




Towards the automatic risk of bias assessment on randomized controlled trials: A comparison of RobotReviewer and humans

Yuan Tian¹ | Xi Yang² | Suhail A. Doi³ | Luis Furuya-Kanamori⁴  | Lifeng Lin⁵  | Joey S. W. Kwong⁶ | Chang Xu⁷ 

¹First School of Clinical Medicine, Anhui Medical University, Hefei, China

²Department of Maternal, Child and Adolescent Health, School of Public Health, Anhui Medical University, Hefei, China

³Department of Population Medicine, College of Medicine, QU Health, Qatar University, Doha, Qatar

⁴UQ Centre for Clinical Research, The University of Queensland, Herston, Queensland, Australia

⁵Department of Epidemiology and Biostatistics, University of Arizona, Tucson, Arizona, USA

⁶Global Health Nursing, Graduate School of Nursing Science, St Luke's International University, Tokyo, Japan

⁷Proof of Concept Center, Eastern Hepatobiliary Surgery Hospital, Third Affiliated Hospital, Second Military Medical University, Naval Medical University, Shanghai, China

Correspondence

Chang Xu, Proof of Concept Center, Eastern Hepatobiliary Surgery Hospital, Third Affiliated Hospital, Second Military Medical University, Naval Medical University, Shanghai, China.
Email: xuchang2016@runbox.com

Funding information

Shanghai Eastern Hepatobiliary Surgery Hospital of Naval Medical University, Grant/Award Number: TF2024YZRH03; National Natural Science Foundation of China, Grant/Award Number: 72204003; Qatar National Research Fund, Grant/Award Number: NPRP-BSRA01-0406-210030

Abstract

RobotReviewer is a tool for automatically assessing the risk of bias in randomized controlled trials, but there is limited evidence of its reliability. We evaluated the agreement between RobotReviewer and humans regarding the risk of bias assessment based on 1955 randomized controlled trials. The risk of bias in these trials was assessed via two different approaches: (1) manually by human reviewers, and (2) automatically by the RobotReviewer. The manual assessment was based on two groups independently, with two additional rounds of verification. The agreement between RobotReviewer and humans was measured via the concordance rate and Cohen's kappa statistics, based on the comparison of binary classification of the risk of bias (low vs. high/unclear) as restricted by RobotReviewer. The concordance rates varied by domain, ranging from 63.07% to 83.32%. Cohen's kappa statistics showed a poor agreement between humans and RobotReviewer for allocation concealment ($\kappa = 0.25$, 95% CI: 0.21–0.30), blinding of outcome assessors ($\kappa = 0.27$, 95% CI: 0.23–0.31); While moderate for random sequence generation ($\kappa = 0.46$, 95% CI: 0.41–0.50) and blinding of participants and personnel ($\kappa = 0.59$, 95% CI: 0.55–0.64). The findings demonstrate that there were domain-specific differences in the level of agreement between RobotReviewer and humans. We suggest that it might be a useful auxiliary tool, but the specific manner of its integration as a complementary tool requires further discussion.

KEYWORDS

automation, rapid review, risk of bias, RobotReviewer

Highlights

What is already known?

- RobotReviewer represents a novel solution designed to automatically evaluate the risk of bias in randomized controlled trials. Despite its potential, the existing body of evidence regarding its reliability remains limited.

What is new?

- This is the largest validation study to date investigating the agreement of RobotReviewer compared to the consensus of human reviewers.

Potential impact for *Research Synthesis Methods* readers

- The findings confirm previous findings of domain-specific differences in the level of agreement between RobotReviewer and humans. In addition the tool has moderate agreement with humans. It might be a useful auxiliary tool in future practice, but the specific manner of its integration as a complementary tool requires further discussion and thought.

1 | INTRODUCTION

Evidence synthesis is a robust method for evaluating the effects of healthcare interventions, especially when applying evidence synthesis to randomized controlled trials (RCTs). In evidence synthesis, the assessment of risk of bias (RoB) is an important step that ensures reviewers consider all potential limitations of included trials and know the extent to which biases affect the results of each trial.¹ However, methodological rigorous procedures makes evidence synthesis time-consuming; it is estimated that the assessment of RoB for each RCT takes researchers an average of 10–60 min to complete,² and the whole process of evidence synthesis research often takes 0.5–2 years.³ This makes current evidence synthesis inflexible to meet any urgent need for emergent public health problems, and can pose a threat to effective clinical decision-making.

To expedite the review process, researchers have integrated artificial intelligence into medical research and developed a range of automated review tools.⁴ RobotReviewer, a free online RoB assessment tool, is one of the tools that aims to reduce the time needed to assess RoB. It was developed by a team of researchers and leverages machine learning for classification and information extraction.⁵ The authors annotated the PDFs of 12,808 trials from the Cochrane Database of Systematic Reviews (CDSR) to train a multi-task machine learning model.⁵ By harnessing the power of machine learning and leveraging a vast dataset, RobotReviewer aims to offer a valuable and accessible tool for RoB assessment.

The key question is whether artificial intelligence instruments are mature enough to be used to deliver

acceptable concordance with humans. While several studies have suggested that human reviewers should check and validate results from RobotReviewer,^{6,7} there remains significant skepticism regarding its reliability for broader implementation, especially considering that its validation has largely been internal within the confines of the CDSR.⁸ It is therefore a priority to evaluate its reliability compared to that of human reviewers. We now compare the concordance between RobotReviewer and humans using a large empirical dataset (SMART Safety),⁹ serving as an external validation for the further application of such automatic instruments in evidence synthesis.

2 | METHODS

2.1 | Data source

This study utilized a subset of the data from a previous study.¹⁰ In summary, we conducted a PubMed search for systematic reviews of adverse events published between January 1, 2015 and January 1, 2020. We included systematic reviews of RCTs focusing on healthcare interventions with adverse events as the exclusive outcome. The term adverse event is defined as “*any untoward medical occurrence in a patient or subject in clinical practice,*” encompassing side effects, adverse effects, adverse reactions, harm, or complications associated with any healthcare intervention.¹¹ The representativeness of the search has been well-confirmed, with its sensitivity ranging between 93.85% and 99.30%.¹² The complete search strategy is detailed in the Supplementary File S1.

Two reviewers (X.Q., C.X.) independently screened the titles and abstracts as well as the full texts of the records using the Rayyan online tool.¹³ In step one, records were excluded only if both reviewers agreed on their exclusion. The remaining records were then subjected to screening again in step two. Any conflicts were resolved through discussion between the two authors. More detailed inclusion and exclusion criteria have been described in the supplementary file (see protocol in Supplementary File S1).

2.2 | Risk of bias assessment: Human reviewers

In order to be consistent with the rating system of RobotReviewer, we used the Cochrane Collaboration's tool (an updated version of RoB 1.0 in 2011¹⁴) for assessing risk of bias for all included trials, including seven domains: (1) random sequence generation, (2) allocation concealment, (3) blinding of participants and personnel, (4) blinding of outcome assessor, (5) incomplete outcome data, (6) selective reporting, and (7) other bias. However, we focused our evaluation on the first four domains due to their relatively objective nature, which also aligned with RobotReviewer.¹⁵ Five individuals with a background in evidence-based medicine conducted the risk of bias assessment (X.Y., R.Z., T.Q., F.Y., Y.Y.). Since the assessment of the risk of bias is somewhat subjective, we completed three rounds of checks to ensure the objectivity of the evaluation results. The first round involved two independent review groups (Group 1: F.Y., T.Q., Y.Y.; Group 2: X.Y.), followed by a third-party (R.Z.) comparison and joint discussions to resolve conflicts until a consensus. Subsequently, the second and third verification rounds were conducted by X.Y. and R.Z.

We followed the Cochrane Collaboration guidelines to perform risk of bias assessments. We used response options aligned with the updated version of RoB 1.0,¹⁴ with categories labeled “Low,” “Unclear,” and “High.” We confirmed the sentence annotations in full-text reviewing and labeled the risk of bias in each domain specifically, such that a tag of “Low” depended on a clear description of procedures to reduce the risk of bias (e.g., randomly assigned via an interactive web system). As for unclear signaling questions, we labeled them as “Unclear,” and the remaining were labeled as “High” if the original articles mentioned that they did not implement this safeguard or this implementation was not reported. We combined “Unclear” and “High” where reviewers selected “Unclear” or “High,” in accordance with assessments from RobotReviewer. While the systematic reviews, as well as the individual studies may have looked at multiple outcomes, each study implements

(or not) these four safeguards (i.e., concealment allocation) independently of the outcomes. Even for blinding of outcome assessors, this is not commonly outcome specific in individual studies and therefore there was no imperative for outcome specific assessments by both RobotReviewer and the human reviewers. As such, there was a study specific assessment only.

2.3 | Risk of bias assessment: RobotReviewer

The RobotReviewer is a free online tool that uses a machine-learning algorithm to automatically evaluate the risk of bias of RCTs.⁵ The algorithm takes a full-text article describing the conduct and findings of a randomized controlled trial as input and generates a binary conclusion indicating whether the study is at low or high/unclear risk. It only accepts PDFs due to the machine learning procedures clinging to the annotations on PDFs. The RobotReviewer allows users to “drag & drop” a document file into the proper spot on the user interface, and the tool will assess the risk of bias of the relevant domain for the uploaded file automatically. The outputs of the tool, including the risk of bias information for each trial, are transferred to a Microsoft Excel worksheet. Since the RobotReviewer did not provide the overall risk bias, the performance between RobotReviewer and human reviewers was compared based on specific domains.

2.4 | Outcomes

Two primary outcomes were pre-determined: (1) The level of agreement between the human reviewers and RobotReviewer for the assessment of risk of bias; (2) Positive percent agreement and negative percent agreement were used to ascertain the capacity of RobotReviewer in identifying risks categorized as “low” and “unclear/high.” Secondary outcomes included the yearly concordance between human reviewers and RobotReviewer.

2.5 | Data analysis

We utilized descriptive statistics (counts and percentages) to summarize the risk of bias information on a safeguard-specific basis. The Cohen's Kappa statistic was used to measure the performance of RobotReviewer.¹⁶ The Kappa statistics measure the level of agreement in six categories from poor to perfect: poor ($\kappa < 0$), slight ($\kappa = 0.0-0.20$), fair ($\kappa = 0.21-0.40$), moderate ($\kappa = 0.41-0.60$), substantial ($\kappa = 0.61-0.80$), and almost perfect ($\kappa = 0.81-1.00$).¹⁶

Positive percent agreement (PPA) and negative percent agreement (NPA) of RobotReviewer compared to human reviewers (as an imperfect reference standard) was computed, which was calculated via true positive rate (TP), true negative rate (TN), false negative rate (FN), and false positive rate (FP). Low risk of bias was deemed “positive” and high/unclear risk of bias was deemed “negative.” The PPA reflects the ability to correctly identify domains as low risk of bias, and the NPA similarly reflects the ability to correctly identify domains as high/unclear risk of bias. We considered true positive or true negative only when RobotReviewer concurred with the humans. False negative refers to the case where RobotReviewer gave a ‘high/unclear risk’ judgment, while human reviewers gave a ‘low risk’ judgment. False positive refers to the case that RobotReviewer gave a ‘low risk’ judgment, while human reviewers gave a ‘high/unclear risk’ judgment. The PPA and NPA were calculated using the following formulas:

$$PPA = \frac{TP}{TP + FN}; NPA = \frac{TN}{TN + FP}.$$

We conducted a comprehensive analysis of the yearly concordance rate (TP + TN) between RobotReviewer and human reviewers. The motivation behind this examination was to assess whether there has been any improvement in the reported performance of RobotReviewer over time.

Sensitivity analysis was conducted by deleting trials with the risk of bias assessment based on Supplementary Materials S1. This procedure ensures better comparability by maintaining consistency in the sources of assessed materials. Due to the nature of subjective health outcomes, inadequate randomization and blinding can lead to bias. Given the potential impact of different outcome types on the assessment of the domain of blinding for outcome assessors, a post-hoc sensitivity analysis was conducted by randomized controlled trials with subjective outcomes recorded in our dataset (e.g., fatigue).

All data analyses were run via Stata/SE 16.0 (Stata Corp LCC, College Station, TX) and Microsoft Office Excel (version 2021, Microsoft Corporation, Redmond, Washington), with $\alpha = 0.05$ as the significance level for a two-sided test.

3 | RESULTS

Overall, 18,636 records from PubMed were identified through the primary search. After removing 1967 duplicates and 15,339 by titles and abstracts, 1330 records remained to be reviewed for eligibility via full-texts. Among these, 151 systematic reviews with 629 meta-

analyses encompassing 2305 trials were included. The list of the included and excluded reviews can be found in the supplementary file (Table S1). After removing 250 duplicate trials and 100 trials that RobotReviewer was unable to assess, 1955 trials were included in the analysis (Figure 1).

3.1 | RoB Assessment: Concordance and agreement between RobotReviewer and human reviewers

Table 1 presents the results of risk of bias assessment for RobotReviewer and human reviewers. For the domain-specific risk of bias. Low risk of bias was identified in 48.5% of the trials by human reviewers while 59.9% by RobotReviewer ($p < 0.01$) for random sequence generation, 63.3% (1237/1955) versus 52.3% (1023/1955) for allocation concealment ($p < 0.01$), 74.5% (1456/1955) versus 68.1% (1332/1955) for blinding of participants and personnel ($p < 0.01$), and 28.8% (563/1955) versus 38.5% (752/1955) for blinding of outcome assessors ($p < 0.01$). The details of the assessments made by human reviewers and by RobotReviewer are presented in Table S2 (supplementary file).

The concordance rate was 63.1% for random sequence generation, 83.3% for allocation concealment, 67.0% for blinding of participants and personnel, and 77.1% for blinding of outcome assessors.

Figure 2 illustrates the concordance rate by year between RobotReviewer and human reviewers. The included RCTs spanned from 1971 to 2020 in terms of publication years. Generally, the concordance of the assessments between RobotReviewer and human reviewers did not suggest improvement over time. For the domain-specific risk of bias, the concordance rate ranged from 59.3% to 95.0% for random sequence generation, from 40.0% to 74.7% for allocation concealment, from 68.7% to 90.6% for blinding of participants and personnel, and from 55.4% to 75.0% for blinding of outcome assessors.

Cohen's kappa showed varying levels of agreement: Fair for allocation concealment ($\kappa = 0.25$, 95% CI: 0.21–0.30) and blinding of outcome assessors ($\kappa = 0.27$, 95% CI: 0.23–0.31); Moderate for random sequence generation ($\kappa = 0.46$, 95% CI: 0.41–0.50) and blinding of participants and personnel ($\kappa = 0.59$, 95% CI: 0.55–0.64).

3.2 | RobotReviewer performance: PPA and NPA

RobotReviewer demonstrated a high PPA for random sequence generation (PPA = 0.84, 95%CI: 0.81–0.86), however, the NPA was much lower (NPA = 0.62, 95%CI: 0.59–0.65). For blinding of participants and personnel,

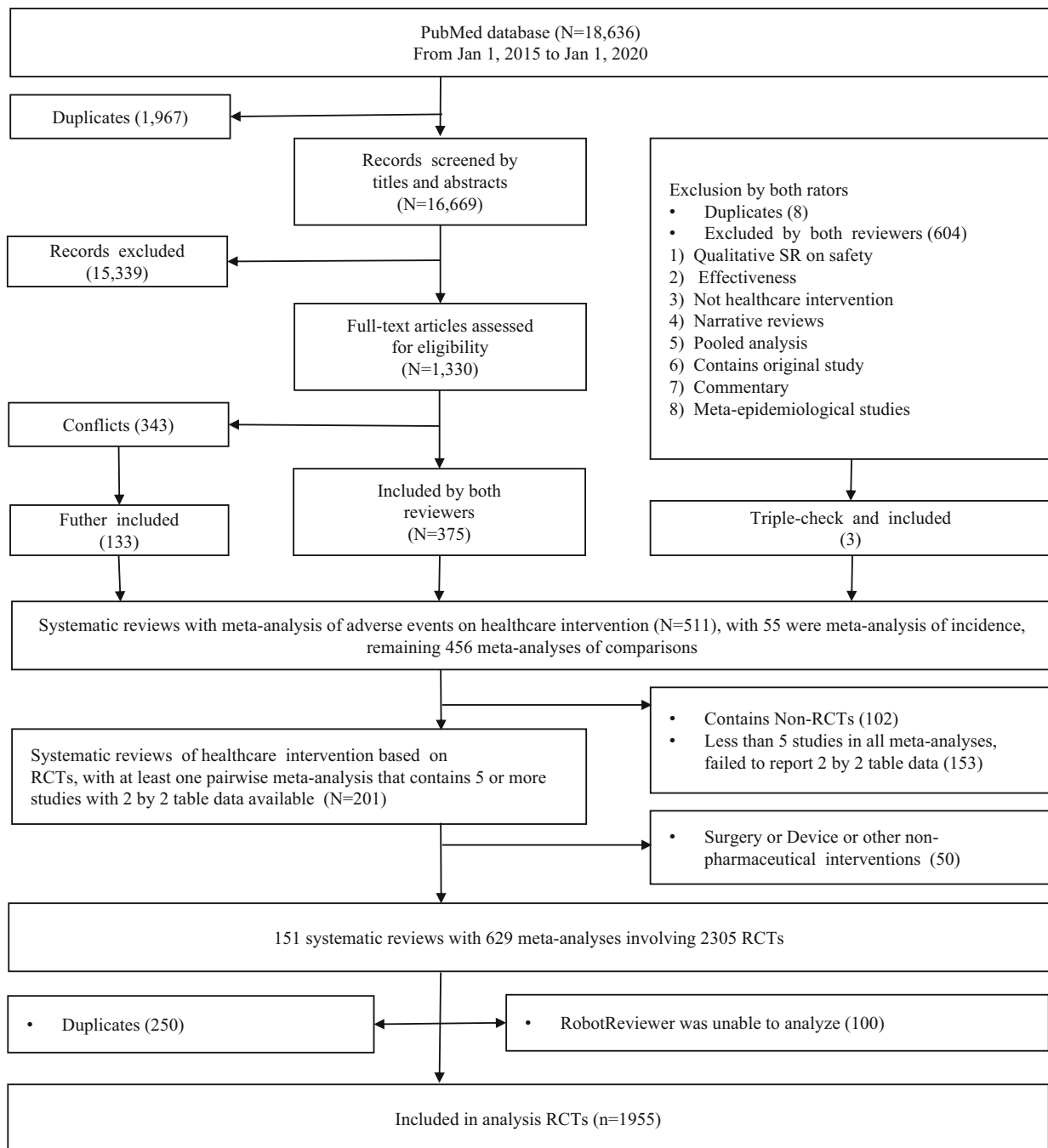


FIGURE 1 Flow diagram of the literature screening.

there was a high PPA (PPA = 0.85, 95%CI: 0.83–0.86) and high NPA (NPA = 0.80, 95%CI: 0.76–0.83). The PPA in the domain of blinding of outcome assessors was lower (PPA = 0.60, 95%CI: 0.55–0.64) while NPA was slightly better (NPA = 0.70, 95%CI: 0.68–0.72). There was lower PPA (PPA = 0.62, 95%CI: 0.59–0.65) and NPA (NPA = 0.65, 95%CI: 0.61–0.68) in the area of allocation concealment. Details are presented in Table 1.

3.3 | Sensitivity analysis

Sensitivity analysis was performed by excluding RCTs with the risk of bias assessed via supplementary materials by human reviewers, and the results remained robust ($N = 1948$); See Table 2. Our post-hoc sensitivity analysis for RCTs with subjective outcomes also showed robust results on Table S3. (supplementary file).

TABLE 1 Agreement between RobotReviewer and human reviewers.

RoB domain	Human reviewers	RobotReviewer	Concordance (%)	Kappa (95%CI)	TP (%)	TN (%)	FP (%)	FN (%)	PPA (95%CI)	NPA (95%CI)
Random sequence generation	948 (48.5%)	1171 (59.9%)	72.7	0.46 (0.41–0.50)	40.6	32.2	19.3	7.9	0.84 (0.81–0.86)	0.62 (0.59–0.65)
Allocation concealment	1237 (63.3%)	1023 (52.3%)	63.1	0.25 (0.21–0.30)	39.3	23.7	13.0	23.9	0.62 (0.59–0.65)	0.65 (0.61–0.68)
Blinding of participants and personnel	1456 (74.5%)	1332 (68.1%)	83.3	0.59 (0.55–0.64)	63.0	20.4	5.2	11.5	0.85 (0.83–0.86)	0.80 (0.76–0.83)
Blinding of outcome assessors	563 (28.8%)	752 (38.5%)	67.0	0.27 (0.23–0.31)	17.1	49.9	21.3	11.7	0.60 (0.55–0.64)	0.70 (0.68–0.72)

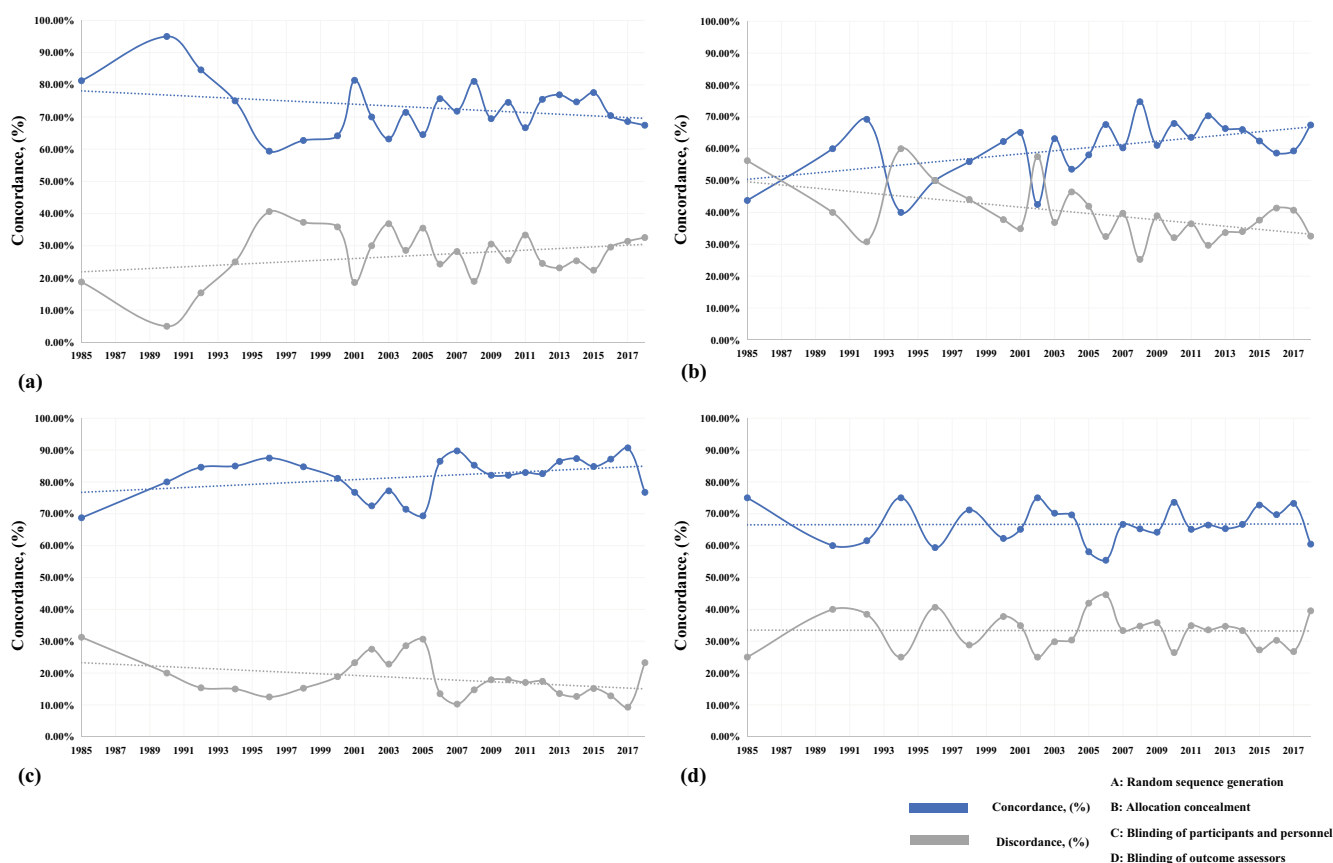


FIGURE 2 Concordance rate between RobotReviewer and human reviewers over time.

4 | DISCUSSION

In this study, we investigated the external validity of RobotReviewer on automatic risk of bias assessment using a large empirical dataset. The findings suggest that there were domain-specific differences in the level of agreement between RobotReviewer and human

reviewers, with moderate agreement for some of the domains (i.e., randomized sequence generation, blinding of participants and personnel), while fair agreements for the remaining domains (i.e., allocation concealment, blinding of outcome assessors).

Our findings regarding random sequence generation align closely with conclusions by Hirt et al.,⁶ showing

TABLE 2 Sensitivity analysis excluding RCTs with the risk of bias assessed via supplementary materials ($N = 1948$).

RoB domain	Human reviewers	Robot-Reviewer	Concordance (%)	Kappa (95%CI)	TP (%)	TN (%)	FP (%)	FN (%)	PPA (95%CI)	NPA (95%CI)
Random sequence generation	942 (48.4%)	1165 (59.8%)	72.7	0.46 (0.41–0.50)	40.5	32.3	19.4	7.9	0.84 (0.81–0.86)	0.63 (0.59–0.65)
Allocation concealment	1231 (63.2%)	1017 (52.2%)	63.0	0.25 (0.21–0.30)	39.2	23.8	13.0	24.0	0.62 (0.59–0.65)	0.65 (0.61–0.68)
Blinding of participants and personnel	1449 (74.4%)	1325 (68.0%)	83.3	0.60 (0.55–0.64)	62.8	20.4	5.2	11.6	0.84 (0.83–0.86)	0.80 (0.76–0.83)
Blinding of outcome assessors	557 (28.6%)	748 (38.4%)	67.0	0.27 (0.22–0.31)	17.0	50.0	21.4	11.6	0.59 (0.55–0.63)	0.70 (0.68–0.72)

moderate agreements and good PPA while lower NPA between RobotReviewer and human reviewers. For the domain of blinding of outcome assessors, the results of this survey are consistent with those of Armijo-Olivo's colleagues, indicating fair Cohen's kappa and approximately 50% PPA and 70% NPA.⁷ Notably, in terms of blinding of participants and personnel or outcome assessors, our assessment yielded higher agreement between RobotReviewer and human reviewers compared to previous similar studies.^{6–8} This difference in performance may likely be due to the fact that the RCTs used in these studies were from different sources. The 'samples' used in Hirt et al.'s⁶ study were RCTs ($n = 190$) from 23 Cochrane reviews on nursing, those used in Armijo-Olivo et al.'s study⁷ were RCTs ($n = 393$) from 43 meta-analyses from a sample of Cochrane reviews for physical therapy, and those used in Gates et al.'s⁸ were RCTs ($n = 1180$) from 13 systematic reviews or methodological studies from their own team. The "samples" used in our study were RCTs ($n = 1955$) from 151 systematic reviews for medication harms. Empirical evidence has shown that only 43% of the published RCTs reported information about medication harms.¹⁷

The concordance rate between RobotReviewer and human reviewers, (63%–77%) is either comparable or superior to evaluations reported for a panel of independent authors (35%–71%).¹⁸ This suggests that RobotReviewer may be a reasonable complementary tool in practical applications. However, the concordance rate remained steady within a certain range without any significant increase over time. It is widely acknowledged that standardized reporting increases the likelihood of a more accurate risk of bias assessment. This observation indicates that reporting randomized controlled trials (RCTs) is not yet fully standardized, despite substantial efforts made to improve reporting. These efforts include the publication of good practice guidelines (e.g., CONSORT 2010¹⁹), the refinement of quality

assessment tools,¹⁴ and the implementation of mandatory registration.²⁰ This highlights the persisting challenges in achieving consistent and standardized RCT reporting.

4.1 | Implications for automatic evidence synthesis practice

The above findings indicate that the performance of RobotReviewer is comparable or even superior to evaluations conducted by a panel of independent authors in certain aspects. It is essential to acknowledge that it cannot entirely replace human evaluations. It exhibits certain limitations that require attention and improvement. First, its assessment of the risk of bias is restricted to a binary classification, offering only "low risk" or "high/unclear risk" labels, although this is not necessarily a problem. In addition, currently the assessment of this tool is confined to only four safeguards, namely random sequence generation, allocation concealment, blinding of participants and personnel, and blinding of outcome assessors. It overlooks other possible safeguards of which at least 36 exist across analytical study designs,²¹ which makes an automated comprehensive judgment of risk of bias less feasible. Moreover, RobotReviewer currently lacks interactivity with users; future improvements could be directed towards a platform similar to ChatPDF, which allows for interaction with users. Each interaction with the system serves as a training opportunity, enabling continuous enhancement of its accuracy. Through this iterative process of user interaction and training, RobotReviewer has the potential to continually improve its performance. The tool also relies on electronic versions of studies and renders it less user-friendly for older or inaccessible studies, consequently reducing its overall utility. Addressing these limitations is essential for enhancing the effectiveness and usability of RobotReviewer.

Regarding the evaluation of RoB, the current recommended practice involves reviewers independently assessing the RoB of a trial and then reaching a consensus, as well as details of automatic tools.¹ The team at RobotReviewer propose an alternative approach by incorporating an automated method to replace one of the two reviewers, ensuring there is still a double independent evaluation.⁵ However, Jardim argued that the tool should not replace one of the two reviewers since an independent assessment by two human reviews is more likely to catch errors and provide valuable feedback to each other,²¹ but RobotReviewer lacks this ability. In future practical applications, it is essential to develop rapid review products like RobotReviewer further. These tools should include clear guidance for reviewers on when and how to apply automation techniques. By providing advice on leveraging automation appropriately, reviewers can strike a balance between the efficiency and assistance offered through automation.

4.2 | Strengths and limitations

To our knowledge, this is the largest validation study to date investigating the agreement and reliability of RobotReviewer compared to the consensus of human reviewers. It is important to underscore that our database has been checked three rounds to ensure the objective of the evaluation. Consequently, manual evaluations are more likely to align with the reference standards. Moreover, by incorporating almost 2000 RCTs and encompassing a wide range of topics without restrictions, we achieved enhanced representativeness and extrapolation potential for the samples.

Some limitations deserve to be emphasized. First, we were unable to quantitatively assess the time used by RobotReviewer since we did not collect specific information on the time required for both automatic and human reviewer assessments. As a recommendation for future research, we propose an approach involving three groups of comparisons: The first group would consist of two human reviewers for independent assessment. The two remaining groups employed a semi-automated approach. In the first arm, a human reviewer and RobotReviewer independently conduct the assessment. In the second arm, a human reviewer checks the results obtained by RobotReviewer. By comparing the time and reliability required for each group, a more comprehensive understanding of the benefits and efficiency of the semi-automated approach can be achieved. Secondly, although our search was not limited in terms of topics, it did impose restrictions on safety outcomes. This limitation may compromise the representativeness of current

studies, as empirical evidence indicates that only 43% of published trials actually report safety data.¹⁷ Finally, outcome specific safeguards in other RoB tools (not considered in this paper) may be a problem for RobotReviewer as it can only perform a study specific assessment. Therefore safeguards such as “*The outcome was objective and/or reliably measured*” or “*Cointerventions that could impact the outcome were comparable between groups or avoided*” may not be feasible, though can easily be handled by human reviewers.²² This is a current limitation of the RobotReviewer.

5 | CONCLUSIONS

Based on current evidence, RobotReviewer may serve as a supplementary evaluation tool in practical applications. However, the specific manner of its integration as an auxiliary tool requires further discussion and consideration. Furthermore, it is worth noting that the reliability of RobotReviewer has not shown obvious improvement over time, posing a big challenge in achieving the perfection and standardization of RCT reports. Addressing this issue remains a crucial area of focus for future efforts. As RobotReviewer continues to evolve as a rapid review tool, a series of ongoing evaluations of its reliability, utility, and potential to enhance human work will be essential. Such evaluations will offer valuable experience and guidance in improving its practicability, and we can ensure its optimal utility in assisting human reviewers in their tasks.

AUTHOR CONTRIBUTIONS

Yuan Tian: Data curation; conceptualization; writing – review and editing; methodology; formal analysis. **Xi Yang:** Writing – original draft; writing – review and editing; data curation; formal analysis. **Suhail A. Doi:** Writing – review and editing. **Luis Furuya-Kanamori:** Writing – review and editing. **Lifeng Lin:** Writing – review and editing. **Joey S. W. Kwong:** Writing – review and editing. **Chang Xu:** Writing – review and editing; data curation; methodology; conceptualization.

ACKNOWLEDGMENTS

The current study was supported by the National Natural Science Foundation of China (72204003) and an institutional funding from Shanghai Eastern Hepatobiliary Surgery Hospital of Naval Medical University (‘TengFei Project’, TF2024YZRH03). Suhail Doi was supported by Program Grant #NPRP-BSRA01-0406-210030 from the Qatar National Research Fund. The funding bodies had no role in any process of the study (i.e., study design,

analysis, interpretation of data, writing of the report, and the decision to submit the article for publication).

CONFLICT OF INTEREST STATEMENT

All authors have completed the ICMJE uniform disclosure form at www.icmje.org/coi_disclosure.pdf (available on request from the corresponding author) and declare: no support from any organization for the submitted work; no financial relationships with any organizations that might have an interest in the submitted work in the previous 3 years; no other relationships or activities that could appear to have influenced the submitted work.

DATA AVAILABILITY STATEMENT

The data of the study are shared via OSF platform <https://osf.io/k6w9q/>.

ORCID

Luis Furuya-Kanamori  <https://orcid.org/0000-0002-4337-9757>

Lifeng Lin  <https://orcid.org/0000-0002-3562-9816>

Chang Xu  <https://orcid.org/0000-0002-2627-1250>

REFERENCES

- Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*. 2021;372:n71.
- 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;15(3):37.
- Shojania KG, Sampson M, Ansari MT, Ji J, Doucette S, Moher D. How quickly do systematic reviews go out of date? A survival analysis. *Ann Intern Med*. 2007;147(4):224-233.
- Blaizot A, Veettil SK, Saidoung P, et al. Using artificial intelligence methods for systematic review in health sciences: a systematic review. *Res Synth Methods*. 2022;13(3):353-362.
- Marshall IJ, Kuiper J, Wallace BC. RobotReviewer: evaluation of a system for automatically assessing bias in clinical trials. *J Am Med Inform Assoc*. 2016;23(1):193-201.
- Hirt J, Meichlinger J, Schumacher P, Mueller G. Agreement in risk of bias assessment between RobotReviewer and human reviewers: an evaluation study on randomised controlled trials in nursing-related Cochrane reviews. *J Nurs Scholarsh*. 2021;53(2):246-254.
- Armijo-Olivo S, Craig R, Campbell S. Comparing machine and human reviewers to evaluate the risk of bias in randomized controlled trials. *Res Synth Methods*. 2020;11(3):484-493.
- Gates A, Vandermeer B, Hartling L. Technology-assisted risk of bias assessment in systematic reviews: a prospective cross-sectional evaluation of the RobotReviewer machine learning tool. *J Clin Epidemiol*. 2018;96:54-62.
- Fan S, Yu T, Yang X, Zhang R, Furuya-Kanamori L, Xu C. The SMART safety: an empirical dataset for evidence synthesis of adverse events. *Data Brief*. 2023;51:109639.
- Xu C, Zhang F, Doi SAR, et al. Influence of lack of blinding on the estimation of medication-related harms: a retrospective cohort study of randomized controlled trials. *BMC Med*. 2024;22(1):83.
- Zorzela L, Golder S, Liu Y, et al. Quality of reporting in systematic reviews of adverse events: systematic review. *BMJ*. 2014;348:f7668.
- Xu C, Yu T, Furuya-Kanamori L, et al. Validity of data extraction in evidence synthesis practice of adverse events: reproducibility study. *BMJ*. 2022;377:e069155.
- Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. *Syst Rev*. 2016;5(1):210.
- Marshall IJ, Kuiper J, Banner E, Wallace BC. Automating biomedical evidence synthesis: RobotReviewer. *Proc Conf Assoc Comput Linguist Meet*. 2017;2017:7-12.
- Wang Y, Ghadimi M, Wang Q, et al. Instruments assessing risk of bias of randomized trials frequently included items that are not addressing risk of bias issues. *J Clin Epidemiol*. 2022;152:218-225.
- Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33(1):159-174.
- Golder S, Loke YK, Wright K, Norman G. Reporting of adverse events in published and unpublished studies of health care interventions: a systematic review. *PLoS Med*. 2016;13(9):e1002127.
- Jordan VM, Lensen SF, Farquhar CM. There were large discrepancies in risk of bias tool judgments when a randomized controlled trial appeared in more than one systematic review. *J Clin Epidemiol*. 2017;81:72-76.
- Schulz KF, Altman DG, Moher D, CONSORT Group. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *BMC Med*. 2010;8:18.
- Speich B, Gryaznov D, Busse JW, et al. Nonregistration, discontinuation, and nonpublication of randomized trials: a repeated metaresearch analysis. *PLoS Med*. 2022;19(4):e1003980.
- Jardim PSJ, Rose CJ, Ames HM, Echavez JFM, van de Velde S, Muller AE. Automating risk of bias assessment in systematic reviews: a real-time mixed methods comparison of human researchers to a machine learning system. *BMC Med Res Methodol*. 2022;22(1):167.
- Ahmed AI, Kaleem MZ, Elshoeibi AM, et al. MASTER scale for methodological quality assessment: reliability assessment and update. *J Evid Based Med*. 2024;17(2):263-266.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Tian Y, Yang X, Doi SA, et al. Towards the automatic risk of bias assessment on randomized controlled trials: A comparison of RobotReviewer and humans. *Res Syn Meth*. 2024;15(6):1111-1119. doi:10.1002/jrsm.1761