Harnessing AI for Quality Assessment in Economic Literature Reviews: Development and Testing of Drummond Checklist Prompts

RWD310

Lim A,[1] Yin C,[1] Evans JS[1]

[1]Costello Medical, Singapore

Objective

To develop and test Artificial Intelligence (AI) prompts for conducting quality assessments of economic evaluations using the 10-item Drummond checklist, and to compare the accuracy of AI versus human assessments.

Background

- Literature reviews are a promising area for AI application because of the structured tasks that AI is able support with.[1, 2]
- While prior studies show promise in AI-assisted data extraction, evidence for AI application in quality assessment of economic evaluations remains limited.[2]
- Given the critical role of quality assessment in literature reviews, this study aimed to evaluate whether carefully engineered prompts can enable GPT-4o to assess methodological quality of economic evaluations using the 10-item Drummond checklist and compared its performance with human reviewers.[3]

Methods

- An overview of the project approach is provided in **Figure 1**.
- All AI quality assessments were conducted using the OpenAI GPT-4o model.
- Human data extractions, used as the reference standard, were quality checked by a second human reviewer across the development and test phases, in alignment with guidance from the University of York's Centre for Reviews and Dissemination.[4]

Development Phase

- Prompts based on the 10-item Drummond checklist were iteratively refined using both a context prompt and the full publication text (or abstract if no full-text was available) as input.
- These prompts were applied to a development set of four non-small cell lung cancer (NSCLC) economic evaluation articles, including two full-text journal articles and two conference abstracts/health technology assessment (HTA) reports.
- After each iteration, F1 scores (harmonic mean of precision and recall; range 0–1) were calculated to evaluate AI performance, with refinement continuing until the model achieved an F1 score of ≥0.70, with subsequent iterations conducted until a lower score than the previous iteration was observed, at which point refinement was stopped.

Test Phase

- The best-performing prompt was applied to an independent test set of five NSCLC articles (three full-text journals and two conference abstracts.
- F1 scores from AI-generated quality assessments were compared to those of human quality assessments of the same articles.
- A detailed item-level analysis was conducted to evaluate AI accuracy for each question in the Drummond checklist.

Results

- AI performance improved through prompt refinement and was maintained in the test phase, with F1 scores ≥0.70 (**Figure 2A**).
- In the development set (n=4), AI F1 scores ranged from 0.77 to 0.84, compared to 0.91 for human assessments.
- Applying the best performing prompt (prompt 4) to the test set (n=5), AI achieved an F1 score of 0.78 versus 0.90 for humans.
- AI demonstrated relatively stronger performance on full-text journal articles, producing results that approached human reviewer accuracy. However, it provided less accurate evaluations for conference abstracts and HTA reports, a pattern consistent with observations in human assessments (**Figure 2B** and **2C**).
- Analysis of individual Drummond checklist items revealed that AI errors were predominantly inaccurate (false positives), where the model fabricated outputs that were not present in the articles (AI hallucinations) (**Figure 3**).
- Notably, the AI had challenges connecting information spread across an article, sometimes overlooking relationships between key details and assumptions and missing their full context, which impacted its overall assessments.

Conclusion

The AI model showed promise for conducting quality assessments using the 10-item Drummond checklist, particularly when applied to full-text journal articles.

However, given the lower F1 scores observed for AI quality assessments compared to humans and the persistence of false positives, maintaining a human-in-the-loop approach is essential to ensure accuracy. Future research should focus on broader validation across multiple disease areas and consider evaluating other variants of the Drummond checklist to enhance applicability and performance.

FIGURE 1

Summary of Project Approach

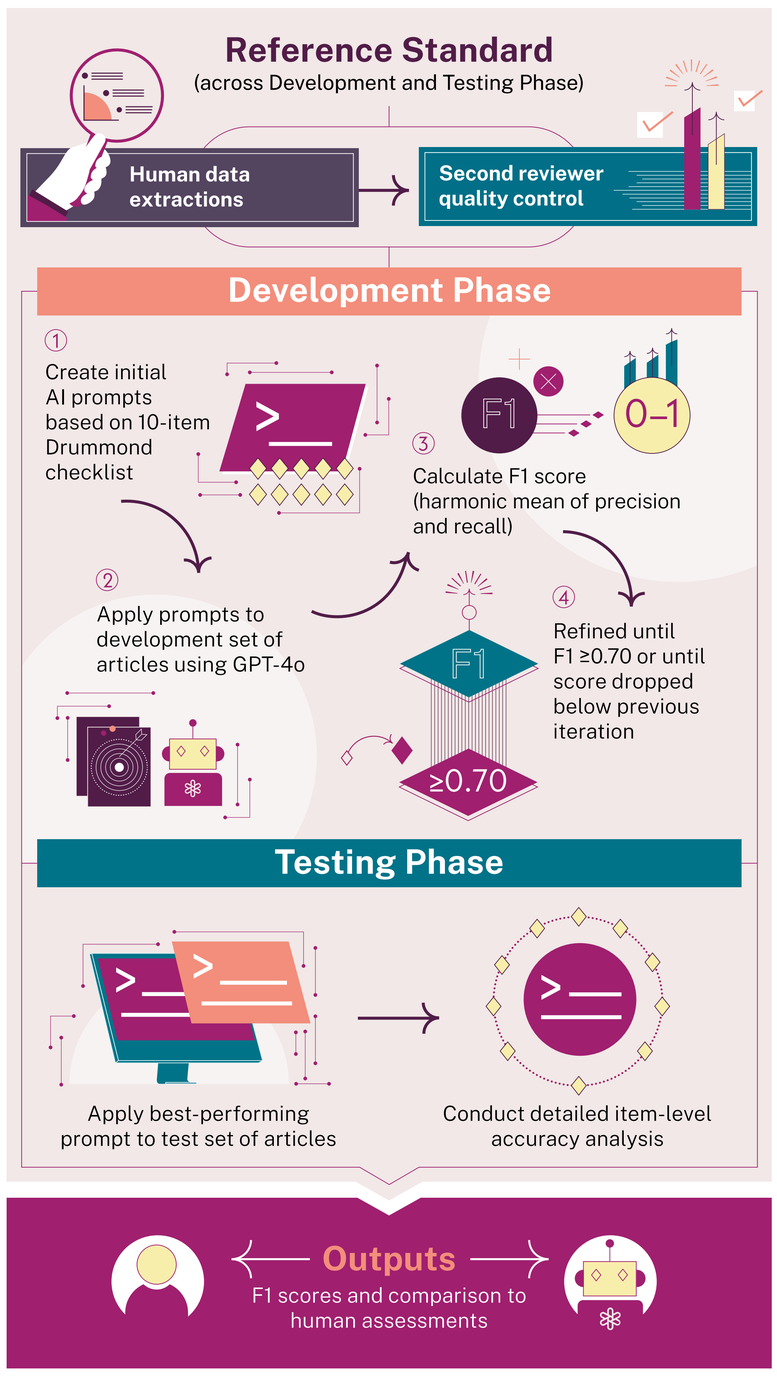
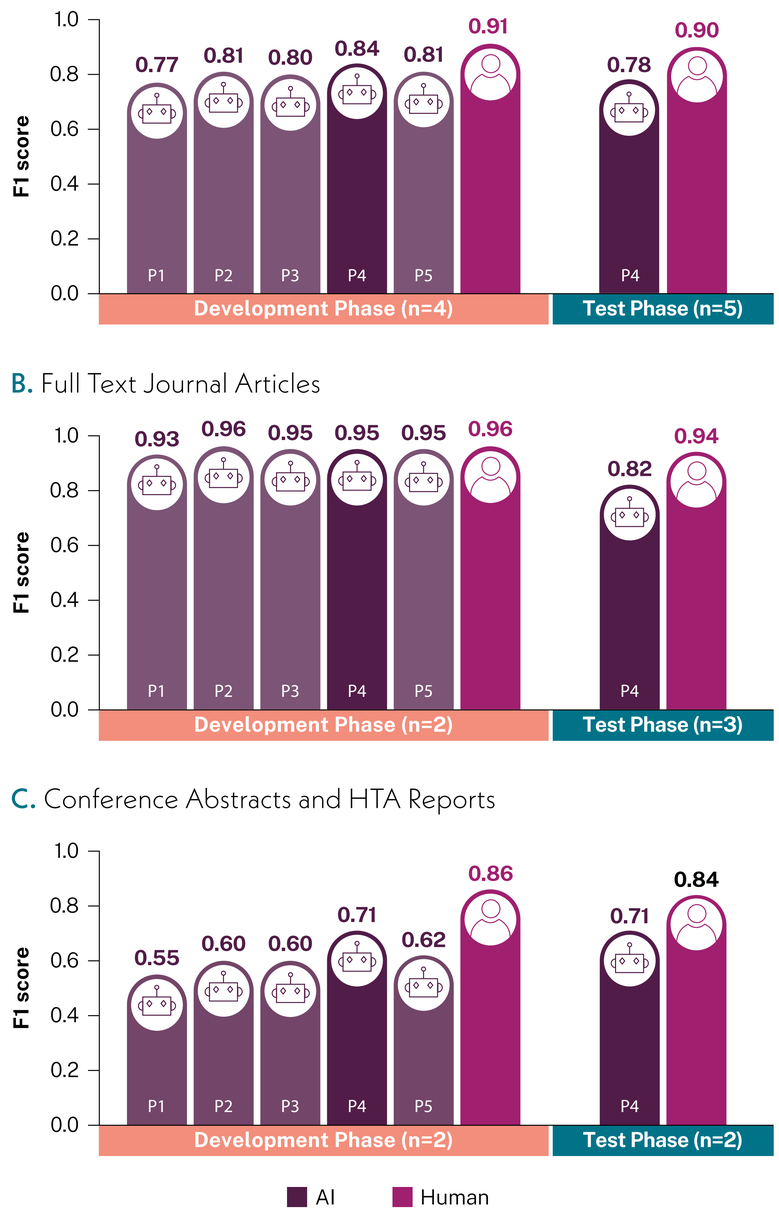


FIGURE 2

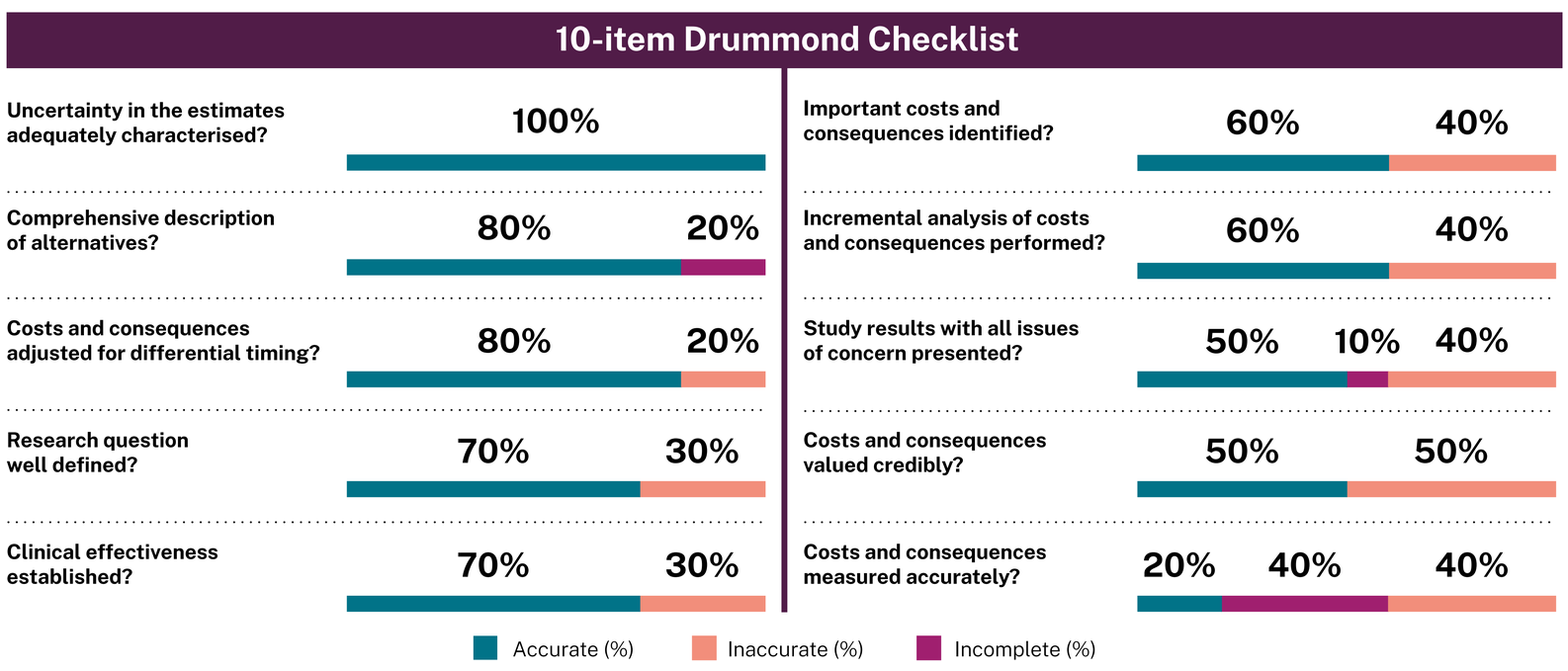
Comparison of F1 Scores Between AI and Human Assessments



Prompt iterations (P1–P5) were evaluated during the development phase; P4 achieved the highest F1 score (Fig. 2A) and was carried forward to the test phase.

FIGURE 3

Item-Level Analysis of AI Assessment (Test Set, n=5)



This analysis was conducted on the test set (n=5 articles) assessing each item of the 10-question Drummond Checklist. AI assessments were classified as accurate when the model correctly identified information present in the article; assessments were classified as inaccurate when the AI identified information absent from the article (false positives); instances where the AI failed to detect information that was present were classified as incomplete (false negatives).

Abbreviations: AI: artificial intelligence; **HTA:** Health Technology Assessment; **NSCLC:** non-small-cell lung cancer; **P:** prompt.

References: [1]Khraisha Q. et al. Res Synth Methods. 2024;15(4):616–626; [2]Scherbakov D. et al. J Am Med Inform Assoc. 2025;32(6):1071–1086; [3]Drummond M. et al. Methods for the economic evaluation of health care programmes. 2015. Available from: https://nibmehub.com/opac-service/pdf/read/Methods%20for%20the%20Economic%20Evaluation%20of%20Health%20Care%20Programmes.pdf. [Last accessed: 04 Sep 2025]; [4]University of York. CRD's guidance for undertaking reviews in health care. 2009. Available from: https://www.york.ac.uk/media/crd/Systematic_Reviews.pdf. [Last accessed: 29 Aug 2025]. **Acknowledgements:** The authors thank Henny Zhan, Costello Medical, for graphic design assistance.

Presented at ISPOR Real-World Evidence Summit 2025 | Tokyo, Japan | 28–30 September