SWAR 37: Automated data extraction for evidence synthesis using Elicit

Objective of this SWAR
To compare the performance of human-led data extraction versus automated data extraction using Elicit in a qualitative systematic review.

Study area: Data extraction, Publications, Qualitative meta synthesis
Sample type: Publications
Estimated funding level needed: Medium

Background
Data extraction has been described as the most time consuming and error prone aspect of evidence synthesis [1]. Research suggests that single reviewer data extraction with verification by a second reviewer takes on average 107 minutes per study, adding considerably to the time taken to complete evidence syntheses [2]. The accuracy of data extraction is not always guaranteed, with up to 70% of systematic reviews having data extraction errors [3]. A recent review indicates that approximately 12% of reviews had data extraction errors that resulted in moderate to large errors in the magnitude of the effects of the interventions studied [4].

Semi-automation, which combines automation of some tasks with human intervention, offers a means of increasing the efficiency of data extraction [5]. Large Language Models (LLMS), a type of machine learning model designed to perform a wide range of text generation and comprehension tasks appear to be most promising. There are several commercially available LLMS including Claude.ai and Elicit, which appear to have high potential to contribute to data extraction. However, there has been limited evaluation of these tools [6].
We have identified only one study which evaluated the use of Claude 2, a commercially available LLM [5]. While the results are promising, the study was restricted to data extraction from a small sample of open access reports of randomized trials, thus limiting the generalizability of findings to other types of evidence syntheses. We are not aware of any studies evaluating the use of Elicit, an AI research assistant designed to semi automate data extraction processes [7]. This Study Within a Review (SWAR) [8] will help to fill this gap.

The host review for this SWAR, ‘How do people with disability or long-term health conditions experience accessing professional support for their parenting roles?’ has been developed by the study team and the review protocol is available on PROSPERO (CRD42023396592) [9].

Interventions and Comparators
Intervention 1: Human-led data extraction completed by two investigators followed by validation for completeness and accuracy by third investigator. This team will use a standardized data extraction form with initial extraction completed independently by two investigators. A third investigator will compare the data extraction against the original PDF files to check for accuracy and completeness. Discrepancies will be resolved through discussion and returning to the original text. Data will be presented in a tabular format.
Intervention 2: Semi-automated data extraction using Elicit. The team will develop a series of prompts (instructions) to guide Elicit in generating accurate and parsimonious outputs. The prompts will be based on clear definitions of each data element. An iterative process of prompt building and testing will be conducted using a random sample (n=10) of articles. Testing will continue until the team are satisfied that the prompts are specific enough to provide accurate and complete responses for each data element. The team will then upload the PDF file of each paper and prompt the model for each data element

Index Type: Full Review; Data extraction

Method for Allocating to Intervention or Comparator:
Cross Over

Outcome Measures

Primary: The primary outcomes are (1) level of concordance of the extracted data between the two data extraction processes and (2) time taken to complete all tasks associated with data extraction and verification.
Secondary: Secondary outcomes to be considered are (1) accuracy and (2) error type made by 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 [5].

Analysis Plans
A blinded investigator who was not involved in the data extraction will compare the results of human only extraction and Elicit data extraction. Where differences occur, they will use the original PDF file to determine which approach is more accurate. We will calculate 95% Clopper-Pearson confidence limits for the proportion of concordant data elements separately for each category of data. To determine the time taken to complete data extraction, the investigators will use a time tracking app while doing the data extraction tasks (including prompt building). We will provide a descriptive exploratory analysis to understand frequency of error types.

Possible Problems in Implementing This SWAR
No problems anticipated.

References
1. Jones AP, Remmington T, Williamson PR, et al. (2005). High prevalence but low impact of data extraction and reporting errors were found in Cochrane systematic reviews. Journal of Clinical Epidemiology 2005;58(7):741-2. doi: 10.1016/j.jclinepi.2004.11.024
2. Li T, Saldanha IJ, Jap J, et al. A randomized trial provided new evidence on the accuracy and efficiency of traditional vs. electronically annotated abstraction approaches in systematic reviews. Journal of Clinical Epidemiology 2019;115:77-89. doi: 10.1016/j.jclinepi.2019.07.005
3. Mathes T, Klaßen P, Pieper D. Frequency of data extraction errors and methods to increase data extraction quality: a methodological review. BMC Medical Research Methodology 2017;17(1):152. doi: 10.1186/s12874-017-0431-4
4. Xu C, Doi SAR, Zhou X, et al. Data reproducibility issues and their potential impact on conclusions from evidence syntheses of randomized controlled trials in sleep medicine. Sleep Medicine Reviews 2022;66:101708. doi: 10.1016/j.smrv.2022.101708
5. Gartlehner G, Kahwati L, Hilscher R, et al. Data extraction for evidence synthesis using a large language model: A proof-of-concept study. Research Synthesis Methods 2024;15(4):576-89. doi: 10.1002/jrsm.1710
6. Thomas J. Large Language Models for health and evidence synthesis: can we trust what they say? Evidence Synthesis Ireland Webinar 23 November 2023. Available from https://evidencesynthesisireland.ie/wp-content/uploads/2023/11/LLMs-for-evidence-synthesis.pdf
7. Elicit. Frequently asked questions. April 2022 Available from: https://elicit.org/faq (Accessed on 9 October 2024).
8. Devane D, Burke NN, Treweek S, et al. Study within a review (SWAR). Journal of Evidence Based Medicine 2022;15(4):328-32.
9. O'Mara V, Honey A, McGrath M. How do parents with disability or long-term health conditions experience accessing professional support for their parenting roles? PROSPERO 2023 CRD42023396592.Available at https://www.crd.york.ac.uk/prospero/display_record.php?ID=CRD42023396592 (Accessed on 9 October 2024).

Publications or presentations of this SWAR design

Examples of the implementation of this SWAR

People to show as the source of this idea: Margaret McGrath, Anne Honey, John V Rider, Evelina Pituch, Veronica O Meara
Contact email address: margaretmcgrath@ucc.ie
Date of idea: 02/APR/2024
Revisions made by: Margaret McGrath, Anne Honey, John V Rider, Evelina Pituch, Veronica O Meara
Date of revisions: 17/SEP/2024