Validation of the MetaSLR RoB Module: A Multi-Agent Generative AI System for Automating Cochrane Risk of Bias 2.0 Assessments

MSR134

Inderpreet Singh-Marwaha, Rajdeep Kaur, Ritesh Dubey, Shubhram Pandey, Barinder Singh
Pharmacoevidence, Mohali, India

Evidence Pharmaco®
Science-led HEOR & Tech-driven AI

INTRODUCTION

- Risk of Bias (RoB) assessment is a critical yet resource-intensive element of systematic literature reviews (SLRs), often limited by inter-rater consistency
- This study validated a Retrieval-Augmented Generation (RAG)-enabled multi-agent Generative AI (GenAI) module within the MetaSLR platform, benchmarking its performance and reliability against subject matter experts (SMEs) for Cochrane RoB 2.0 assessments

METHODS

- The MetaSLR RoB module employed a multi-agent generative AI architecture in which specialized sub-agents autonomously evaluated signalling questions (SQs) through a dynamic checklist framework, whereby responses to preceding SQs informed subsequent assessment logic (**Figure 1**)
- Domain-level and overall risk-of-bias judgments ('low risk,' 'some concerns,' or 'high risk') were generated from SQ responses according to the decision framework outlined in the Cochrane Risk of Bias 2.0 tool for randomized controlled trials (RCT)[1]
- The tool was validated using 36 RCTs identified from two historical SLRs. The concordance with SME consensus assessments was evaluated using performance metrics including observed agreement, sensitivity (recall), specificity, and F1 score
- Inter-rater reliability between the GenAI tool and SME was assessed using Cohen's κ coefficients (weighted for both domain-level and overall judgments) and Gwet's AC1/AC2 coefficients, with higher values indicating stronger agreement beyond chance
- Systematic directional bias in AI-generated assessments was evaluated using a directional bias metric [defined as Δ = mean (AI score) − mean (SME score)], where positive values indicate more conservative AI judgments (i.e., a tendency to assign higher risk-severity) relative to SMEs
- Efficiency gains from MetaSLR's RoB module were assessed by comparing aggregate and per-study completion times for AI- versus SME-based RoB evaluations

RESULTS

- Across the 36 RCTs, the evaluation included 792 paired SQ level observations, 180 paired domain-level judgments, and 36 paired overall risk-of-bias assessments
- Strong adherence to SQ decision logic was observed, with 78.2% agreement between AI and SME ratings. Performance metrics demonstrated high specificity (91.9%), moderate sensitivity (57.0%), and an F1 score of 54.4%. Inter-rater reliability was substantial (κ = 0.637; Gwet's AC1 = 0.727)
- Agreement between AI and SME assessments was highest for 'low risk' classifications at both domain and overall levels. Discordant cases primarily reflected AI-driven upward risk reclassification (e.g., from 'low risk' to 'some concerns' or 'high risk'), resulting in approximately 2.6-fold and 2-fold higher frequencies of 'high risk' ratings at domain and overall levels, respectively **(Figures 2–3)**
- Weighted Gwet's AC2 coefficients, which account for class imbalance, indicated substantial reliability at the domain level (AC2 = 0.742; agreement = 67.2%; specificity = 64.1%) and moderate reliability at the overall level (AC2 = 0.453; agreement = 55.6%; specificity = 78.1%) **(Figure 4).**
- Directional bias analysis indicated that AI assessments were more conservative than SME ratings (overall RoB Δ = 0.19; domain-level Δ = 0.07), with no evidence of systematic underestimation of risk **(Figure 5)**
- GenAI reduced total assessment time by 46% (9.1 hours saved, including adjudication), decreasing mean per-study completion time from 15 minutes to 15 seconds **(Figure 5).**

Figure 1: RAG + Architecture of a Multi-Agent RAG Framework for Automated Risk of Bias (RoB 2.0) Assessments

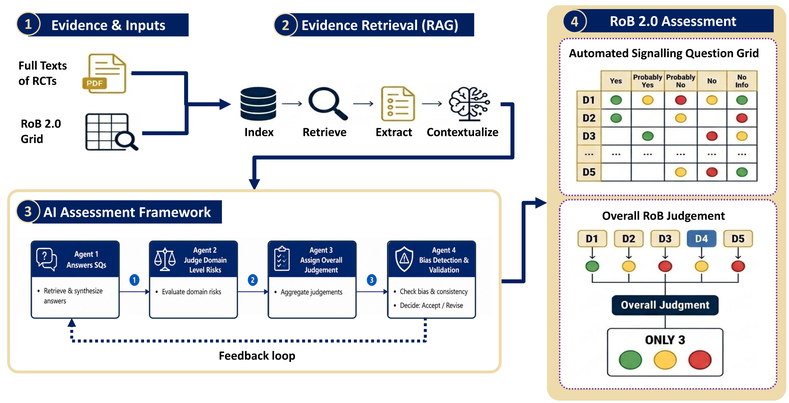


Figure 2: Distribution of Risk-of-Bias Ratings Across Domains and Overall Judgments

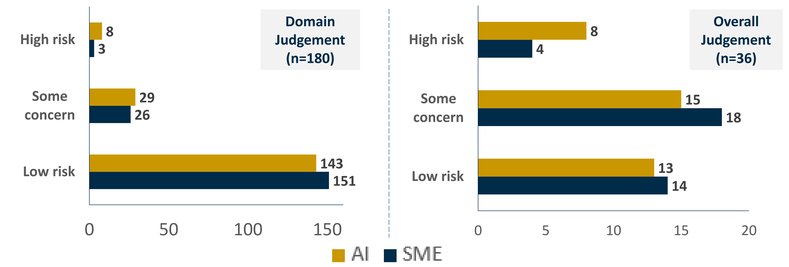
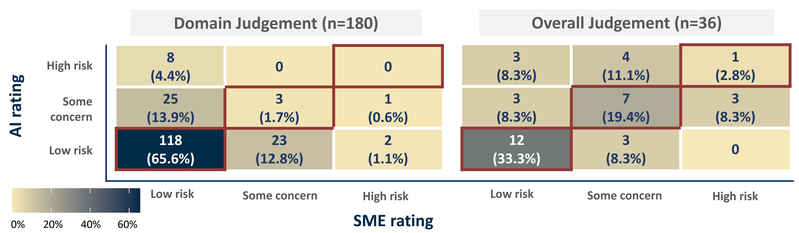


Figure 3: Inter-Rater Agreement Matrices (Confusion Matrices) Comparing AI and SME Assessments for (A) Domain-Level and (B) Overall Risk of Bias Judgments



Note: Diagonal cells represent exact AI-SME concordance; off-diagonal cells quantify disagreement, while cells above the diagonal indicate AI conservative bias (higher risk assigned by AI); shading intensity reflects the percentage of total observations

Figure 4: Observed and Chance-Adjusted Agreement Statistics for AI and SME RoB Assessments

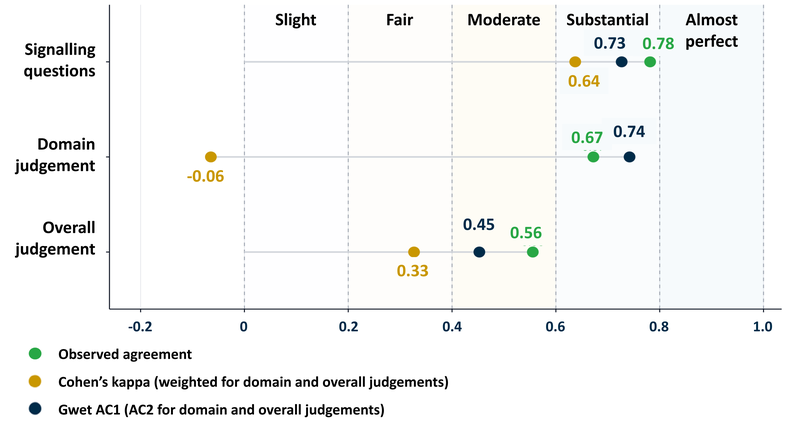
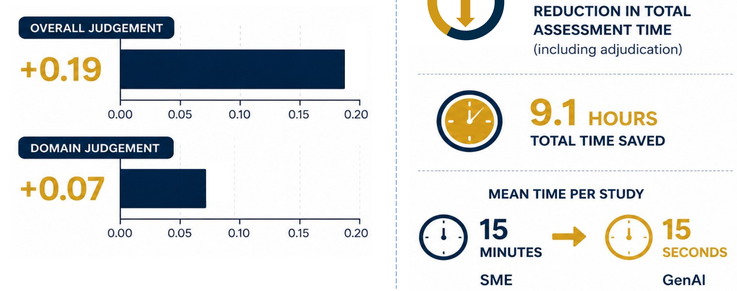


Figure 5: Directional Bias and Efficiency Outcomes of MetaSLR RoB module



LIMITATIONS

- The generalizability of these findings is constrained by the moderate sample size from two SLRs (n=36 RCTs), necessitating further validation across diverse therapeutic areas and type of reviews
- The MetaSLR RoB module validation demonstrated a systematic conservative bias in AI led RoB assessments, resulting in moderate exact concordance for overall risk judgments, reinforcing the need for human-in-the-loop (HITL) adjudication
- While SME consensus was used as the reference standard, evaluation of exact concordance is limited by inherent inter-rater variability in complex expert appraisals; the impact of reviewer subjectivity will be examined in future studies

CONCLUSIONS

- ❖ The multi-agent RAG-enabled MetaSLR RoB module demonstrated high methodological validity in adhering to Cochrane conditional logic (high specificity)
- ❖ While exact concordance with SMEs was moderate, the GenAI based system systematically exhibited a conservative bias in its bias assessments
- ❖ These findings support the system's implementation as a reliable, high-fidelity quality appraisal tool for HITL governed evidence synthesis

References:

1. Sterne JAC, Savović J, Page MJ, Elbers RG, Blencowe NS, Boutron I, Cates CJ, Cheng H-Y, Corbett MS, Eldridge SM, Hernán MA, Hopewell S, Hróbjartsson A, Junqueira DR, Jüni P, Kirkham JJ, Lasserson T, Li T, McAleenan A, Reeves BC, Shepperd S, Shrier I, Stewart LA, Tilling K, White IR, Whiting PF, Higgins JPT. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ 2019; 366: l4898.

Correspondence: Barinder Singh; barinder.singh@pharmacoevidence.com

Disclosure: ISM, RK, RD, SP, and BS, the authors declare that they have no conflict of interest

ISPOR, May 17-20, 2026, Philadelphia, PA, United States