Leveraging artificial intelligence to streamline study quality assessment in systematic literature reviews

Maria Arregui, PhD[1]; Maria Koufopoulou, MSc[2]; Sarah Cadarette, MPH[3]; Erika Wissinger, PhD[3]

[1]Cencora, Hannover, Germany. [2]Cencora, London, United Kingdom. [3]Cencora, Conshohocken, PA, USA.

Background

- Systematic literature reviews (SLRs) are vital to evidence-based decision-making in health economics and outcomes research (HEOR). As critical components of health technology assessments (HTAs), SLRs synthesize existing research, facilitating informed decisions in healthcare policy and clinical practice.
- The reliability and validity of findings in SLRs are paramount, heavily relying on the rigorous evaluation of individual study quality. Globally, HTA agencies mandate quality assessment as a fundamental aspect of SLRs. Although methodological requirements vary across agencies, there is a universal emphasis on the necessity for thorough quality assessment and critical appraisal of studies to ensure the integrity and validity of evidence.[1]
- Quality assessment of the studies included in an SLR involves using validated tools tailored to specific study designs. Examples include the Cochrane Risk of Bias Tool 1 (RoB 1) tool for randomized controlled trials (RCTs),[2] the Newcastle-Ottawa Scale for cohort and case-control studies,[3] and the Motheral Checklist for retrospective studies.[4]
- Each tool comprises a set of predefined criteria designed to identify potential biases and methodological flaws relevant to the study type. Despite their structured nature, the traditional assessment process is labor-intensive and time-consuming, often requiring significant human resources and expertise.
- Recent advancements in artificial intelligence (AI) present promising opportunities to enhance the quality assessment process in SLRs. AI-powered tools can rapidly analyze large volumes of data, potentially streamlining tasks and reducing time investment. The United Kingdom's national HTA agency, the National Institute for Health and Care Excellence (NICE), position statement underscores AI's potential to automate various steps in literature search and review processes; however, cautious adoption is advised, as these applications continue to evolve. Efforts by Cochrane and the Guidelines International Network to develop guidance for AI use in evidence synthesis will provide valuable practices for organizations.[5]
- In this context, our study explores using an internal, closed-system AI tool to assess study quality within an SLR framework.

Objective

- The primary objective was to evaluate the accuracy of the closed-system AI tool in comparison to traditional manual assessments conducted by trained systematic reviewers.
- By examining the concordance between AI-generated assessments and those of human reviewers, we aimed to highlight the potential benefits and limitations of integrating AI into SLR processes. Through this work, we seek to contribute to the evolving landscape of AI applications in HEOR, ultimately enhancing the efficiency of SLRs.

Methods

- **Quality assessment tools:** We used quality assessment tools specifically selected according to the study designs of the studies included in the SLR. These tools were the RoB 1 tool for RCTs, the Newcastle-Ottawa Scale for prospective cohort and case-control studies, and the Motheral Checklist for retrospective cohort and registry studies. Tool selection was guided by the recommendations of the Centre for Research and Dissemination Guidance for Reviews and the Cochrane Handbook for Systematic Reviews of Interventions (version 6.5).[6]
- **Handling multiple publications:** For studies with multiple related publications, the primary publication with the most comprehensive reporting of methodology and outcomes was prioritized for the quality assessment process.
- **Exclusion of conference abstracts:** Studies presented solely in conference abstract form were excluded from the quality assessment due to the limited information offered by such publications.
- **Review process:** An experienced reviewer initially assessed the quality of each study, and these assessments were subsequently validated by a second reviewer to ensure accuracy and consistency. Once consensus was reached between the reviewers, tailored prompts were used to conduct the quality assessment using the AI tool.
- **AI tool task:** PDF files of the studies were uploaded to the AI system one by one. After uploading each study, the appropriate prompt for the specific study design was selected and provided to the AI tool. The AI was then tasked with analyzing each study publication, responding to quality inquiries, and generating detailed responses supported by verbatim text excerpts from the publications.
- **Accuracy and reliability:** To evaluate the accuracy and reliability of the AI tool, its responses were qualitatively compared with those of trained systematic reviewers. For each study design, the percentage of agreement was calculated by determining the total number of answers, which was obtained by multiplying the number of studies by the number of quality questions assessed per study. The number of instances where the AI tool's answers matched those of the reviewer was recorded. The agreement percentage was then calculated by dividing the number of matching answers by the total number of answers and multiplying by 100 to express it as a percentage.
- **Figure 1** provides an overview of the methodology employed in this study, illustrating the workflow from the receipt of publications to the generation of comprehensive responses.

Figure 1. Workflow overview: from publication input into the closed-system AI tool to comparative analysis of quality assessments between AI and human reviewers

Results

- **Table 1** displays findings categorized by study type, quality assessment tools used with a brief description of the tool, the number of quality items assessed, and main assessment outcomes.

Table 1. Summary of quality assessment results: Comparison of AI and human reviewer evaluations across different study types

Study type (number of studies)	Quality assessment tool used	Description of the tool (number of quality items assessed)	AI vs reviewer agreement	Main discrepancy
RCTs (6)	Cochrane Risk-of-Bias Tool	Evaluates key aspects such as randomization, allocation concealment, blinding, outcome reporting, and intention-to-treat analysis, etc (7)	81%	Interpretation of allocation concealment
Retrospective observational (4)	Motheral Checklist	Assesses quality by examining data source reliability, clarity of research questions, eligibility criteria, adequacy of statistical analyses, etc (10)	83%	Data source reliability and validity
Prospective observational (18)	Newcastle-Ottawa Scale	Focuses on patient selection, comparability of cohorts, outcome assessment, follow-up adequacy, etc (7)	82%	Comparability of exposed and non-exposed cohorts

- **Study quality overview:** The quality of the included studies was generally good, with no studies categorized as having a high risk of bias overall. Across RCTs, common concerns included inadequate concealment of the allocation sequence and lack of blinding to treatment assignment. In prospective observational studies, concerns focused on cohort representativeness and blinding of outcome assessments. In retrospective studies, the largest source of concern was the reliability and validity of the data source used.
- **RCTs:** For the 6 RCTs assessed, the AI-generated responses aligned with the systematic reviewer's evaluations in 81% of cases. Discrepancies were primarily observed in the interpretation of allocation concealment, indicating an area for potential refinement in the AI's assessment capabilities. Notably, this was an area of diminished quality. **Figure 2** provides an illustrative example of the AI tool's assessment of random allocation methods within a given RCT.

Figure 2. Example of the AI tool's assessment of random allocation generation for RCTs using the Cochrane tool

a. Systematic reviewer-designed prompt input to AI tool

Instructions: *Please assess the quality of Study XX by thoroughly reviewing the attached PDF. Respond to the following questions with detailed analyses, ensuring that your answers are substantiated by verbatim excerpts from the publication.*

Was the method used to generate random allocations adequate?

1. ***Yes:*** *A random component was used in the sequence generation process.*
   Examples: Computer-generated random numbers, random number tables, coin tossing, shuffling cards or envelopes, dice throwing, drawing lots.
2. ***No:*** *No random element was used, or the sequence is predictable.*
   Examples: Alternation, date-based methods (birth or admission), patient record numbers, decisions by clinicians or participants, availability-based allocation, any systematic/haphazard method.
3. ***No information:*** *The study is described as randomized without further detail.*
4. ***Judgment calls:*** *In some situations, respond with "Probably yes" or "Probably no."*
   Guidance: For a large trial run by an experienced clinical trials unit, absence of specific information might lead to "Probably yes."

b. AI tool response

Answer: Yes

The method used to generate random allocations in the study was adequate. The publication provides specific details about the randomization process, indicating the use of permuted-block randomization, a recognized method that incorporates a random component in sequence generation. Here is a generalized version of the relevant information:

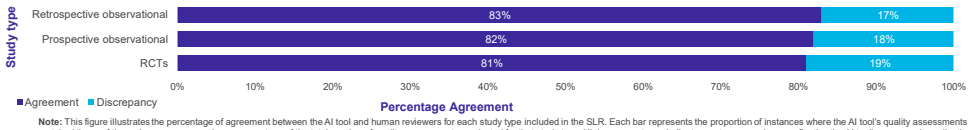
"We used permuted-block randomization (block size four) to assign patients (2:1:1) via an interactive voice and web response system to [Treatment A, Treatment B, or Treatment C].[a] Stratification factors included [Factor 1 (eg, Status X vs Status Y) and Factor 2 (eg, Status Z vs Status W)].[a]

Note: Panel A shows the quality assessment prompt for evaluating random allocation methods in a specific RCT using the RoB 1 tool. **Panel B** displays the AI tool's response.

[a] Specific details of treatments and stratification factors have been generalized to ensure compliance with copyright standards.

- **Prospective cohort studies:** The AI tool exhibited an 82% agreement with the systematic reviewer for the 4 studies assessed using the Newcastle-Ottawa Scale. Discrepancies mainly involved the comparability of exposed and nonexposed cohorts.
- **Retrospective observational studies:** Among the 18 retrospective cohort and registry studies evaluated using the Motheral Checklist, the AI tool achieved an 83% concordance rate with the systematic reviewer. Discrepancies often pertained to assessments of data source reliability and validity, suggesting specific methodological nuances where human oversight remains critical.
- **Figure 3** illustrates the percentage of agreement between the AI tool and human reviewers for each study type included in the SLR.

Figure 3. AI vs reviewer agreement for each study type



Note: This figure illustrates the percentage of agreement between the AI tool and human reviewers for each study type included in the SLR. Each bar represents the proportion of instances where the AI tool's quality assessments matched those of the reviewers, expressed as a percentage of the total number of quality assessments conducted for that study type. Higher percentages indicate greater concordance, reflecting the AI tool's accuracy in replicating human evaluations.

Conclusions

- Overall, the closed-system AI tool reliably delivered detailed responses, significantly reducing the time required for quality assessment compared to traditional manual methods. The generally good quality of the studies may have contributed to the high accuracy of the AI tool. In both RCTs and retrospective observational studies, the AI tool performed worse on items that showed lower quality as assessed by systematic reviewers. While the AI tool performed well at identifying reported information, lack of information was more difficult for the AI to evaluate.
- This case study underscores the potential value of AI-powered tools in streamlining the quality assessment of studies within SLRs, as evidenced by the high agreement rates with trained systematic reviewers across diverse study types.
- Nonetheless, human oversight remains indispensable to ensure robustness and effectively address any potential interpretation discrepancies.

References

1. Wright C, Swanston L, Nicholson L, Marjenberg Z. HTA360: A comparative assessment of systematic literature review requirements for health technology assessment, globally. *Value Health*. 2023;26(12):S389-S390. **2.** Higgins JPT, Green S, eds. *Cochrane Handbook for Systematic Reviews of Interventions. Version 5.1.0.* Cochrane; 2011. Accessed April 2025. https://methods.cochrane.org/bias/risk-bias-tool **3.** Wells GA, Shea B, O'Connell D, et al. The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses. Ottawa Hospital Research Institute; 2025. Accessed April 2025. https://www.ohri.ca/programs/clinical_epidemiology/oxford.asp **4.** Motheral B, Brooks J, Clark MA, et al. A checklist for retrospective database studies--report of the ISPOR Task Force on Retrospective Databases. *Value Health*. 2003;6(2):90-97. doi:10.1046/j.1524-4733.2003.00242.x **5.** National Institute for Health and Care Excellence (NICE). Use of AI in evidence generation: NICE position statement. NICE; 2024. Accessed April 2025. https://www.nice.org.uk/about/what-we-do/our-research-work/use-of-ai-in-evidence-generation--nice-position-statement **6.** Higgins JPT, Thomas J, Chandler J, eds. *Cochrane Handbook for Systematic Reviews of Interventions. Version 6.5.* Cochrane; 2024. Accessed April 2025. www.training.cochrane.org/handbook

Presented at: ISPOR 2025; May 13-16, 2025; Montreal, Quebec, Canada
This study was funded by Cencora.