SWAR 67: AI-assisted automated data extraction for evidence synthesis using Elicit

Objective of this SWAR
To compare the performance of human-led and Elicit-based automated data extraction in a scoping review of qualitative evidence synthesis.

Study area: Data Extraction
Sample type: Review Authors
Estimated funding level needed: Low

Background
Evidence synthesis relies on accurate and transparent data extraction from the included studies. This step is critical for ensuring the validity of qualitative and quantitative conclusions but is also one of the most time-consuming and resource-intensive stages of the review process.[1] Manual data extraction requires sustained attention and methodological expertise and is vulnerable to human error, inconsistency, and fatigue, particularly when large bodies of literature are involved.[2]

Recent advances in artificial intelligence (AI) and natural language processing have led to the development of tools designed to support or partially automate components of evidence synthesis.[3] Among these, AI-assisted platforms such as Elicit aim to extract structured information directly from full-text articles, potentially reducing reviewer workload and accelerating review timelines. Such tools may be especially appealing for scoping reviews and methodological studies, where the objective is to map existing evidence rather than produce pooled effect estimates.

Despite growing interest in AI-supported review workflows, there are few empirical evaluations of automated data extraction. Existing studies have primarily focused on citation screening or abstract-level tasks, while fewer investigations have examined the accuracy, reliability, and error patterns associated with automated extraction from full-text articles.[4] Moreover, there is a need for transparent comparisons between human-led and AI-assisted extraction processes to better understand the strengths, limitations, and appropriate use cases of these technologies.

Scoping reviews provide a suitable framework for this type of methodological evaluation because they allow exploratory and comparative analyses without requiring formal assessment of intervention effectiveness. Embedding a comparative study in a scoping review enables systematic examination of how automated extraction performs relative to human reviewers across predefined data elements, while also identifying common sources of discrepancy.

This Study Within a Review (SWAR) [5] aims to contribute to the emerging evidence base on AI-assisted evidence synthesis by comparing human-led data extraction with automated extraction using Elicit. By characterizing accuracy and error types in a scoping review context, the findings are intended to inform future methodological development and guide responsible integration of AI tools into evidence synthesis workflows.

Interventions and Comparators
Intervention 1: Human reviewer: A trained reviewer will manually extract predefined data elements from the full texts of all included records.
Intervention 2: Elicit extraction: Automated data extraction will be performed using Elicit to extract the same predefined data elements from the full texts of all included records.

Index Type: Data extraction

Method for Allocating to Intervention or Comparator:
Allocation to the intervention and comparator conditions will be determined by the data extraction approach rather than by participant-level randomization. For each included study, data will be

extracted independently using two parallel methods: (1) manual data extraction performed by a human reviewer and (2) automated data extraction conducted using Elicit. To minimize cross-contamination between approaches, the prompts used for Elicit-based extraction will be developed by a researcher who is not involved in the manual data extraction process. The human reviewer will be blinded to the outputs of the automated extraction until both extraction processes are complete.

Outcome Measures
Primary: Level of concordance of the extracted data between the two data extraction processes. Secondary: Secondary outcomes will include accuracy and error types associated with each data extraction process. Error type will be classified based on the framework proposed by Gartlehner et al. and will include major errors, minor errors, false data ('hallucination') and missed or omitted data.[6]

Analysis Plans
Investigators will systematically compare data obtained through human-led and semi-automated extraction methods. When discrepancies are identified, the original full-text study reports will be reviewed to adjudicate extraction accuracy. Descriptive exploratory analyses will be performed to assess the frequency and distribution of error types.

Possible Problems in Implementing This SWAR
Use of a single human reference assessor: The SWAR employs one human assessor as the reference point for evaluating data extraction accuracy. This approach introduces uncertainty because the AI-assisted extraction will be benchmarked against the judgments of a single individual rather than a collectively agreed reference. Individual variability, including fatigue and subjective interpretation, may influence extraction decisions. Incorporating multiple independent assessors and consensus-based adjudication would strengthen the reliability of the reference standard.

Variability in human judgment: Human data extraction is inherently influenced by cognitive and temporal factors. Changes in attention, decision thresholds, or interpretative consistency over time may affect extraction outcomes. Such variability may, in turn, shape the observed differences between manual and automated extraction approaches.

Dependence on algorithm training characteristics: The effectiveness of the automated extraction method is closely tied to the composition of the data used to develop the underlying model. Limited representation of particular methodologies, specialized language, or research originating from certain regions may lead to uneven performance across included studies.

Tool-specific applicability of findings: The results of this investigation reflect the performance of a specific AI tool and model configuration. Given the diversity of AI-based extraction systems, which differ in design and training strategies, these findings should not be assumed to apply to other tools without direct evaluation.

Contextual limitations of the research field: The assessment is conducted in a narrowly defined subject area. Differences in terminology, reporting standards, and evidentiary norms across disciplines may influence automated extraction performance. As a result, extrapolation of these findings to other research domains or to evidence synthesis more broadly should be undertaken cautiously.

References
1. Nussbaumer-Streit B, Ellen M, Klerings I, Sfetcu R, Riva N, Mahmić-Kaknjo M, et al. Resource use during systematic review production varies widely: a scoping review. Journal of Clinical Epidemiology 2021;139:287–96. doi: 10.1016/j.jclinepi.2021.05.019.
2. Buscemi N, Hartling L, Vandermeer B, Tjosvold L, Klassen TP. Single data extraction generated more errors than double data extraction in systematic reviews. Journal of Clinical Epidemiology 2006;59:697–703. doi: 10.1016/j.jclinepi.2005.11.010.

3. Li L, Mathrani A, Susnjak T. Transforming evidence synthesis: A systematic review of the evolution of automated meta-analysis in the age of AI. Research Synthesis Methods 2026:1-48. doi: 10.1017/rsm.2025.10065.
4. Tóth B, Berek L, Gulácsi L, Péntek M, Zrubka Z. Automation of systematic reviews of biomedical literature: a scoping review of studies indexed in PubMed. Systematic Reviews 2024;13:174. doi: 10.1186/s13643-024-02592-3.
5. Devane D, Burke NN, Treweek S, et al. Study within a review (SWAR). Journal of Evidence-Based Medicine 2022;15(4):328-32. doi: 10.1111/jebm.12505.
6. Gartlehner G, Kahwati L, Hilscher R, Thomas I, Kugley S, Crotty K, et al. Data extraction for evidence synthesis using a large language model: A proof-of-concept study. Research Synthesis Methods 2024;15:576–89. doi: 10.1002/jrsm.1710.

Publications or presentations of this SWAR design

Examples of the implementation of this SWAR

People to show as the source of this idea: Michał Walaszek, Zofia Kachlik, Jakub Ruszkowski
Contact email address: michal.walaszek@gumed.edu.pl
Date of idea: 14/01/2026
Revisions made by:
Date of revisions: