Evaluation of Machine Learning-Assisted Risk of Bias Assessments for Supporting Systematic Reviews: A Case Study of a Review in Movement Disorders

SA22

Michaela Lunan-Taylor; Sathushan Thurairajah; Jose S. Marcano-Belisario; Louise Hartley

RTI Health Solutions, Manchester, United Kingdom

BACKGROUND

- Systematic literature reviews (SLRs) constitute a foundational methodology in health economics and outcomes research, underpinning evidence-based decision-making.
- As the demand for timely, efficient, and reliable evidence generation has increased, artificial intelligence (AI) technologies have been progressively integrated to support and automate key stages of the SLR process. Recent research has suggested that automated risk of bias (ROB) assessments can achieve moderate to high performance compared with humans[1-3] and thereby could equate to time savings within the SLR process.
- Consequently, AI ROB assessment needs to be rigorously tested to ensure the robustness, transparency, and reproducibility of systematic reviews, as well as appropriate integration of AI use within established workflows.

OBJECTIVES

- To evaluate AI-assisted ROB assessment using the Cochrane Risk of Bias 2 (ROB2) tool in an SLR of randomized-controlled trials (RCTs) of movement disorders, assessing (1) performance (accuracy, recall, precision, and F1 score), (2) time requirements, and (3) implications for reviewer workflow and oversight.

METHODS

- We conducted a clinical systematic review of RCTs in movement disorders and tested ROB2 in the AI-assisted evidence synthesis tool, Nested Knowledge, using Adaptive Smart Tags.
- Prompts for quality appraisal were iteratively developed, piloted, and finalized prior to full assessment based on all questions within the ROB2 tool.
- Data were extracted across narrative text, tables, and figures from publications, abstracts, and ClinicalTrials.gov records.
- AI-generated extractions were quality checked by human researchers, who corrected errors and supplemented missing data. For studies reported across multiple publications, data were collated at the study level to remove duplication.
- True positives = correct extraction from any of the included references for a study. False positives = incorrect data. False negatives = missed or incomplete data. True negatives = appropriately unreported and appropriately assessed as "No information" by the AI.
- Time spent on prompt development and piloting, as well as on quality checking and supplementing AI-generated extractions, was recorded. These time requirements were compared with benchmark estimates for fully manual ROB2 assessment (based on typical researcher-reported averages rather than review-specific timings). Time comparisons were made per study.

See Figure 1

References

1. Arregui M, et al. Value Health. 2025;28(S1):SA28.
2. Taneri PE. Cochrane Evid Synth Methods. 2025 Aug 31;3(5):e70044.
3. Jardim PSJ, et al. BMC Med Res Methodol. 2022;22(1):167.

The power of **knowledge**
The value of **understanding**

RESULTS

SLR Results

- The SLR included 39 references, corresponding to 24 unique studies included for data extraction.

AI Performance (Table 1)

- AI-assisted ROB2 assessment demonstrated high recall and F1 score, with moderate accuracy and precision.

Table 1. Accuracy, Recall, Precision, and F1 of AI ROB2 Assessment

	ROB2
Accuracy	0.71
Recall	0.94
Precision	0.71
F1	0.81

Table 2. Average Time for ROB by Assessment Approach and Workflow Stage

	Template data extraction sheet (human)/prompt building and piloting (AI)	ROB2 Assessment	QC	Additional QC of human extracted data	Merging individual article data to a single study
Benchmark manual extraction (typical average)	1 hour total	30 minutes per article	15 minutes per article	NA	NA (completed during assessment)
AI extraction	6 hours total	NA	8 minutes per article	3 minutes per article	5 minutes per linked article

NA = not applicable; QC = quality check.

Correct AI Assessment

- AI aligned with human assessments for all studies for the following ROB2 items:
  - Was the allocation sequence random?
  - Was an appropriate analysis used to estimate the effect of assignment to intervention?
  - Was the method of measuring the outcome inappropriate?
  - Were outcome assessors aware of the intervention received by study participants?
- AI aligned with human reviewer judgments for > 90% of studies for the following items:
  - Did baseline differences between intervention groups suggest a problem with the randomization process?
  - Were there deviations from the intended intervention that were likely to have affected the outcome?
  - Were data for this outcome available for all, or nearly all, participants randomized?
  - Could assessment of the outcome have been influenced by knowledge of intervention received?
  - Is it likely that assessment of the outcome was influenced by knowledge of intervention received?

Issues/Faults

- Common faults included the following:
  - **Blinding:** Frequent misclassification of participant/caregiver/investigator blinding (47/48 assessments, 97%), despite extraction of statements indicating double blinding.
  - **Nonadherence:** Frequent misclassification of whether nonadherence could have affected outcomes (23/24 assessments, 96%), despite identification of relevant text.
  - **Outcome measurement/ascertainment:** Misclassification of whether outcome measurement could have differed between groups; outcome descriptions were often cited as justification for an incorrect "yes" response (15/24 assessments, 63%).
  - **Selective reporting:** Misclassification of whether the numerical result was selected from multiple eligible measurements/timepoints within the outcome domain; "yes" was frequently assigned despite reporting of prespecified timepoints (12/24 assessments, 50%).

Timing (Table 2)

- Although these errors were consistent, AI-assisted assessment reduced total time per study after accounting for QC and consolidation of linked publications.
  - AI-assisted assessment was faster initially but required significant time for prompt development and piloting.
  - Prompt development took ~6 hours (vs. ~1 for manual templates); QC took ~8 minutes/article (vs. ~15 manually).
  - Additional time (~5 minutes/article) was needed to merge linked publications.
  - Overall, total time per study indicated time savings with AI: ~31 minutes (AI + QC + linking) vs. ~48 minutes (benchmark manual approach).

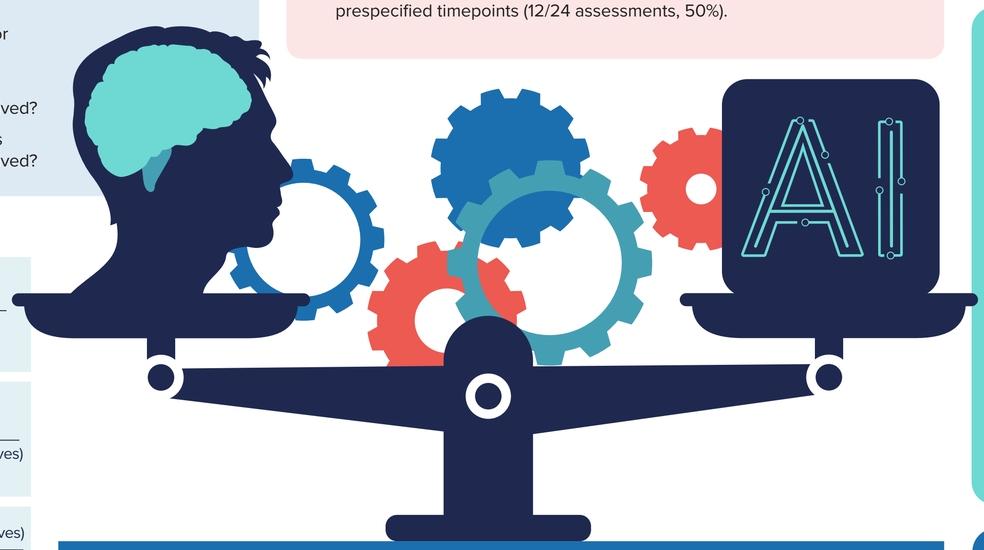
Figure 1.

Recall
Proportion of relevant data that is correctly identified by the AI out of all relevant data available
$$\frac{\text{True positives}}{(\text{True positives} + \text{False negatives})}$$

Precision
Proportion of relevant data among that identified by the AI
$$\frac{\text{True positives}}{(\text{True positives} + \text{False positives})}$$

Accuracy
Proportion of correctly identified data out of all data captured by the AI
$$\frac{(\text{True positives} + \text{True negatives})}{\text{True positives} + \text{True negatives} + \text{False positives} + \text{False negatives}}$$

F1 score
Harmonic mean of precision and recall
$$\frac{2 \times (\text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})}$$



LIMITATIONS

- Time comparisons were benchmarked against typical manual averages rather than review-specific data due to feasibility constraints.
- ROB2 includes conditional (dependent) signaling questions; Nested Knowledge's workflow did not support conditional logic, which may have contributed to correlated errors and/or inflated apparent performance for some items.
- AI is constantly improving and adapting; results may differ if this review were performed today.
- Nested Knowledge has since created a Smart Critical Appraisal tool that may prove more or less effective than the method applied in this study.

DISCUSSION

- AI-assisted ROB2 assessment showed high recall (0.94) and F1 score (0.81), supporting use as a drafting and screening aid. Accuracy and precision were moderate (both 0.71), indicating that human quality checking remains necessary.
- Errors were concentrated in specific questions (i.e., blinding, nonadherence, selective reporting). In many instances, the AI extracted relevant evidence from the source article but applied an incorrect ROB2 response, suggesting limitations in translating extracted text into judgments for these signaling questions. Further targeted prompt refinement as well as improved translation of extracted evidence into ROB2 judgments may improve performance.
- Timing results indicate a trade-off between upfront configuration and downstream efficiency. Prompt development required ~6 hours; however, per-article QC was shorter than a benchmark manual approach (~8 vs. ~15 minutes), and total time per study was lower after consolidation of linked publications (~31 vs. ~48 minutes). These comparisons use benchmark manual timings (typical averages) rather than review-specific head-to-head measurements; time savings, therefore, are indicative and are most likely to accrue in larger reviews and through repeated use of prompts in other SLRs, whereas QC and linking articles remain key determinants of total effort.

CONCLUSIONS

- AI-assisted ROB2 assessment demonstrated high recall and reduced total reviewer time per study when implemented with human-in-the-loop QC. However, systematic misclassification for specific ROB2 items indicates that human oversight remains essential. These findings support targeted integration of AI into SLR workflows, with enhanced prompts and reviewer checks focused on recurring error modes.

RTI Health Solutions
2nd Floor, The Pavilion, Towers Business Park, Wilmslow Road, Didsbury, Manchester, United Kingdom
Telephone: +44(0)161.447.9033

Contact: Michaela Lunan-Taylor (mlunan@rti.org)

Presented at: ISPOR 2026; 17-20 May 2026; Philadelphia, PA, USA