Artificial Intelligence for Rapid Data Extraction: A Case Study on the Economic and Caregiver Burden of Motor Neuron Disease

SA8

Kim Wager,[1] **Tomas Rees**,[1] **Maxandre Jacqueline**,[1] **Jungwon Byun**[2]

[1]Oxford PharmaGenesis, Oxford, UK; [2]Elicit PBC, Oakland, CA, USA

Scan here to download the poster

ISPOR main topic/taxonomy: Study Approaches
ISPOR subtopics: Literature Review & Synthesis, Artificial Intelligence, Machine Learning

BACKGROUND

- Systematic identification of evidence is a fundamental principle of health economic and outcomes research, but systematic literature reviews (SLRs) can be time-consuming to develop.
- Artificial intelligence (AI) tools may be able to improve the efficiency of SLRs while maintaining accuracy, including for the data extraction stage of the review; however, rigorous evaluations comparing AI performance with human performance for data extraction are limited, especially when considering diverse data domains.
- In this study, we assessed the performance of an AI research platform (Elicit) on a review related to motor neuron disease (MND).

METHODS

- We designed a semantic literature search and used AI-assisted screening against agreed PICO (Population, Intervention Comparison, Outcome) criteria to identify relevant publications in MND, spanning diverse data types including clinical characteristics, heterogeneous cost categories and multidimensional caregiver burden measures.
- We performed data extraction across 167 publications using 15 predefined data fields spanning study characteristics, economic components and caregiver burden measures.
  - Highly detailed, structured prompts were iteratively developed to guide extractions of the required depth using a training set excluded from the analysis.
- For AI data extraction, we used Elicit, a large language model (LLM)-based literature review platform, which enables interactive title and abstract screening and data extraction from full text with sentence-level citations.
- AI performance was compared against a human ground-truth subset of these (N = 21 open access publications) using a 4-dimension scoring rubric assessing completeness, accuracy, detail level and data integrity (0–4 points each, maximum 16 points per field), which were combined into an overall performance score, given as a percentage.
  - Completeness: Did AI attempt to get all the relevant information?
  - Accuracy: For information extracted from the right location, was it correct?
  - Detail level: Was the right amount of detail provided?
  - Data integrity: Was the information from the correct location in the paper and not fabricated?
- To avoid double-penalizing AI for the same error, completeness was interpreted as capturing what information was present or attempted, while detail level assessed the depth of reporting.
- Performance patterns were analysed across all 15 fields to identify AI strengths and limitations across three different data extraction domains: study characteristics, economic measures and caregiver measures.
- Additionally, for each field, we rated the human ground truth answer on the same 4-dimension rubric and used this to categorize the AI answers as superior, equivalent or inferior to the human extraction.

RESULTS

Overall AI performance

- AI demonstrated strong overall performance across data extraction tasks, with mean scores ranging from 77.7% to 95.2% across the 15 fields evaluated (**Figure 1**).
- Overall mean AI performance was high (≥ 90%) for 11 of 15 extraction fields, moderate (80–89%) for 3 fields and lower (< 80%) for 1 field (sample size) (**Figure 1**).
- Most extractions (80.6%) achieved high performance scores (≥ 80%), with 57.1% scoring perfectly (100%) and failures (< 40%) occurring in only 1.3% of cases (**Figure 2**).

Performance by domain

- Accuracy scores were consistently the strongest component, followed closely by completeness, while detail level and data integrity showed greater variability (**Figure 3**).

Caregiver measures

- Caregiver measures represented the strongest performance domain for AI, with all 5 caregiver measures achieving mean performance scores of over 91.4% (caregiver costs: 94.9%, burden scores: 94.6%, quality of life [QoL]: 92.9%, hours: 92.9%, work impact: 91.4%) (**Figure 1**).

Economic measures

- Economic measures showed high mean performance scores, with cost-related measures ranging from 85.7% to 92.9% across indirect costs (92.9%), direct non-medical costs (92.0%), direct medical costs (90.5%), total annual costs (89.9%) and cost drivers (85.7%) (**Figure 1**).

Study characteristics

- Country identification performed strongly (95.2%), demonstrating the proficiency of AI with clearly defined, structured data elements.
- Data sources and study design achieved mean performance scores of 93.5% and 90.5%, respectively, indicating reliable identification of study methodological details.
- However, more complex variables showed slightly lower performance scores, with MND stage/severity (81.2%) and sample size (77.7%) proving to be the most challenging extraction tasks (**Figure 1**).
- Sample size extraction provides a good example of an extraction task that can be unexpectedly challenging. Papers required determining whether to count surveyed families or affected individuals, distinguishing disease rating scales from staging categories and aggregating subpopulations. These interpretive tasks require contextual reasoning rather than simple extraction. Further prompt instructions specifying decision rules for common ambiguities, or detailed extraction protocols for human reviewers, would address these challenges, reflecting the normal iterative process of extraction protocol refinement in systematic reviews.

Performance variability

- Standard deviations (SDs) varied considerably across extraction fields. Sample size extraction showed the highest variability (SD = 24.1 percentage points [p.p.]), suggesting inconsistent AI performance across different study reporting styles.
  - Conversely, data sources showed the lowest variability (SD = 8.7 p.p.), indicating consistent extraction reliability.

AI versus human comparison

- In 315 head-to-head comparisons (15 fields across 21 papers), AI and human data extraction performed equivalently in 173 instances (54.9%), human extraction was superior in 102 instances (32.4%) and AI extraction was superior in 40 instances (12.7%).
  - Most variables showed high equivalence rates, particularly for caregiver measures and economic costs.
  - AI outperformed humans most often on study characteristics (20.0% of comparisons) versus economic measures (11.4%) or caregiver measures (6.7%).
  - Three variables (indirect costs, caregiver burden score, caregiver QoL) showed no AI advantage (**Figure 4**).

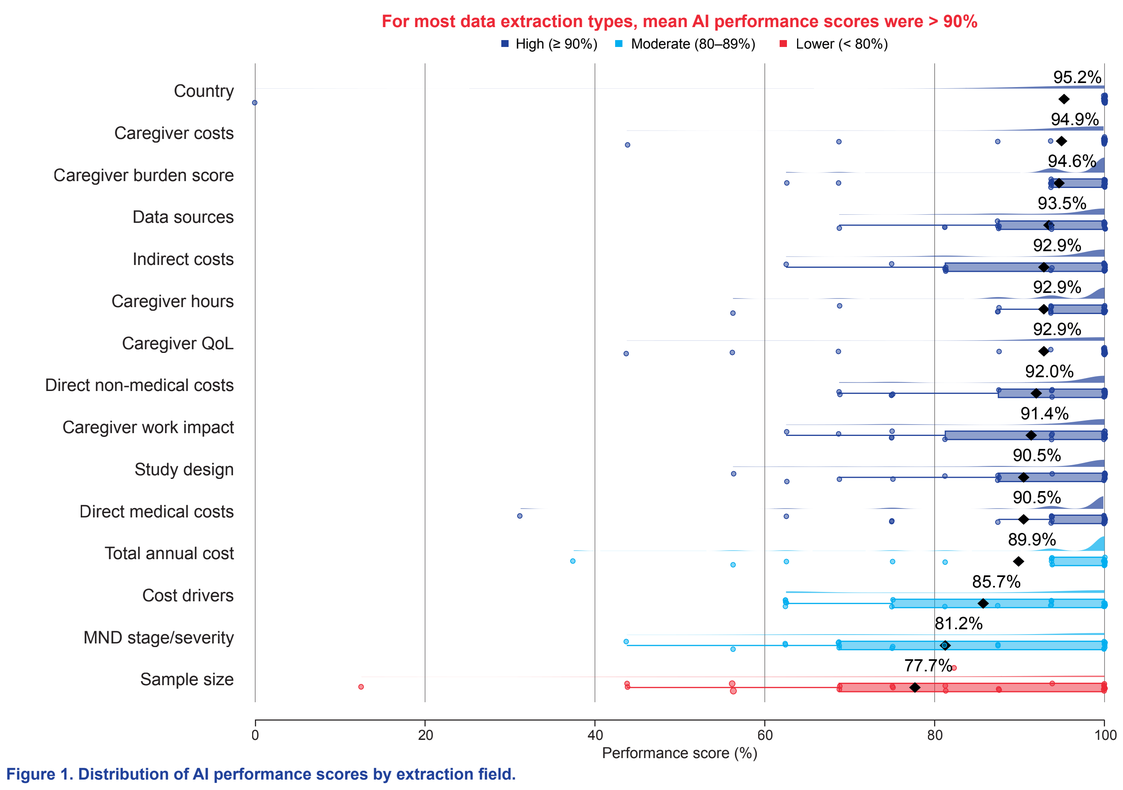

Figure 1. Distribution of AI performance scores by extraction field.

Black diamonds show mean performance scores. AI performance was assessed for completeness, accuracy, detail level and data integrity (0–4 points each, maximum 16 points per field), which were combined into an overall performance score. Boxes represent interquartile range, and density curves show the smoothed distribution of performance scores across papers.

AI, artificial intelligence; MND, motor neuron disease; QoL, quality of life.

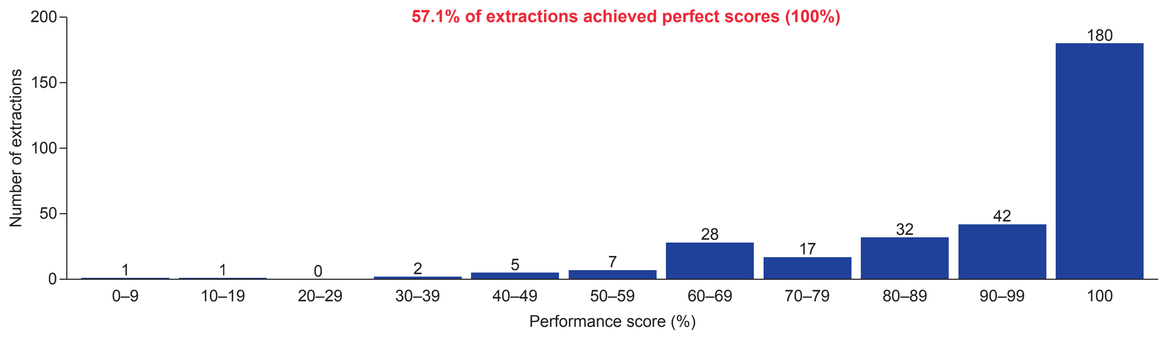

Figure 2. Distribution of AI performance scores.

AI performance was assessed for completeness, accuracy, detail level and data integrity (0–4 points each, maximum 16 points per field), which were combined into an overall performance score.

AI, artificial intelligence.

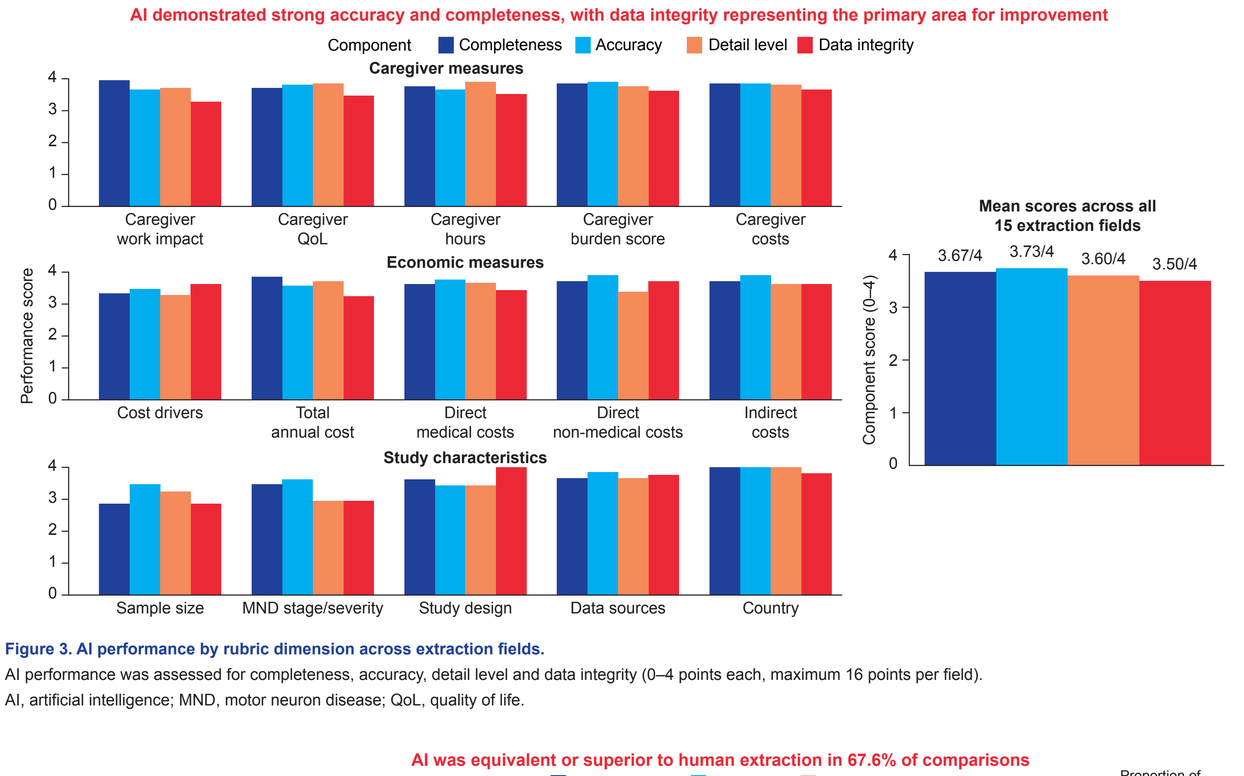

Figure 3. AI performance by rubric dimension across extraction fields.

AI performance was assessed for completeness, accuracy, detail level and data integrity (0–4 points each, maximum 16 points per field).

AI, artificial intelligence; MND, motor neuron disease; QoL, quality of life.

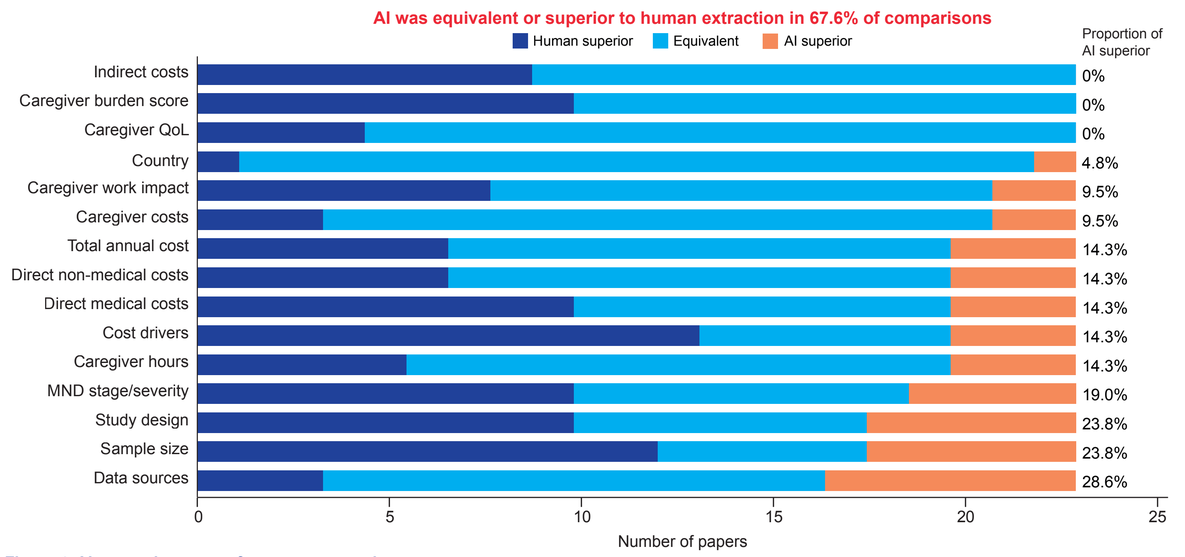

Figure 4. AI versus human performance comparison.

AI, artificial intelligence; MND, motor neuron disease; QoL, quality of life.

DISCUSSION

Performance variability

- The data revealed a pronounced performance ceiling effect, with median scores of 100% for 9 of 15 variables, indicating that AI achieved perfect extraction for most papers on most tasks.
  - However, this high median performance masked substantial variability, and data extraction quality is highly dependent on factors such as document structure, data presentation format and terminology consistency across studies.
- The mean AI performance score of individual papers across all 15 fields ranged from 78.0% to 95.0%, indicating strong performance but a need for robust quality assurance protocols.

Behavioural patterns

- Both human and AI inaccuracies often stemmed from incomplete rather than incorrect extraction.
- Analysis revealed characteristic AI ‘shortcuts’ where the model appeared to simplify complex data presentations.
  - AI frequently reported summary statistics (means) while omitting disaggregated values, percentages without corresponding counts or point estimates without confidence intervals or *p* values.
- This pattern suggests that AI may prioritize efficiency over comprehensiveness when processing complex tabular data, a known challenge for LLMs.

Domain-specific challenges

- Terminology ambiguity or complexity affected extraction accuracy, as demonstrated by the MND stage/severity variable.
  - Such semantic nuances highlight the importance of precise variable definitions in AI-assisted extraction protocols.

Prompt engineering trade-offs

- Structured prompting proved to be a double-edged tool. While scaffolding enhanced detail extraction for conventional research papers, it occasionally constrained AI responses when encountering non-standard publication types such as editorials or ethnographic studies, suggesting the need for adaptive prompting strategies.

Comparative error profiles

- The evaluation revealed distinct error modes between human and AI extraction. While AI performed well for accuracy and completeness, it showed particular vulnerabilities in table and figure interpretation tasks and handling of longitudinal studies, representing key areas for targeted improvement.
- AI sometimes extracted information from inappropriate manuscript sections, representing a limitation in distinguishing supporting context from primary evidence.

Conclusions

- Elicit AI combined with expert prompting performed well but there remains a need for robust quality assurance protocols.
- AI can support data extraction, but we continue to recommend human oversight; all extractions should be checked for missing and misreported extractions.

Presented at International Society for Pharmacoeconomics and Outcomes Research (ISPOR) Europe 2025, 9–12 November 2025, Glasgow, UK