HTA401

AI-Assisted Data Extraction for Systematic Literature Reviews in HTA: Accuracy and Efficiency within a NICE-Aligned Workflow

Praptee Ghimire[1], Ben Lanning, PhD[1], Tejas Pethkar[1], Anna Macbeth[1], Jessica Radford[1], Priscila Mazzola, PhD[1], Jean-Etienne Poirrier, PhD, MBA[2]

[1] Kintiga, United Kingdom, [2]Kintiga, Belgium.

Objectives

Systematic literature review (SLR) data extraction is a resource-intensive step within health technology assessment (HTA), and subject to inter-reviewer variability.[1,2] Recent guidance from the National Institute for Health and Care Excellence (NICE) recognises the potential role of artificial intelligence (AI) in evidence generation, while emphasising the need for transparent, justified, and validated use with appropriate human oversight.[3] This study evaluated the accuracy and efficiency of an automated AI-assisted data extraction tool implemented within a NICE-aligned workflow incorporating human quality control (QC).

Methods

A sample of journal articles was selected across three evidence domains: clinical, economic and quality of life. For each domain, predefined questions were used to extract study characteristics, population characteristics, interventions and comparators, outcomes, safety data, costs and quality-of-life data, as applicable (Figure 1).

The same questions were applied using an AI-assisted extraction tool. The AI-generated outputs were compared with reference extractions produced by human reviewers. Each output was assessed at the data-point level and classified as correct or incorrect. An output was classified as correct if it did not require human re-extraction, although minor editing or reformatting was permitted. Outputs containing missed or inaccurate information that required re-extraction were classified as incorrect.

Accuracy was calculated as the proportion of assessed data points classified as correct, overall and separately for each evidence domain. To assess efficiency, the estimated time required for AI-assisted extraction and human quality control was compared with the estimated time required for a dual-human extraction workflow applied to an equivalent dataset.

Figure 1: Accuracy assessment workflow



Results: Accuracy by evidence domain

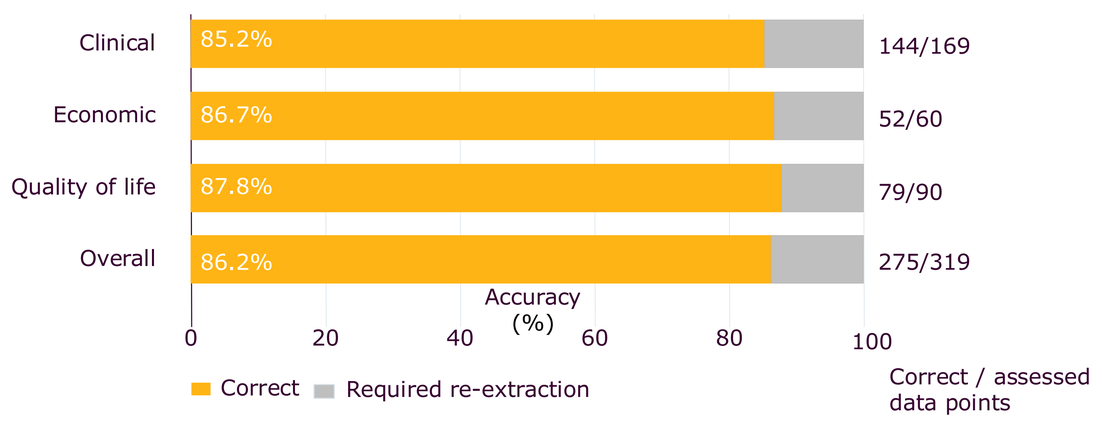
Across 30 publications (Table 1), the AI-assisted tool correctly extracted 275 of 319 data points, corresponding to an overall accuracy of 86.2%. Accuracy was consistent across clinical (85.2%), economic (86.7%) and quality-of-life (87.8%) evidence (Figure 2). The AI-assisted workflow was estimated to reduce extraction-task time by 30%–40% relative to dual-human extraction, despite the inclusion of human quality control. This estimate included the additional time required for human quality control of AI-generated outputs.

Table 1: Evaluation dataset for the assessment of accuracy

Evidence domain	Publications evaluated	Extraction questions assessed	Examples of data fields
Clinical	10	17	Study design, eligibility criteria, population characteristics, interventions, outcomes, adverse events
Economic	10	6	Economic model, population, costs, QALYs, ICERs
Quality of life	10	9	Study design, patient characteristics, QoL instruments, QoL outcomes

Abbreviations: ICER, incremental cost-effectiveness ratio; QALY, quality-adjusted life year; QoL, quality of life

Figure 2: Accuracy by evidence domain



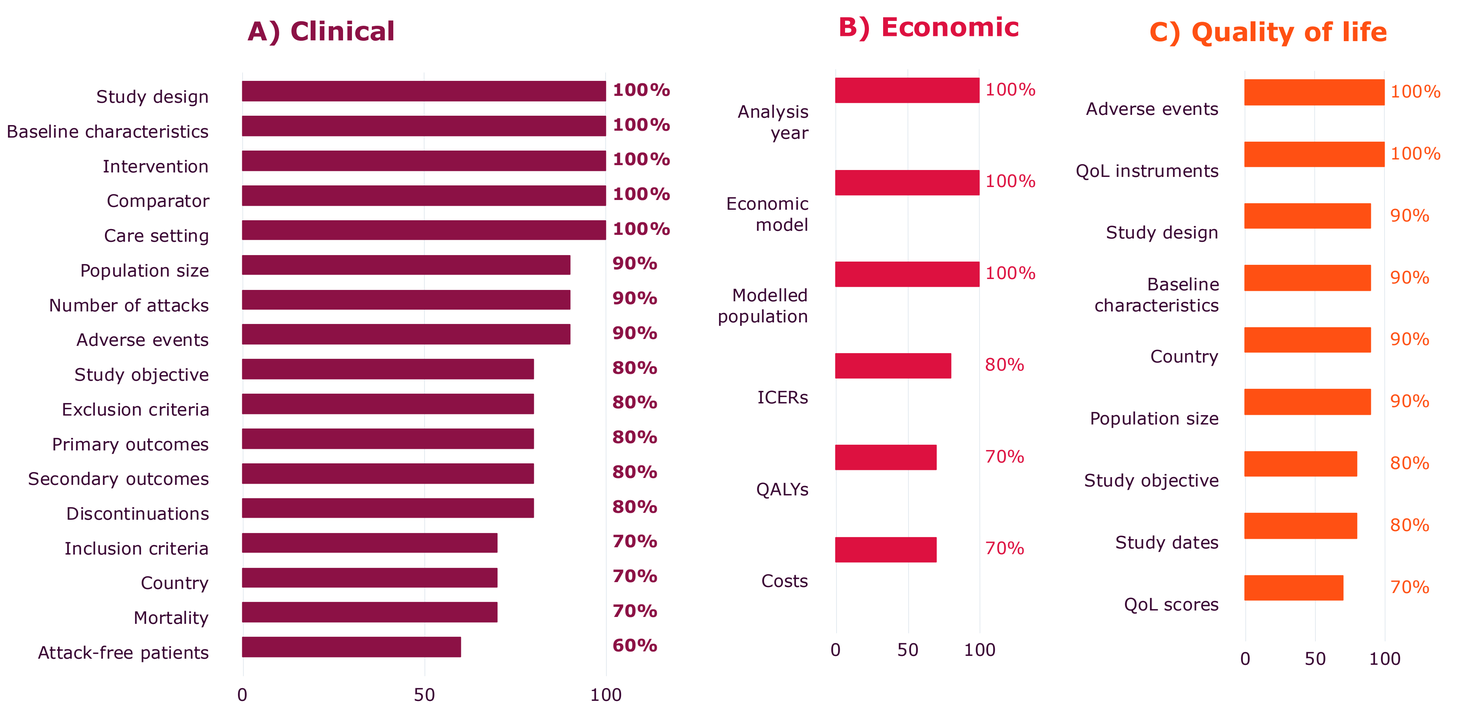
Results: Accuracy by extraction category

Clinical evidence comprised the largest assessment, with 144 of 169 data points extracted correctly. The tool correctly extracted all assessed information relating to study design, baseline population characteristics, interventions and comparators. Accuracy was also high for population size, clinical outcomes and safety data. The lowest accuracy was observed for the number of patients who remained attack-free, for which 6 of 10 outputs were correct (Figure 3A).

For economic evidence, 52 of 60 data points were extracted correctly. The tool correctly extracted the analysis year, model structure and modelled population for all 10 publications. Eight of 10 ICER outputs were correct. Accuracy was lower for comprehensive extraction of QALYs and costs, with 7 of 10 outputs classified as correct for each category. Documented errors included omitted cost information from tables, incomplete extraction of QALY data and instances in which QALY-related information was placed in the ICER field rather than the designated QALY field (Figure 3B).

For quality-of-life evidence, 79 of 90 data points were extracted correctly, representing the highest overall accuracy among the three evidence domains. The tool correctly identified all assessed adverse-event information and quality-of-life instruments. Accuracy was lower for complete extraction of numerical quality-of-life results, with 7 of 10 outputs classified as correct (Figure 3C).

Figure 3: Accuracy by extraction category across A) clinical, B) economic and C) quality of life evidence domains



Note: All panels use a common 0%–100% scale. Categories are ordered by observed accuracy within each evidence domain.
Abbreviations: ICER, incremental cost-effectiveness ratio; QALY, quality-adjusted life years; QoL, quality of life

Conclusion

Within a NICE-aligned SLR workflow incorporating human QC, AI-assisted data extraction demonstrated high overall accuracy across clinical, economic and quality-of-life evidence, while reducing estimated extraction-task time. The observed accuracy was achieved within a framework that retained human review of AI-generated outputs, indicating that efficiency gains were obtained without removing established quality-control processes. The findings suggest that AI-assisted extraction can be incorporated into HTA-focused evidence-synthesis workflows as a structured support tool rather than a replacement for human reviewer oversight. In addition to reducing resource requirements, the use of predefined extraction questions and structured outputs may contribute to greater consistency in the capture and presentation of study data across reviewers and evidence domains. This is particularly relevant for HTA submissions, where transparency, reproducibility and methodological rigour are essential. Further evaluation across broader evidence bases and SLR applications will help establish the generalisability of these findings and further define the impact of AI-assisted extraction on review quality, consistency and efficiency.

Access decisions, built for a changing Europe. Kintiga partners with global pharma and biotech teams to make the European access decisions that matter most – earlier, smarter and with the senior expert insights to defend them.

References:

1. Borah R, Brown AW, Capers PL, Kaiser KA. Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. BMJ Open. 2017 Feb 27;7(2):e012545. doi: 10.1136/bmjopen-2016-012545. PMID: 28242767; PMCID: PMC5337708. **2.** Michelson M, Reuter K. The significant cost of systematic reviews and meta-analyses: A call for greater involvement of machine learning to assess the promise of clinical trials. Contemp Clin Trials Commun. 2019 Aug 25;16:100443. doi: 10.1016/j.conctc.2019.100443. Erratum in: Contemp Clin Trials Commun. 2019 Sep 12;16:100450. doi: 10.1016/j.conctc.2019.100450. PMID: 31497675; PMCID: PMC6722281. **3.** National Institute for Health and Care Excellence (NICE). Use of AI in evidence generation: NICE position statement [Internet]. London: NICE; 2024 Aug 15. (Corporate document ECD11). Available from: https://www.nice.org.uk/corporate/ecd11.