Evaluating the performance of an artificial intelligence-powered tool for assessing systematic literature reviews using the preferred reporting items for systematic reviews and meta-analyses (PRISMA) 2020 checklist

Maria Arregui, PhD[1]; Evelyn Gomez Espinosa, BSc, PhD[2]; Erika Wissinger, PhD[3]; Maria Koufopoulou, MSc[2]

[1]Cencora, Hannover, Germany; [2]Cencora, London, UK; [3]Cencora, Conshohocken, PA, USA

Background

- Systematic literature reviews (SLRs) are fundamental to evidence-based decision-making in health economics and outcomes research (HEOR), supporting health technology assessment (HTA) submissions, reimbursement decisions, guideline development, and strategic evidence planning. High-quality SLR reporting must be transparent, complete, and reproducible.
- The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 statement provides an internationally endorsed framework of 27 reporting items to standardize documentation of review rationale, methods, and findings and is widely used to evaluate reporting completeness in HTA and peer-review contexts.[1,2]
- Assessing PRISMA adherence is a checklist-based, rules-driven activity that requires reviewers to verify whether each reporting criterion is explicitly addressed. This manual process is time- and resource-intensive because reviewers must navigate manuscripts and supplementary files to locate supporting information for each item.
- As the volume and velocity of evidence synthesis increase, there is growing interest in approaches that can improve efficiency and consistency without compromising methodological rigor. Advances in artificial intelligence (AI) may support structured, auditable tasks such as checklist-based compliance checks.
- From an HTA perspective, responsible use of AI in evidence generation emphasizes well-bounded use cases, transparency, validation, and human oversight, with AI intended to augment (not replace) expert judgment—particularly for interpretive decisions.[3,4]

Objective

- To evaluate whether a customized generative AI (genAI) chatbot, operating entirely within Cencora's secure internal environment (no external data sharing or model training), can replicate expert human assessments of PRISMA 2020 adherence across published SLRs.
- To quantify item-level concordance between genAI and human reviewers and to identify PRISMA domains that are most amenable to automation vs those that require human oversight.

Methods

- **Overview:** The study workflow is summarized in **Figure 1**.
- **Sample:** Six published SLRs were selected across multiple therapeutic areas.
- **Independent assessments:** Each SLR was assessed by (1) a human reviewer experienced in PRISMA-guided appraisal and (2) a genAI chatbot.
- **GenAI configuration and tasking:** Structured prompts were aligned to all 27 PRISMA 2020 items, operationalized into 42 PRISMA-derived questions. For each question, the genAI chatbot was instructed to (a) identify relevant locations (e.g., page/table/figure), (b) extract verbatim supporting text, and (c) judge fulfillment. **Figure 2** provides an illustrative prompt excerpt.

Methods (cont.)

- **Human procedure and quality assurance:** The human reviewer completed full PRISMA-based evaluations with supporting text documentation. A second reviewer cross-validated 20% of evaluations to check consistency and accuracy.
- **Concordance and analysis:** GenAI outputs were compared with human assessments. Concordance was classified as full (same judgment and same evidence), partial (same judgment but different evidence), or none (different judgment and divergent evidence). Agreement was summarized per PRISMA section and overall.

Figure 1. Workflow summary

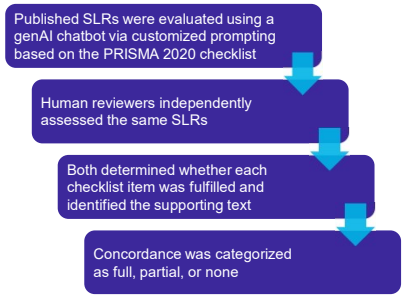


Figure 2. Example prompt used to assess PRISMA 2020 reporting in an SLR

Instructions: *Evaluate the provided SLR against the following PRISMA 2020 checklist items. For each item, extract the location where the item is reported (e.g., "Table X," "Page Y", "Figure Z") and the verbatim text corresponding to that item. If the item is not reported or unclear, indicate this explicitly.*

…

Eligibility criteria*: Specify the inclusion and exclusion criteria for the review and how studies were grouped for the syntheses.*

Information sources*: Specify all databases, registers, websites, organizations, reference lists, and other sources searched or consulted to identify studies. Specify the date when each source was last searched or consulted.*

Search strategy: *Present the full search strategies for all databases, registers, and websites, including any filters and limits used.*

…

Note: Only a subset of the 27 PRISMA 2020 items is displayed in the figure for brevity; formatting and output-structure instructions are omitted from the figure.

Results

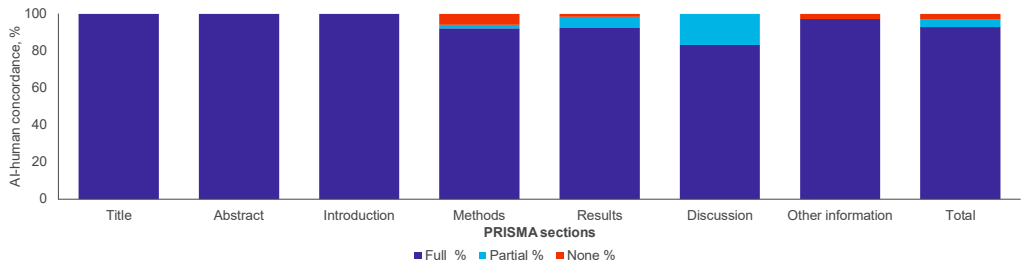
- **Overall PRISMA adherence:** The 6 included SLRs showed complete or near-complete reporting against the PRISMA 2020 checklist.
- **Overall concordance:** Across all 42 PRISMA-derived questions (252 total item assessments across 6 SLRs), the genAI chatbot achieved 93% full agreement with human reviewers (Table 1; Figure 3).
- **"Title, Abstract, Introduction, and Other Information":** The genAI chatbot achieved 100% agreement with human reviewers for these domains, which typically contain explicit, well-structured reporting elements.
- **"Methods":** Among the 17 "Methods" criteria, the chatbot achieved full concordance for 12 items. Discrepancies were concentrated in synthesis-related items (e.g., data preparation and tabulation/handling of extracted data).
- **"Results":** For the 11 "Results" items, the chatbot demonstrated strong performance with full agreement on 6 items and near-complete agreement on the remaining items, reflecting robust identification of structured results elements even when narrative formatting varied.
- **"Discussion":** Disagreements were most frequent in the "Discussion" domain, where items often require interpretive judgment (e.g., whether limitations of evidence and the review process and implications for policy/future research were explicitly and adequately addressed).

Table 1. AI-Human Concordance by PRISMA Section

PRISMA section (n)	Total n across 6 SLRs	AI-human concordance		
		Full, n (%)	Partial, n (%)	None, n (%)
"Title" (1)	6	6 (100)	0 (0.0)	0 (0.0)
"Abstract" (1)	6	6 (100)	0 (0.0)	0 (0.0)
"Introduction" (2)	12	12 (100)	0 (0.0)	0 (0.0)
"Methods" (17)	102	94 (92.2)	2 (2.0)	6 (5.9)
"Results" (11)	66	61 (92.4)	4 (6.1)	1 (1.5)
"Discussion" (4)	24	20 (83.3)	4 (16.7)	0 (0.0)
"Other information" (6)	36	35 (97.2)	0 (0.0)	1 (2.8)
Total (42)	252	234 (92.9)	10 (4.0)	8 (3.2)

N – number of questions.

Figure 3. Summary of AI-human agreement across PRISMA sections



Note: Agreement rates are expressed as a percentage based on the concordance of AI and human responses to PRISMA checklist items across 6 SLRs.

Conclusions

- A genAI chatbot operating in a secure internal environment demonstrated high concordance with expert human reviewers when assessing PRISMA 2020 reporting across 6 published SLRs.
- Agreement was highest for domains with structured, explicitly reported elements (e.g., "Title/Abstract/Introduction" and "Other Information"), supporting use of genAI for well-bounded, auditable checklist tasks.
- Lower concordance occurred for synthesis-related "Methods" items and interpretive "Discussion" items, indicating that human oversight remains essential for nuanced judgments and narrative interpretation.
- These findings suggest genAI-enabled PRISMA assessment could reduce reviewer burden, provided results are transparently reported and embedded within a human-led quality assurance workflow.

References

1. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*. 2021;372:n71.
2. Page MJ, Moher D, Bossuyt PM, et al. PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. *BMJ*. 2021;372:n160.
3. National Institute for Health and Care Excellence. Use of artificial intelligence in evidence generation: NICE position statement. Published August 15, 2024. https://www.nice.org.uk/corporate/ecd11
4. Trowman R, Boysen M, Migliore A, Valiotis G. Advancing the use of artificial intelligence in health technology assessment activities: insights and next steps from the 2025 HTAi Global Policy Forum. *Int J Technol Assess Health Care*. 2026;42(1):e5. doi:10.1017/S0266462325103395.

Presented at: Presented at: ISPOR US 2026 Conference; May 17-20, 2026; Philadelphia, PA, USA. This study was funded by Cencora.

cencora