MSR187

Quality Assessment in a Systematic Literature Review Using an Artificial Intelligence Model

Martin P,[1] Di Risio M,[1] Colman O,[2] Gregory C,[2] Hughes L,[2] Lunn L[3]

[1]Knight Therapeutics Inc., Montréal, Quebec, Canada; [2]Costello Medical, Boston, MA, US; [3]Costello Medical, Manchester, UK

Objective

To evaluate artificial intelligence (AI) quality assessment (QA) performance for randomized controlled trials (RCTs) using a modified Cochrane risk of bias (RoB) 1.0 tool.[1]

Background

- Assessing study quality is a critical component of systematic literature reviews (SLRs), yet it remains a highly manual, resource-intensive step.
- AI models offer new opportunities to increase efficiency of QA, but their performance in evaluating nuanced RoB domains, and their susceptibility to misinterpretation, are not well characterized.
- We evaluated an AI model using a high-volume, methodologically diverse set of attention deficit hyperactivity disorder (ADHD) RCTs, varying in population, intervention, and study design, to understand its accuracy in RoB assessment.

Methods

- AI prompts were developed to assess study quality using the Cochrane RoB 1.0 tool,[1] with additional questions addressing cross-over designs and study funding. Prompts were refined for best output by human judgement and run in GPT-4o (temperature: 0.7) for all questions (**Figure 1**).
- A human verified AI outputs against publications to indicate a true positive (accurate, as reported in publication), false positive (data reported which are not present in publication), or false negative (data not reported which are present in publication). Results were used to compute recall (relevant data accurately identified [score range 0–1]) and precision (correct outputs [score range 0–1]).
- F1 scores (harmonic mean of precision and recall [score range 0–1]) were calculated. A predefined threshold of ≥0.70 was considered a 'good' F1 score.

Results

- Among 32 studies, median recall was 1.0 (range: 0.31–1.0), and median precision was 0.68 (range: 0.31–0.94). Most studies (91% [29/32]) had a 'good' F1 score of at least 0.70. Two studies scored close to the threshold, with F1 scores of 0.63 and 0.67. The remaining study score (for a clinicaltrials.gov record) was an outlier at 0.31 (**Figure 2**). This likely arose from limitations in the AI model's ability to interpret and preserve table structure, with rows and column misalignment during extraction resulting in incorrect interpretations.
- The highest question-specific F1 score was 0.97 for the checklist item assessing whether treatment groups were balanced in prognostic factors at baseline. The lowest F1 score was 0.51, which was recorded for whether the analysis used an intention-to-treat (ITT) population (**Figure 3**).
- Overall, lower F1 scores were primarily driven by the occurrence of false positives, or 'hallucinations' (29% of responses), rather than false negatives (3% of responses).

Conclusion

The AI model demonstrated high recall, efficiently identifying QA checklist items when present. Precision varied across records, reflecting common misinterpretations of data. The frequency of false positives underscores the importance of careful human review of all output. Although most studies met the F1 threshold, AI appeared less consistent in extracting and interpreting data from a clinical trial record than manuscript publications. These findings support the value of AI in QA of RCTs but emphasize the need for continuous human-in-the-loop verification to ensure accuracy.

FIGURE 1

Summary of prompt engineering and testing process

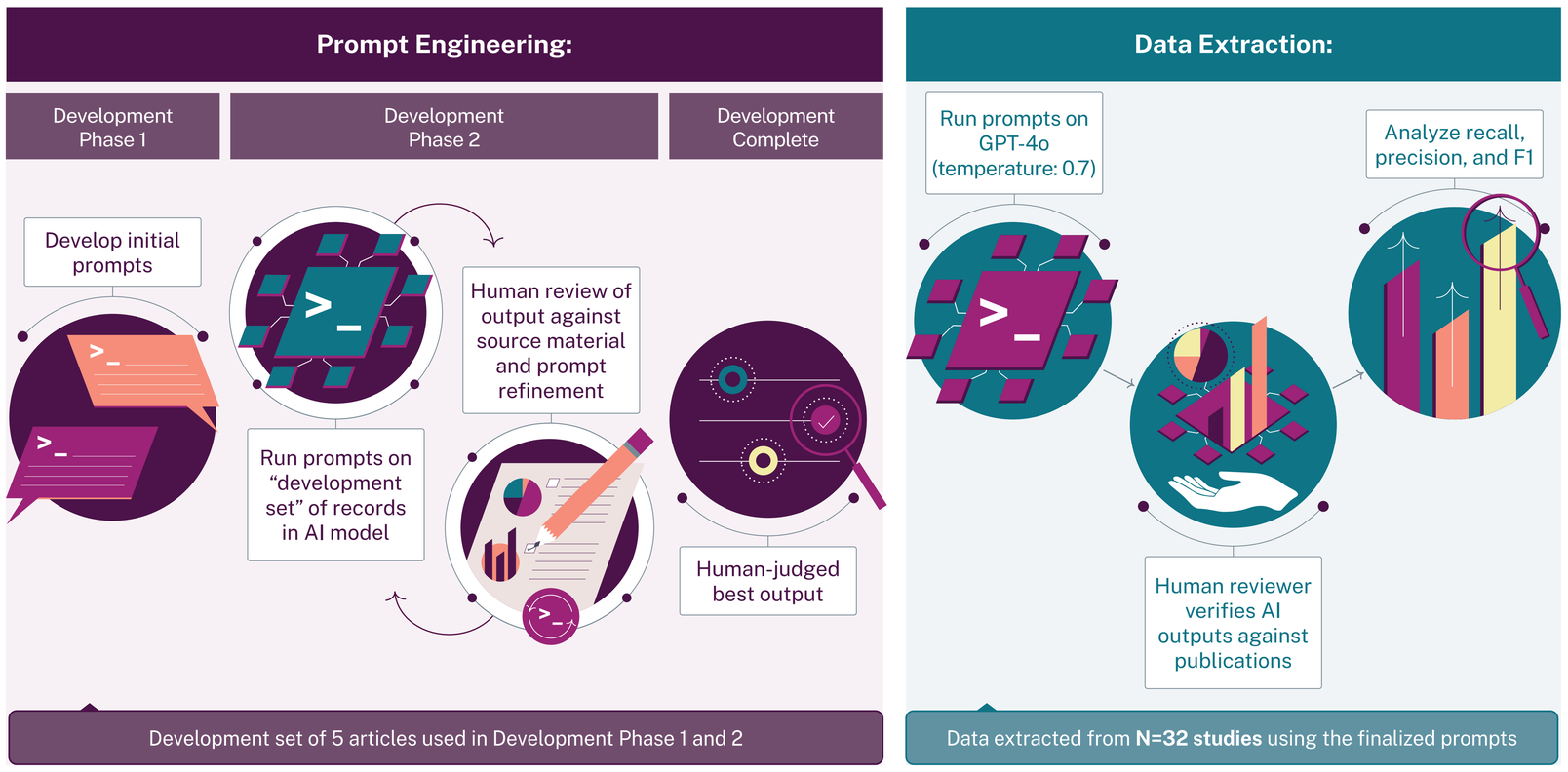


FIGURE 2

F1 score by study

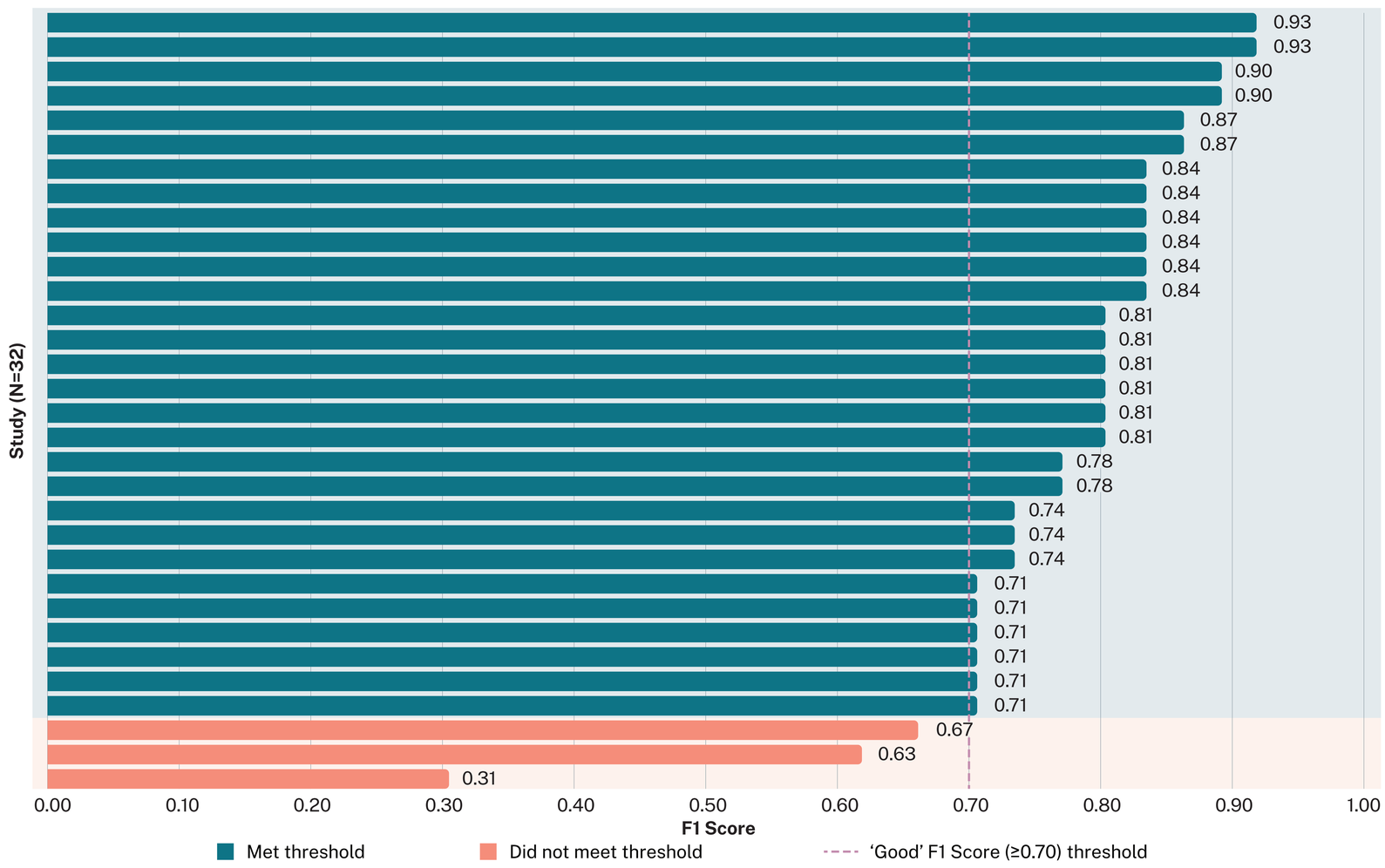
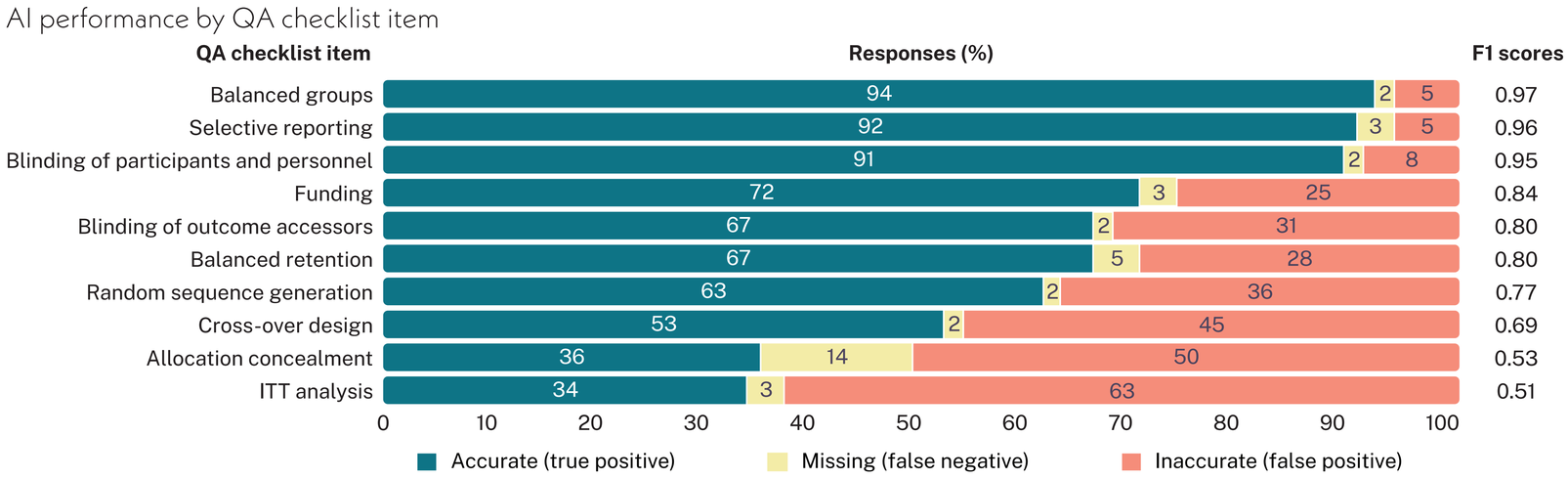


FIGURE 3

AI performance by QA checklist item



Footnote: Each checklist item comprises of a multiple-choice question and free response details.

Abbreviations: ADHD: attention deficit hyperactivity disorder; **AI:** artificial intelligence; **ITT:** intention-to-treat; **QA:** quality assessment; **RCT:** randomized controlled trial; **RoB:** risk of bias; **SLR:** systematic literature review.

References: [1]Higgins JP. et al. BMJ 2011;343:d5928

Acknowledgements: The authors thank Sanjana Prakash, Costello Medical, for graphic design assistance. We also thank Steph Beaver and Sophie Schoeni for review and editorial assistance in the preparation of this poster.

Presented at ISPOR International 2026 | Philadelphia, PA, US | 17–20 May