EE446

Evaluating the Performance of an Artificial Intelligence (AI)-Powered Tool for Assessing Quality of Published Economic Evaluations: A Comparison With Human Reviewers Using the Drummond Checklist

Maria Arregui, PhD[1]; Maria Koufopoulou, MSc[2]

[1]Cencora, Hannover, Germany; [2]Cencora, London, United Kingdom

Background

- Economic evaluations are essential tools for guiding healthcare decision-making, offering insights into the cost-effectiveness, affordability, and value of healthcare interventions. They play a pivotal role in shaping policy development, funding allocation, and drug-pricing strategies. As healthcare systems worldwide face increasing demands and limited resources, ensuring the quality and reliability of economic evaluations is vital to avoid inefficient spending and ensure equitable access to care.[1]
- To uphold methodological rigor, the Drummond checklist[2] is widely used to assess the quality of economic evaluations. This tool comprises 35 criteria spanning 3 core domains (*study design, data collection, and analysis*) providing a structured framework for evaluating the robustness of economic studies.[2] Systematic literature reviews (SLRs), which underpin health technology assessments (HTAs), rely heavily on such quality assessments (QAs) to ensure trustworthy conclusions. Globally, HTA agencies mandate QA as a fundamental component of SLRs.[3]
- However, applying QA tools to large volumes of studies remains labour-intensive and time-consuming. Recent advances in artificial intelligence (AI), particularly in natural language processing (NLP) and machine learning (ML), offer promising avenues to streamline this process.[4]
- The transformative potential of AI in evidence synthesis has been acknowledged by prominent organizations across the global health research landscape. In August 2024, the United Kingdom's National Institute for Health and Care Excellence (NICE) published a formal position statement recognizing the potential of AI in evidence generation. NICE acknowledged that AI methods (including ML and generative AI) can automate aspects of literature search and review, offering efficiency gains in systematic reviews and HTAs. However, NICE emphasized that adoption must be careful, transparent, and aligned with existing regulations and standards to ensure trustworthiness, methodological rigor, and human oversight.[5]
- Similarly, in June 2025, Cochrane – alongside partners including the Campbell Collaboration, the Joanna Briggs Institute, and the Collaboration for Environmental Evidence – released the RAISE (*Responsible AI use in Evidence Synthesis*) framework.[6] This 3-paper guidance outlines best practices for integrating AI into SLRs, covering performance evaluation, ethical and regulatory considerations, and governance structures. The framework aims to support transparent, reliable, and responsible use of AI across the evidence synthesis ecosystem.[6]
- These initiatives reflect a growing consensus that AI can transform evidence synthesis, but it must be deployed with care and accountability. In this context, our study explores the use of an internal, closed-system AI tool to assess the quality of economic evaluations using the Drummond checklist.

Objective

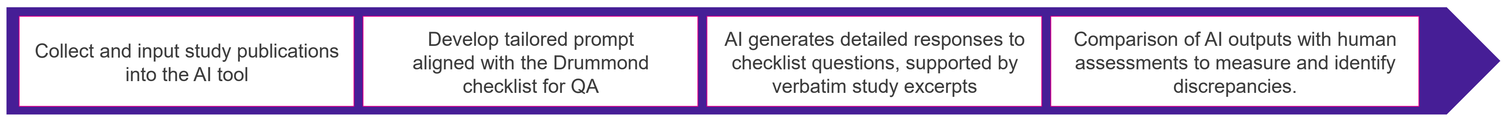
- The primary objective of this study was to evaluate the performance of an AI-powered tool in assessing the quality of economic evaluations using the Drummond checklist. Specifically, the study aimed to compare the AI-generated QAs with those conducted by trained human reviewers to identify areas of agreement and discrepancies.

Methods

- **Study selection:** An SLR was conducted to identify published economic evaluations relevant to an HTA submission. Eight peer-reviewed full-text manuscripts were included in the SLR, comprising 7 cost-effectiveness analyses (2 of which were cost-utility analyses) and 1 cost-minimization analysis.
- **Review process:** An experienced systematic reviewer initially conducted a detailed assessment of each study using the Drummond checklist. A second reviewer cross-validated these evaluations to ensure consistency and accuracy. Upon achieving consensus among human reviewers, a tailored prompt was developed to guide the AI tool's evaluation process.
- **AI tool implementation:** Each study publication was individually uploaded to the closed-system AI tool. A single standardized prompt was used, based on the Drummond checklist, to evaluate the quality of the studies. The AI tool analyzed the content of each study and generated responses to the checklist questions. These responses included detailed justifications supported by verbatim excerpts from the study texts.
- **Comparative analysis:** To assess the AI tool's performance, its responses were directly compared to human reviewers' assessments. Agreement rates were calculated for each checklist item and overall across all 8 studies. Agreement was expressed as a percentage, calculated by dividing the number of matching responses by the total number of responses, then multiplying by 100.
- **Workflow overview:** The workflow for this study is summarized in **Figure 1**, which outlines the process from the selection of study publications to the comparison of QAs between the AI tool and human reviewers.

Methods (cont'd)

Figure 1. Workflow summary: AI-assisted QA process



Results

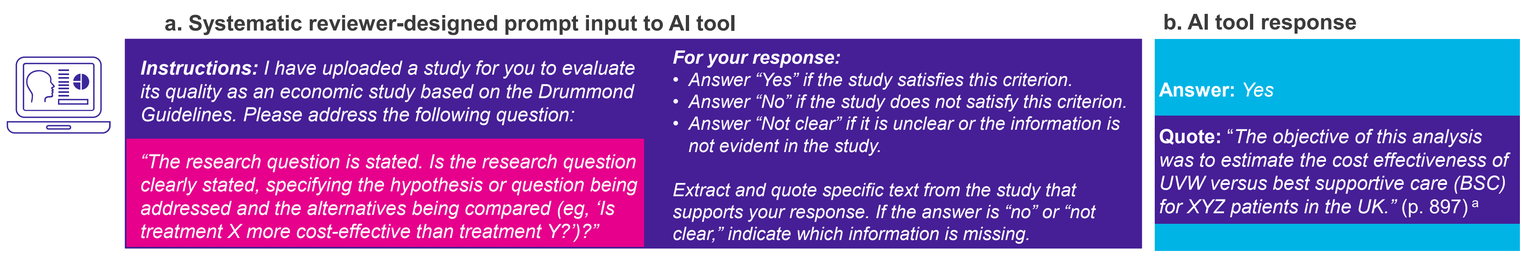
- **Overall quality of included studies:** Of the 8 economic evaluations assessed, 6 demonstrated strong methodological quality, meeting the majority of Drummond checklist criteria, with less than 3 items either unmet or insufficiently reported. In contrast, 2 studies exhibited more substantial limitations, failing to meet 9 and 10 criteria, respectively. Recurring issues across studies included limited reporting on the characteristics of subjects from whom valuations were obtained (2 studies), failure to report quantities of resource use separately from unit costs (3 studies), and inadequate details on statistical tests and confidence intervals (2 studies) and sensitivity analysis approaches (2 studies).
- **Agreement rates:** The AI tool demonstrated strong concordance with human reviewers, with agreement rates ranging from 65.7% to 100% and a median of 94.3%. Agreement was highest in the *Study Design* domain, where responses were nearly identical across all studies. Discrepancies were more frequent in the 2 studies with lower quality, underscoring the difficulty of interpreting ambiguous or incomplete information. **Figure 3** presents a visual summary of agreement and discrepancy rates across the 3 domains of the Drummond checklist.
- **Discrepancies:** Differences between AI-generated and human reviewer assessments were observed in 6 of the 8 studies, with most occurring in the *Data Collection* and *Analysis and Interpretation of Results* domains of the Drummond checklist. Question 19 – which addresses currency adjustments for inflation or conversion – had the highest number of disagreements (n=3). Additional discrepancies were found in Questions 9, 12, 13, 16, and 27 (each with 2 disagreements), covering areas such as effectiveness study details, benefit valuation methods, and sensitivity analysis approaches. Other checklist items had isolated disagreements, reflecting study-specific nuances. **Table 1** presents a detailed summary of agreement rates, along with key areas of divergence across the 3 checklist domains.
- **Sources of discrepancy:** Discrepancies between AI and human reviewer assessments were primarily driven by ambiguous or incomplete reporting within the study texts. Notably, there was no consistent relationship between study design and the level of disagreement; instead, discrepancies were more closely linked to variations in reporting quality and study complexity. The AI tool tended to produce optimistic assessments, often selecting "*Yes*" for checklist items that human reviewers judged as "*No*". In contrast, human reviewers demonstrated stronger contextual interpretation.
- **"*Not applicable*" responses:** The AI tool correctly identified "*not applicable*" responses in most cases, particularly for questions like Question 10 ("*Methods of synthesis or meta-analysis are given*"), Question 14 ("*Productivity changes are reported separately*"), and Question 15 ("*The relevance of productivity changes to the study question is discussed*").
- **Figure 2** provides an illustrative example of the AI tool's assessment process using the Drummond checklist. Panel A shows the tailored prompt crafted by the systematic reviewer to evaluate a specific quality criterion: whether the research question is clearly stated and supported by relevant details. This prompt directs the AI to extract and analyse targeted text from the study publication. Panel B presents the AI's corresponding response, which includes a concise judgment ("*Yes*") accompanied by verbatim excerpts from the study to justify its evaluation.

Table 1. Agreement rates between human and AI reviewers across Drummond checklist domains

Drummond checklist domain (number of questions)	AI-reviewer agreement rate	Key discrepancies (number of disagreements)
Study Design (7 questions)	**96%**	**Q3:** The viewpoint(s) of the analysis are clearly stated and justified (n=1). **Q6:** The form of economic evaluation used is stated (n=1).
Data Collection (14 questions)	**86%**	**Q9:** Details of the design and results of effectiveness study are given (if based on a single study) (n=2). **Q12:** Methods to value benefits are stated (n=2). **Q13:** Details of the subjects from whom valuations were obtained were given (n=2). **Q16:** Quantities of resource use are reported separately from their unit costs (n=2). **Q19:** Details of currency of price adjustments for inflation or currency conversion are given (n=3).
Analysis and Interpretation (14 questions)	**93%**	**Q27:** The approach to sensitivity analysis is given (n=2).

Key: Q – question.

Figure 2. Illustrative example of the AI tool's assessment process

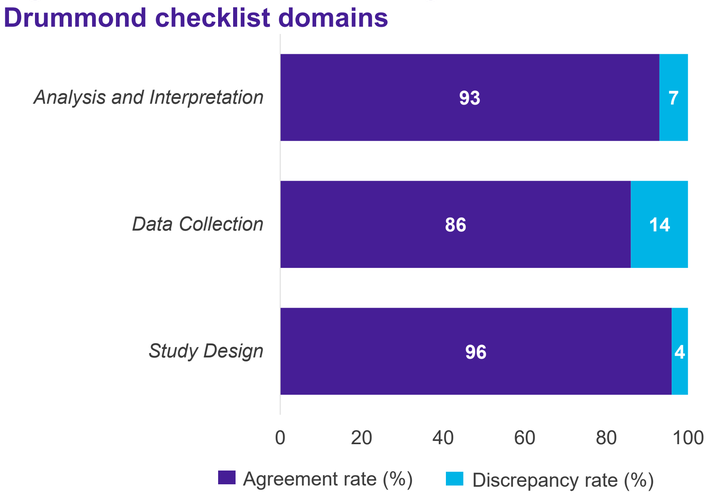


Note: Panel A illustrates the QA prompt designed by the systematic reviewer to evaluate a specific aspect of an economic study using the Drummond checklist. **Panel B** presents the AI tool's corresponding response, including its judgment and supporting excerpts from the study text.
[a] Specific details of the treatment (eg, "UVW") and patient population (eg, "XYZ") have been generalized in this figure to protect proprietary information and ensure compliance with copyright and confidentiality standards.

Conclusions

- This study demonstrated the strong potential of an AI-powered tool to streamline the QA of economic evaluations using the Drummond checklist. The AI tool exhibited high agreement rates with human reviewers. Its ability to generate detailed, evidence-based responses significantly reduced the time required for QAs compared to traditional manual methods.
- While the AI tool excelled at identifying and extracting reported information, it faced limitations in interpreting ambiguous or incomplete data; areas where human reviewers provided essential contextual judgment. Discrepancies were primarily driven by reporting quality, reinforcing the need for human oversight in evaluating complex or nuanced content.
- Ultimately, AI-powered tools should complement, not replace, human reviewers. By accelerating the QA process while preserving expert judgment, AI can support healthcare decision-makers in conducting robust evaluations of economic evidence, paving the way for informed policy and funding decisions.

Figure 3. AI vs human reviewer agreement across Drummond checklist domains



Note: This figure illustrates the percentage of agreement between the AI tool and human reviewers across 3 question domains of the Drummond checklist: *Study Design, Data Collection*, and *Analysis and Interpretation of Results*. Agreement rates are expressed as a percentage based on the concordance of AI and human responses to QA questions across 8 studies.

References: 1. Turner H, Archer R, Downey L, et al. An introduction to the main types of economic evaluations used for informing priority setting and resource allocation in healthcare: Key features, uses, and limitations. *Front Public Health*. 2021;9:722927. https://doi.org/10.3389/fpubh.2021.722927 **2.** Drummond MF, Sculpher MJ, Torrance GW, O'Brien BJ, Stoddart GL. *Methods for the Economic Evaluation of Health Care Programmes*. 3rd ed. Oxford University Press; 2005. **3.** 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. **4.** O'Connor AM, Tsafnat G, Gilbert SB, et al. Automation and artificial intelligence in systematic reviews: an evaluation of the evidence. *Syst Rev*. 2024;13:82. https://doi.org/10.1186/s13643-024-02682-2 **5.** National Institute for Health and Care Excellence (NICE). Use of AI in evidence generation: NICE position statement. August 2024. https://www.nice.org.uk/position-statements/use-of-ai-in-evidence-generation-nice-position-statement **6.** Cochrane, Guidelines International Network, Campbell Collaboration. RAISE: Responsible AI Use in Evidence Synthesis. June 2025. https://www.cochrane.org/events/recommendations-and-guidance-responsible-ai-evidence-synthesis

Presented at: ISPOR Annual European Congress
9-12 November 2025 | Glasgow, Scotland, UK
Funded by Cencora