ARTIFICIAL INTELLIGENCE (AI) TIME AND MOTION

QUANTIFYING THE EFFICIENCY AND ACCURACY OF AI IN HTA-STANDARD SLR WORKFLOWS

F. SHOUGHARI, A. HANSSON HEDBLOM, K. FREEMYER, C. BARWOOD, K. KALLMES, AND D. PATIL

Poster: **MSR75**

INTRODUCTION

HTA bodies (e.g., NICE) and regulatory authorities (e.g., FDA) are increasingly exploring and accepting the use of **artificial intelligence (AI)** and **large language models (LLMs)** to synthesize and evaluate evidence for new health technologies. In response, industry has rapidly adopted these tools, accelerating and enhancing systematic literature reviews (SLRs). Collaboration between industry, HTA agencies, and regulators is essential to ensure responsible and purposeful implementation.

To maintain transparency and technical accuracy, improvements over manual SLRs must be objectively demonstrated.

OBJECTIVE

The aim of this study is to quantify the efficiencies and accuracy of different levels of AI applied to an SLR workflow.

METHOD

A review question and PICOS criteria were developed, searches ran and title/abstract screening and extraction undertaken in the three following ways:

1. **Fully human:** manual configuration, screening and extraction
2. **Hybrid AI:** human configuration, AI making recommendations for screening and extraction, human final decision
3. **Fully AI:** human configuration, AI screening and extraction

Excel was used for the manual tasks and AutoLit (Nested Knowledge) for the AI. Following completion of the respective workflows, each approach was assessed in terms of:

- **Time** taken to complete the task
- **Recall:** percentage of publications correctly identified as irrelevant
- **Precision:** percentage of publications correctly identified as relevant
- **Accuracy:** percentage of publications correctly identified as either relevant or irrelevant.

For screening, 'correct' was defined as aligned with the human adjudicator. The quality of the extraction was assessed through manual QC.

TECHNICAL DETAILS OF THE AI MODULES

Hybrid AI screener (Robot Screener)

Machine learning–based module trained on human decisions to provide include/exclude recommendations, acting as one reviewer, with a human as the second reviewer and adjudicator.

Full AI screener (Smart Screener)

LLM-driven module that screens abstracts or full texts using user-defined criteria, provides traceable assessments, and allows inclusion thresholding.

Full AI extraction (Adaptive Smart Tags)

LLM-based module that automatically recommends or extracts user-defined qualitative or quantitative data from abstracts or full texts with full traceability for export and analysis.

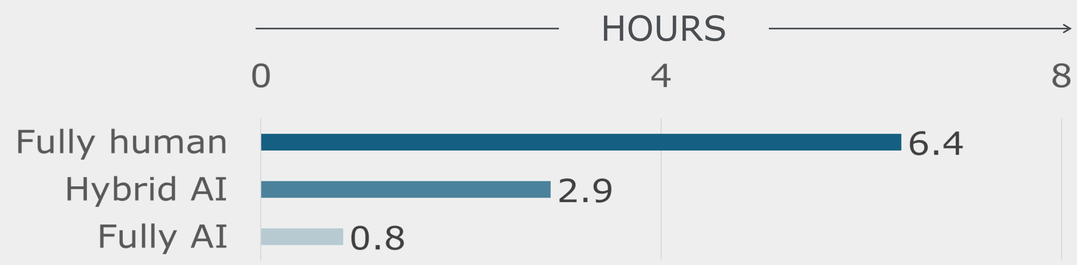
RESULTS

A total of 234 publications were identified in the search. In the **fully human workflow**, the respective reviewers included 50 and 41 studies. Upon adjudication a total of 31 studies were included. In the **hybrid AI workflow**, the AI suggested 26 studies to be included, of which 25 were approved by the human and 6 studies excluded by the AI were reincluded by the human. In the **fully AI workflow**, the AI included 15 studies, all of which were accepted by the human, however 16 studies were missed for inclusion and were added in the final publication selection.

EFFICIENCIES

As shown in **Figure 1**, there are substantial time savings with progressive levels of AI usage. Across dual screening of 234 papers and extraction of 10 papers, the **hybrid AI** and **fully AI** approaches were 55% and 87% faster than the **fully human** approach.

Figure 1: Time required for screening (234 studies) and extraction (10 studies) with progressive levels of AI usage



Abbreviations: AI – Artificial intelligence
Note: Time estimates include arbitration and setting up the project workspace in Excel (Fully human) and Nested Knowledge (hybrid AI and fully AI)

ACCURACY

Screening

As shown in **Table 1**, the **fully human approach** had high recall (89%) but lower precision (62%), suggesting humans are more inclusive but less consistent in excluding irrelevant studies, suggesting that this approach is comprehensive but inefficient. The **hybrid AI approach** had the highest overall accuracy (97%), strong recall (81%) and high precision (96%). **The fully AI approach** achieved perfect precision (100%) but low recall (48%), meaning it rarely includes irrelevant studies but may miss relevant ones.

Extraction

Taking human data extraction as the gold standard, the AI extraction was 96% consistent with manual human extraction. AI extractions were of particularly high quality for qualitative text extraction.

Table 1: Recall, precision and accuracy across progressive levels of AI usage for screening

Approach	Recall	Precision	Accuracy
Fully human	**89%**	62%	91%
Hybrid AI	81%	96%	**97%**
Fully AI	48%	**100%**	93%

Abbreviations: AI – Artificial intelligence

CONCLUSIONS

Appropriate use of AI in SLRs leads to substantial time savings and high quality output.

- Fully AI screening is ideal for precision-critical review tasks as it only includes relevant references, but less suitable when comprehensive recall is needed.
- The hybrid AI approach performs strongly across recall, precision, and accuracy, with higher total accuracy than either humans or full AI.
- AI and humans are more effective together than either alone.
- Combining human judgment with AI precision produces fast and reliable SLRs which is becoming increasingly acceptable to HTA and regulatory bodies.

Want to know more? Contact: Fadel.Shoughari@fiecon.com