, and BP performed the investigation. WZ, DX, CB, JL, JL, BP, and JH conducted formal analysis. JT, JH, HL, MS and LG developed the methodology. JT and LG provided supervision. LG handled the visualization. All authors contributed to the review and editing of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Additional data.
[DOCX File , 264 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Accuracy of large language models compared with reviewers in signaling question.
[PNG File , 143 KB-Multimedia Appendix 2]

References

1. Moosapour H, Saeidifard F, Aalaa M, Soltani A, Larijani B. The rationale behind systematic reviews in clinical medicine: a conceptual framework. J Diabetes Metab Disord. Aug 8, 2024;20(1):919-929. [doi: 10.1007/s40200-021-00773-8] [Medline: 34178868]
2. Guyatt G, Oxman AD, Akl EA, Kunz R, Vist G, Brozek J, et al. GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. J Clin Epidemiol. 2011;64(4):383-394. [doi: 10.1016/j.jclinepi.2010.04.026] [Medline: 21195583]
3. Higgins JPT, Altman DG, Gøtzsche PC, Jüni P, Moher D, Oxman AD, Cochrane Bias Methods Group, et al. Cochrane Statistical Methods Group. The cochrane collaboration's tool for assessing risk of bias in randomised trials. BMJ. 2011;343:d5928. [FREE Full text] [doi: 10.1136/bmj.d5928] [Medline: 22008217]
4. 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] [Medline: 32562833]
5. 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] [Medline: 34537386]
6. Minozzi S, Gonzalez-Lorenzo M, Cinquini M, Berardinelli D, Cagnazzo C, Ciardullo S, et al. University of Milan Post Graduate Course on Systematic Review Working Group. Adherence of systematic reviews to cochrane RoB2 guidance was frequently poor: a meta epidemiological study. J Clin Epidemiol. 2022;152:47-55. [FREE Full text] [doi: 10.1016/j.jclinepi.2022.09.003] [Medline: 36156301]
7. Ramkumar PN, Kunze KN, Haeberle HS, Karnuta JM, Luu BC, Nwachukwu BU, et al. Clinical and research medical applications of artificial intelligence. Arthroscopy. 2021;37(5):1694-1697. [FREE Full text] [doi: 10.1016/j.arthro.2020.08.009] [Medline: 32828936]

8. Dahmen J, Kayaalp ME, Ollivier M, Pareek A, Hirschmann MT, Karlsson J, et al. Artificial intelligence bot ChatGPT in medical research: the potential game changer as a double-edged sword. Knee Surg Sports Traumatol Arthrosc. 2023;31(4):1187-1189. [doi: 10.1007/s00167-023-07355-6] [Medline: 36809511]
9. Lai H, Ge L, Sun M, Pan B, Huang J, Hou L, et al. Assessing the Risk of bias in randomized clinical trials with large language models. JAMA Netw Open. 2024;7(5):e2412687. [FREE Full text] [doi: 10.1001/jamanetworkopen.2024.12687] [Medline: 38776081]
10. Lai H, Liu J, Bai C, Liu H, Pan B, Luo X, et al. ADVANCED Working Group. Language models for data extraction and risk of bias assessment in complementary medicine. NPJ Digit Med. 2025;8(1):74. [FREE Full text] [doi: 10.1038/s41746-025-01457-w] [Medline: 39890970]
11. Jackson DJ, Bacharier LB, Mauger DT, Boehmer S, Beigelman A, Chmiel JF, et al. National Heart, Lung,Blood Institute AsthmaNet. Quintupling inhaled glucocorticoids to prevent childhood asthma exacerbations. N Engl J Med. Mar 08, 2018;378(10):891-901. [FREE Full text] [doi: 10.1056/NEJMoa1710988] [Medline: 29504498]
12. Jonsdottir H, Amundadottir OR, Gudmundsson G, Halldorsdottir BS, Hrafnkelsson B, Ingadottir TS, et al. Effectiveness of a partnership-based self-management programme for patients with mild and moderate chronic obstructive pulmonary disease: a pragmatic randomized controlled trial. J Adv Nurs. Nov 2015;71(11):2634-2649. [doi: 10.1111/jan.12728] [Medline: 26193907]
13. Bourbeau J, Julien M, Maltais F, Rouleau M, Beaupré A, Bégin R, et al. Chronic Obstructive Pulmonary Disease axis of the Respiratory Network Fonds de la Recherche en Santé du Québec. Reduction of hospital utilization in patients with chronic obstructive pulmonary disease: a disease-specific self-management intervention. Arch Intern Med. Mar 10, 2003;163(5):585-591. [doi: 10.1001/archinte.163.5.585] [Medline: 12622605]
14. Ono T, Goto H, Sakai T, Nitta F, Mizuki N, Takase H, et al. Japan VKH Disease Treatment Study Group. Comparison of combination therapy of prednisolone and cyclosporine with corticosteroid pulse therapy in Vogt-Koyanagi-Harada disease. Jpn J Ophthalmol. Mar 2022;66(2):119-129. [doi: 10.1007/s10384-021-00878-w] [Medline: 34689288]
15. Aziminekoo E, Mohseni Salehi MS, Kalantari V, Shahrokh Tehraninejad E, Haghollahi F, Hossein Rashidi B, et al. Pregnancy outcome after blastocyst stage transfer comparing to early cleavage stage embryo transfer. Gynecol Endocrinol. 2015;31(11):880-884. [FREE Full text] [doi: 10.3109/09513590.2015.1056141] [Medline: 26437606]
16. Collin C, Davies P, Mutiboko IK, Ratcliffe S, Sativex Spasticity in MS Study Group. Randomized controlled trial of cannabis-based medicine in spasticity caused by multiple sclerosis. Eur J Neurol. Mar 2007;14(3):290-296. [doi: 10.1111/j.1468-1331.2006.01639.x] [Medline: 17355549]
17. Pressler A, Christle JW, Lechner B, Grabs V, Haller B, Hettich I, et al. Exercise training improves exercise capacity and quality of life after transcatheter aortic valve implantation: a randomized pilot trial. Am Heart J. 2016;182:44-53. [doi: 10.1016/j.ahj.2016.08.007] [Medline: 27914499]
18. Ireland MJ, Clough B, Gill K, Langan F, O'Connor A, Spencer L. A randomized controlled trial of mindfulness to reduce stress and burnout among intern medical practitioners. Med Teach. 2017;39(4):409-414. [doi: 10.1080/0142159X.2017.1294749] [Medline: 28379084]
19. Beigel JH, Tomashek KM, Dodd LE, Mehta AK, Zingman BS, Kalil AC, et al. ACTT-1 Study Group Members. Remdesivir for the treatment of Covid-19 - final report. N Engl J Med. 2020;383(19):1813-1826. [FREE Full text] [doi: 10.1056/NEJMoa2007764] [Medline: 32445440]
20. Hu FW, Huang YT, Lin HS, Chen CH, Chen MJ, Chang CM. Effectiveness of a simplified reablement program to minimize functional decline in hospitalized older patients. Geriatr Gerontol Int. 2020;20(5):436-442. [doi: 10.1111/ggi.13891] [Medline: 32102119]
21. Novotna A, Mares J, Ratcliffe S, Novakova I, Vachova M, Zapletalova O, et al. Sativex Spasticity Study Group. A randomized, double-blind, placebo-controlled, parallel-group, enriched-design study of nabiximols* (Sativex(®)), as add-on therapy, in subjects with refractory spasticity caused by multiple sclerosis. Eur J Neurol. 2011;18(9):1122-1131. [doi: 10.1111/j.1468-1331.2010.03328.x] [Medline: 21362108]
22. Bart BA, Boyle A, Bank AJ, Anand I, Olivari MT, Kraemer M, et al. Ultrafiltration versus usual care for hospitalized patients with heart failure: the relief for acutely fluid-overloaded patients with decompensated congestive heart failure (RAPID-CHF) trial. J Am Coll Cardiol. 2005;46(11):2043-2046. [FREE Full text] [doi: 10.1016/j.jacc.2005.05.098] [Medline: 16325039]
23. Pinto-Fraga J, López-Miguel A, González-García MJ, Fernández I, López-de-la-Rosa A, Enríquez-de-Salamanca A, et al. Topical Fluorometholone Protects the Ocular Surface of Dry Eye Patients from Desiccating Stress: A Randomized Controlled Clinical Trial. Ophthalmology. 2016;123(1):141-153. [doi: 10.1016/j.ophtha.2015.09.029] [Medline: 26520171]
24. Dissanayake E, Tani Y, Nagai K, Sahara M, Mitsuishi C, Togawa Y, et al. Skin care and synbiotics for prevention of atopic dermatitis or food allergy in newborn infants: A 2 × 2 factorial, randomized, non-treatment controlled trial. Int Arch Allergy Immunol. 2019;180(3):202-211. [doi: 10.1159/000501636] [Medline: 31394530]
25. Frattarelli JL, Leondires MP, McKeeby JL, Miller BT, Segars JH. Blastocyst transfer decreases multiple pregnancy rates in in vitro fertilization cycles: a randomized controlled trial. Fertil Steril. 2003;79(1):228-230. [FREE Full text] [doi: 10.1016/s0015-0282(02)04558-2] [Medline: 12524098]

26. Griffiths H, Duffy F, Duffy L, Brown S, Hockaday H, Eliasson E, et al. Efficacy of mentalization-based group therapy for adolescents: the results of a pilot randomised controlled trial. BMC Psychiatry. 2019;19(1):167. [FREE Full text] [doi: 10.1186/s12888-019-2158-8] [Medline: 31170947]
27. RECOVERY Collaborative Group, Horby P, Lim WS, Emberson JR, Mafham M, Bell JL, et al. Dexamethasone in hospitalized patients with Covid-19. N Engl J Med. 2021;384(8):693-704. [FREE Full text] [doi: 10.1056/NEJMoa2021436] [Medline: 32678530]
28. Beautrais AL, Gibb SJ, Faulkner A, Fergusson DM, Mulder RT. Postcard intervention for repeat self-harm: randomised controlled trial. Br J Psychiatry. 2010;197(1):55-60. [doi: 10.1192/bjp.bp.109.075754] [Medline: 20592434]
29. Huhn M, Leucht C, Rothe P, Dold M, Heres S, Bornschein S, et al. Reducing antipsychotic drugs in stable patients with chronic schizophrenia or schizoaffective disorder: a randomized controlled pilot trial. Eur Arch Psychiatry Clin Neurosci. 2021;271(2):293-302. [FREE Full text] [doi: 10.1007/s00406-020-01109-y] [Medline: 32062728]
30. Ooi DSQ, Ling JQR, Sadananthan SA, Velan SS, Ong FY, Khoo CM, et al. Branched-chain amino acid supplementation does not preserve lean mass or affect metabolic profile in adults with overweight or obesity in a randomized controlled weight loss intervention. J Nutr. 2021;151(4):911-920. [FREE Full text] [doi: 10.1093/jn/nxaa414] [Medline: 33537760]
31. Kawanishi C, Aruga T, Ishizuka N, Yonemoto N, Otsuka K, Kamijo Y, et al. ACTION-J Group. Assertive case management versus enhanced usual care for people with mental health problems who had attempted suicide and were admitted to hospital emergency departments in Japan (ACTION-J): a multicentre, randomised controlled trial. Lancet Psychiatry. 2014;1(3):193-201. [doi: 10.1016/S2215-0366(14)70259-7] [Medline: 26360731]
32. Liljeroos M, Ågren S, Jaarsma T, Årestedt K, Strömberg A. Long term follow-Up after a randomized integrated educational and psychosocial intervention in patient-partner dyads affected by heart failure. PLoS One. 2015;10(9):e0138058. [FREE Full text] [doi: 10.1371/journal.pone.0138058] [Medline: 26406475]
33. Botella-Carretero JI, Iglesias B, Balsa JA, Zamarrón I, Arrieta F, Vázquez C. Effects of oral nutritional supplements in normally nourished or mildly undernourished geriatric patients after surgery for hip fracture: a randomized clinical trial. JPEN J Parenter Enteral Nutr. 2008;32(2):120-128. [doi: 10.1177/0148607108314760] [Medline: 18407904]
34. Tauber S, Cupp G, Garber R, Bartell J, Vohra F, Stroman D. Microbiological efficacy of a new ophthalmic formulation of moxifloxacin dosed twice-daily for bacterial conjunctivitis. Adv Ther. 2011;28(7):566-574. [doi: 10.1007/s12325-011-0037-x] [Medline: 21681652]
35. González-Ortega I, Vega P, Echeburúa E, Alberich S, Fernández-Sevillano J, Barbeito S, et al. A multicentre, randomised, controlled trial of a combined clinical treatment for first-episode psychosis. Int J Environ Res Public Health. 2021;18(14):7239. [FREE Full text] [doi: 10.3390/ijerph18147239] [Medline: 34299697]
36. van Anholt RD, Sobotka L, Meijer EP, Heyman H, Groen HW, Topinková E, et al. Specific nutritional support accelerates pressure ulcer healing and reduces wound care intensity in non-malnourished patients. Nutrition. 2010;26(9):867-872. [doi: 10.1016/j.nut.2010.05.009] [Medline: 20598855]
37. Bischoff EWMA, Akkermans R, Bourbeau J, van Weel C, Vercoulen JH, Schermer TRJ. Comprehensive self management and routine monitoring in chronic obstructive pulmonary disease patients in general practice: randomised controlled trial. BMJ. 2012;345:e7642. [FREE Full text] [doi: 10.1136/bmj.e7642] [Medline: 23190905]
38. Coultas D, Frederick J, Barnett B, Singh G, Wludyka P. A randomized trial of two types of nurse-assisted home care for patients with COPD. Chest. 2005;128(4):2017-2024. [doi: 10.1378/chest.128.4.2017] [Medline: 16236850]
39. Curran M, Tierney AC, Collins L, Kennedy L, McDonnell C, Jurascheck AJ, et al. Steps ahead: optimising physical activity in adults with cystic fibrosis: a pilot randomised trial using wearable technology, goal setting and text message feedback. J Cyst Fibros. 2023;22(3):570-576. [FREE Full text] [doi: 10.1016/j.jcf.2022.11.002] [Medline: 36402730]
40. Fan VS, Gaziano JM, Lew R, Bourbeau J, Adams SG, Leatherman S, et al. A comprehensive care management program to prevent chronic obstructive pulmonary disease hospitalizations: a randomized, controlled trial. Ann Intern Med. 2012;156(10):673-683. [FREE Full text] [doi: 10.7326/0003-4819-156-10-201205150-00003] [Medline: 22586006]
41. FitzGerald JM, Becker A, Sears MR, Mink S, Chung K, Lee J, et al. Canadian Asthma Exacerbation Study Group. Doubling the dose of budesonide versus maintenance treatment in asthma exacerbations. Thorax. 2004;59(7):550-556. [FREE Full text] [doi: 10.1136/thx.2003.014936] [Medline: 15223858]
42. Harrison TW, Oborne J, Newton S, Tattersfield AE. Doubling the dose of inhaled corticosteroid to prevent asthma exacerbations: randomised controlled trial. Lancet. 2004;363(9405):271-275. [doi: 10.1016/s0140-6736(03)15384-6] [Medline: 14751699]
43. Hernández C, Alonso A, Garcia-Aymerich J, Serra I, Marti D, Rodriguez-Roisin R, et al. NEXES consortium. Effectiveness of community-based integrated care in frail COPD patients: a randomised controlled trial. NPJ Prim Care Respir Med. 2015;25:15022. [FREE Full text] [doi: 10.1038/npjpcrm.2015.22] [Medline: 25856791]
44. Johnson-Warrington V, Rees K, Gelder C, Morgan MD, Singh SJ. Can a supported self-management program for COPD upon hospital discharge reduce readmissions? A randomized controlled trial. Int J Chron Obstruct Pulmon Dis. 2016;11:1161-1169. [FREE Full text] [doi: 10.2147/COPD.S91253] [Medline: 27330284]
45. Jorm AF, Kitchener BA, Fischer JA, Cvetkovski S. Mental health first aid training by e-learning: a randomized controlled trial. Aust N Z J Psychiatry. 2010;44(12):1072-1081. [doi: 10.3109/00048674.2010.516426] [Medline: 21070103]

46. Kessler R, Casan-Clara P, Koehler D, Tognella S, Viejo JL, Dal Negro RW, et al. COMET: a multicomponent home-based disease-management programme routine care in severe COPD. Eur Respir J. 2018;51(1):1701612. [FREE Full text] [doi: 10.1183/13993003.01612-2017] [Medline: 29326333]
47. Lenferink A, van der Palen J, van der Valk PDLPM, Cafarella P, van Veen A, Quinn S, et al. Exacerbation action plans for patients with COPD and comorbidities: a randomised controlled trial. Eur Respir J. 2019;54(5):1802134. [FREE Full text] [doi: 10.1183/13993003.02134-2018] [Medline: 31413163]
48. Liang J, Abramson MJ, Russell G, Holland AE, Zwar NA, Bonevski B, et al. Interdisciplinary COPD intervention in primary care: a cluster randomised controlled trial. Eur Respir J. 2019;53(4):1801530. [FREE Full text] [doi: 10.1183/13993003.01530-2018] [Medline: 30792342]
49. Mitchell KE, Johnson-Warrington V, Apps LD, Bankart J, Sewell L, Williams JE, et al. A self-management programme for COPD: a randomised controlled trial. Eur Respir J. 2014;44(6):1538-1547. [FREE Full text] [doi: 10.1183/09031936.00047814] [Medline: 25186259]
50. Oborne J, Mortimer K, Hubbard RB, Tattersfield AE, Harrison TW. Quadrupling the dose of inhaled corticosteroid to prevent asthma exacerbations: a randomized, double-blind, placebo-controlled, parallel-group clinical trial. Am J Respir Crit Care Med. 2009;180(7):598-602. [doi: 10.1164/rccm.200904-0616OC] [Medline: 19590019]
51. Rice-McDonald G, Bowler S, Staines G, Mitchell C. Doubling daily inhaled corticosteroid dose is ineffective in mild to moderately severe attacks of asthma in adults. Intern Med J. 2005;35(12):693-698. [doi: 10.1111/j.1445-5994.2005.00972.x] [Medline: 16313543]
52. Sánchez-Nieto JM, Andújar-Espinosa R, Bernabeu-Mora R, Hu C, Gálvez-Martínez B, Carrillo-Alcaraz A, et al. Efficacy of a self-management plan in exacerbations for patients with advanced COPD. Int J Chron Obstruct Pulmon Dis. 2016;11:1939-1947. [FREE Full text] [doi: 10.2147/COPD.S104728] [Medline: 27574418]
53. Titova E, Steinshamn S, Indredavik B, Henriksen AH. Long term effects of an integrated care intervention on hospital utilization in patients with severe COPD: a single centre controlled study. Respir Res. 2015;16(1):8. [FREE Full text] [doi: 10.1186/s12931-015-0170-1] [Medline: 25645122]
54. Walters J, Cameron-Tucker H, Wills K, Schüz N, Scott J, Robinson A, et al. Effects of telephone health mentoring in community-recruited chronic obstructive pulmonary disease on self-management capacity, quality of life and psychological morbidity: a randomised controlled trial. BMJ Open. 2013;3(9):e003097. [FREE Full text] [doi: 10.1136/bmjopen-2013-003097] [Medline: 24014482]
55. Wang LH, Zhao Y, Chen LY, Zhang L, Zhang YM. The effect of a nurse-led self-management program on outcomes of patients with chronic obstructive pulmonary disease. Clin Respir J. 2020;14(2):148-157. [doi: 10.1111/crj.13112] [Medline: 31769181]
56. Rice KL, Dewan N, Bloomfield HE, Grill J, Schult TM, Nelson DB, et al. Disease management program for chronic obstructive pulmonary disease: a randomized controlled trial. Am J Respir Crit Care Med. 2010;182(7):890-896. [doi: 10.1164/rccm.200910-1579OC] [Medline: 20075385]

Abbreviations

COPD: chronic obstructive pulmonary disease
IRR: interrater reliability
LLM: large language model
RCT: randomized controlled trial
RoB2: Risk-of-Bias tool
SR: systematic review

Edited by J Sarvestan, T Leung; submitted 22.12.24; peer-reviewed by X Wang, GK Gupta, M Veliyathnadu Ebrahim, K Varikara; comments to author 26.02.25; revised version received 08.04.25; accepted 05.05.25; published 24.06.25

Please cite as:
Huang J, Lai H, Zhao W, Xia D, Bai C, Sun M, Liu J, Liu J, Pan B, Tian J, Ge L
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
URL: https://www.jmir.org/2025/1/e70450
doi: 10.2196/70450
PMID: 40554779