Evaluation of Machine Learning–Assisted Data Extraction for Supporting Systematic Reviews: A Case Study of a Review in Movement Disorders

SA11

Michaela Lunan-Taylor; Sathushan Thurairajah; Jose S. Marcano-Belisario; Louise Hartley

RTI Health Solutions, Manchester, United Kingdom

BACKGROUND

- Systematic literature reviews (SLRs) constitute a foundational methodology in health economics and outcomes research, underpinning evidence-based decision-making.
- As the demand for timely, efficient, and reliable evidence generation has increased, artificial intelligence (AI) technologies have been progressively integrated to support and automate key stages of the SLR process. Among these stages, data extraction remains one of the most time- and resource-intensive components of evidence generation.
- Consequently, AI data extraction needs to be rigorously tested to ensure the robustness, transparency, and reproducibility of systematic reviews, as well as their appropriate integration within established workflows.

OBJECTIVE

- Our objective was to evaluate the use of AI-assisted data extraction in a clinical systematic review of randomized-controlled trials (RCTs) in movement disorders, while considering consistency (measured by accuracy, recall, precision and F1), time spent, and future implications for SLR researchers.

METHODS

- We conducted a clinical systematic review of RCTs in movement disorders and data extraction was performed within the AI-assisted evidence synthesis tool, Nested Knowledge, using Smart Meta-Analytical Extraction functionality.
- Prompts for data extraction were iteratively developed, piloted, and finalized prior to full extraction. The prompts covered 5 predefined data domains: study characteristics, treatment description, baseline patient characteristics, clinical outcomes, and safety outcomes.

See Figure 1

- Data were extracted across narrative text, tables, and figures from publications, abstracts, and ClinicalTrials.gov records.
- AI-generated extractions were quality checked by human researchers, who corrected errors and supplemented missing data.
- For studies reported across multiple publications, data were collated at the study level to remove duplication.
- True positives = correct extraction from any of the included references for a study. False positives = incorrect data. False negatives = missed or incomplete data. True negatives = appropriately unreported.
- Time spent on prompt development, piloting quality control (QC) checks, and supplementation was tracked and compared with typical manual extraction times per article.

See Figure 2

References

1. Wagner G, et al. J Inf Technol. 2022;37(2): 209-26.
2. Helms Andersen T, et al. Cochrane Evid Synth Methods. 2025;3(4):e70036.
3. Jackson B, et al., 2025. ISPOR - Leveraging AI for Evidence Synthesis: Assessing the Accuracy and Efficiency of AI-Supported Data Extraction and Reporting.

The power of **knowledge**
The value of **understanding**

RESULTS

SLR Results

The SLR included 39 references, corresponding to 24 unique studies included for data extraction.

AI Performance

AI-assisted data extraction demonstrated high accuracy, recall, and precision for study characteristics and baseline patient characteristics. In contrast, extraction performance for clinical outcomes and safety data was characterized by substantially lower recall, indicating a higher frequency of missed relevant data rather than incorrect extractions. Extraction performance metrics by data domain are summarized in **Table 1.**

Table 1. Accuracy, Recall, Precision, and F1 of AI Data Extraction by Data Domain

Category	Study characteristics	Treatment description	Patient characteristics	Outcomes	Safety
Accuracy	0.94	0.80	0.93	0.90	0.47
Recall	0.99	0.84	0.93	0.40	0.31
Precision	0.95	0.93	0.88	0.59	0.71
F1	0.97	0.89	0.90	0.47	0.43

Table 2. Average Time for Data Extraction by Extraction Approach and Workflow Stage

Category	Template data extraction sheet (human)/prompt building and piloting (AI)	Extraction	QC	Additional QC of human extracted data	Merging individual article data to a single study
Average human extraction	5 hours total	1 hour per article	30 minutes per article	NA	NA (completed during data extraction)
AI extraction	12 hours total	NA	38 minutes per article	33 minutes per article	5 minutes per linked article

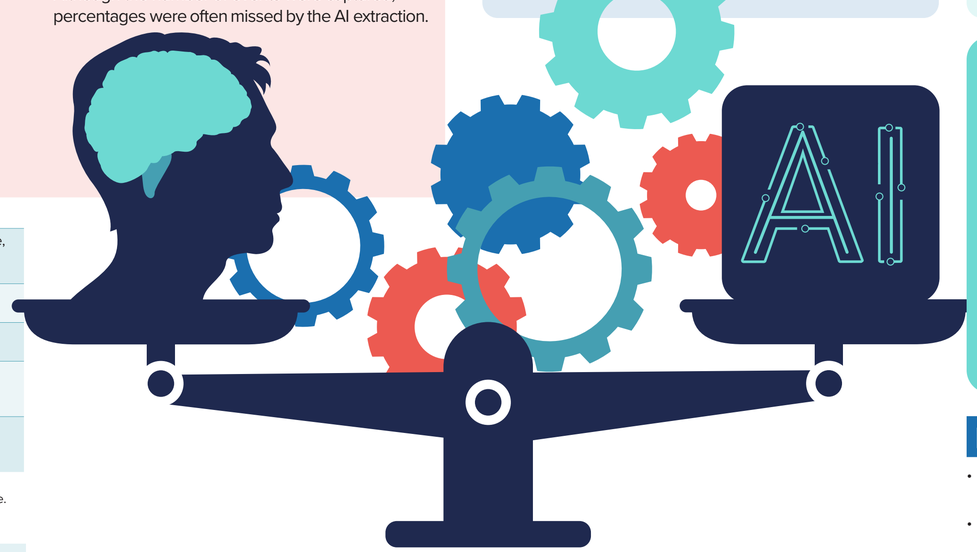
NA = not applicable.

Issues/Faults

- Merging individual treatment arm data.
- Difficulty correctly identifying treatment arms in crossover study designs.
- Incorrectly categorizing ethnicity data in patient characteristics despite the availability of both tabular and textual data within the publications.
- Incorrect handling of summary statistics, with means and medians frequently interchanged.
- Missing outcomes and safety data.
- Despite the presence of relevant data in both tables and narrative text, the AI often misclassified AEs, such as incorrectly categorizing headache data as dysphagia or choking as respiratory infections.
- Misclassification of tools (e.g., pain numerical rating scale was extracted as pain visual analog scale and TWSTRS as a generalized severity scale). The AI demonstrated a tendency to fixate on specific lexical items, such as "pain," and subsequently link the associated data to the first outcome in which that term appeared.
- Although the number of events were captured, percentages were often missed by the AI extraction.

Unanticipated Benefits

- Finding details in areas where humans would typically miss (e.g., footnotes or abstracts).
- Automatically converting units (e.g., days to months).



Timing (Table 2)

- AI-assisted extraction was faster initially but required significant time for prompt development, piloting, and QC checks.
- Prompt development took ~12 hours (vs. ~5 for manual templates); QC took ~71 minutes/article (vs. ~30 manually).
- Additional time (~5 minutes/article) was needed to merge linked publications.
- Overall, total time per article was similar: ~92 minutes (AI + QC) versus ~98 minutes (manual).

Study characteristics	17 data points, including study type, blinding details, study phase, and geographical location
Treatment description	7 data points, including treatment name, dose, and schedule
Baseline patient characteristics	29 data points, including age, sex, and ethnicity
Clinical outcomes	82 data points, including TWSTRS score, pain VAS, and SF-36
Safety	17 data points, including total number of AEs and specific AEs (e.g., headache and dysphagia)

AE = adverse event; SF-36 = SF-36 Health Survey; TWSTRS = Toronto Western Spasmodic Torticollis Rating Scale; VAS = visual analog scale.

Recall
Proportion of relevant data that is correctly identified by the AI out of all relevant data available

$$\frac{\text{True positives}}{\text{(True positives + False negatives)}}$$

Precision
Proportion of relevant data among that identified by the AI

$$\frac{\text{True positives}}{\text{(True positives + False positives)}}$$

Accuracy
Proportion of correctly identified data out of all data captured by the AI

$$\frac{\text{(True positives + True negatives)}}{\text{True positives + True negatives + False positives + False negatives}}$$

F1 score
Harmonic mean of precision and recall

$$\frac{2 \times (\text{Precision} \times \text{Recall})}{\text{(Precision + Recall)}}$$

Key Implications for Health Technology Assessment (HTA) Practice

- **AI currently functions best as a support tool rather than a replacement for human data extraction in HTA relevant systematic reviews.** Performance was strong for structured, objective data, but complex outcomes and safety data required substantial human oversight.
- **Low recall for clinical outcomes and safety endpoints poses a material risk for HTA decision-making.** Missed or misclassified data may compromise evidence completeness if AI outputs are not rigorously validated.
- **Evaluation of AI adoption should consider the full evidence synthesis workflow.** Initial speed gains from AI extraction were largely offset by time required for prompt development, quality checking, and reconciliation of linked publications.
- **AI may add complementary value by enhancing data completeness and standardization.** Benefits such as identifying less prominent data sources and automatic unit conversions suggest potential gains when AI is integrated with human review.
- **Targeted use of AI may be most appropriate in reviews with highly standardized outcomes.** Reviews with heterogeneous outcomes, complex designs, or safety-critical endpoints are likely to require sustained human involvement.

DISCUSSION

- AI-assisted data extraction demonstrated strong performance for objective and atomic data elements.[1-3] However, performance for clinical outcome and safety data was substantially lower, driven primarily by low recall associated with missed and misclassified data points. This reinforces the necessity of a human-in-the-loop approach.
- Errors (e.g., outcome tools, adverse events) suggest reliance on surface cues rather than deep context, indicating limits of prompt design alone.
- Despite limitations, AI showed strengths in objective extraction and identified some overlooked data, supporting its role as a complementary tool.
- Time savings were minimal overall, as gains in speed were offset by effort in prompting, piloting, and QC, highlighting the need to assess end-to-end workflow impact, not just extraction speed.

LIMITATIONS

- This SLR focused on a rare disorder with heterogeneous outcomes; AI may perform better in areas with highly standardized outcomes and reporting.
- Time comparisons used typical manual averages rather than review-specific data due to feasibility constraints.
- Publisher restrictions on AI use may limit broader adoption in systematic reviews.
- AI is constantly improving and adapting; results may differ if this review were performed today.

CONCLUSION

- Overall, the findings strongly support a human-in-the-loop approach, where AI acts as an assistive technology rather than an autonomous extractor. Human oversight was essential in the extraction process, not only for error correction but also for resolving linked studies and ensuring data coherence across publications.
- From a methodological perspective, these results underscore the importance of evaluating AI-assisted extraction across complete systematic review workflows rather than focusing solely on extraction accuracy or speed.

Contact: Michaela Lunan-Taylor (mlunan@rti.org)

Presented at: ISPOR 2026; May 17-20, 2026; Philadelphia, PA, USA