SWAR 28: Semi-automated data extraction for evidence syntheses using Claude 2 and Claude 3

Objective of this SWAR
To assess the use of a semi-automated data extraction process with Claude 2 and Claude 3 in systematic reviews, by investigating the concordance between a human-only and a semi-automated data extraction process and comparing the time needed for each process. A secondary objective is to assess the accuracy and types of error of each data extraction process.

Study area: Data extraction
Sample type: Publications
Estimated funding level needed: Low

Background
Data extraction (i.e., the process of manually extracting data from included studies into standardized tables) is a crucial, but labour-intensive and error-prone part of evidence synthesis.[1] A randomized trial of different data extraction strategies found that single investigator data extraction and verification by a second investigator took an average of 107 minutes per study, while dual independent data extraction took 172 minutes.[2] Data extraction errors can undermine the validity of evidence syntheses, affecting narrative summaries, meta-analyses and conclusions. A methodology review has revealed a high rate of data extraction errors (up to 63%) in systematic reviews.[3] The error rate varied depending on the type and complexity of the data.[3, 4] Causes of data extraction errors are multifaceted, including inaccuracies such as overlooking available data, misclassification (e.g., confusing standard deviations and standard errors), misinterpretation due to ambiguous reporting in primary studies, or straightforward data entry mistakes. Time constraints and language barriers can further heighten the risk of data extraction errors,[5-7] and other problems include the possibility that a correct extraction of a data element by one person could be erroneously changed by a second person checking the completeness and accuracy of the data extraction.

Artificial intelligence (AI) might increase the efficiency of data extraction and reduce errors. Previous research on methods for semi-automating data extraction has mostly focused on natural language processing (NLP) using statistical models such as naive Bayes or support vector machines.[8] These models require training data and often encounter difficulties in extracting information from articles in portable document format (PDF), especially tables or graphs. In general, the training of NLP models to extract data is time-consuming and resource intensive. The October 2023 version of a living systematic review on automated and semi-automated data extraction methods includes 76 publications since 2005.[8] Most studies addressed data extraction from abstracts alone; only 19 addressed extraction from full-text PDFs. Therefore, research on this topic suggests that tools for automated or semi-automated data extraction based on NLP are not yet mature enough for practical use.[6]

Large Language Models (LLMs) have brought new possibilities to increase efficiency of data extraction. In a recent proof-of-concept study, we used Claude 2 [9] to assess the performance of a LLM for data extraction.[10] We used data previously extracted by a single investigator and reviewed for accuracy against the source PDF by a second investigator from a convenience sample of 10 English-language open-access reports of randomized trials included in a systematic review on targeted immune modulators for the treatment of plaque psoriasis. We selected 16 distinct types of data, posing varying degrees of difficulty (160 data elements across the 10 studies) and iteratively developed prompts for each data element. We used the browser version of Claude 2 to upload each PDF and prompted the model for each data element. Across 160 data elements, Claude 2 demonstrated an overall accuracy of 96.3% with a high test-retest reliability (replication 1: 96.9%; replication 2: 95.0% accuracy). Overall, Claude 2 made six errors on 160 data items, with the most common (n=4) being missed data items. Importantly, Claude 2's ease of use was high; it required no technical expertise or training data for effective operation.

However, this proof-of-concept study has several limitations. We limited study designs to open-access publications of randomised trials, did not evaluate continuous outcome data elements, results from multiple study arms, or data that were not reported as primary outcomes. Furthermore,

our reference standard dataset contained only three instances where data we were interested in were not reported in the primary study report. This SWAR [11] will expand on that proof-of-concept study. The first review in which this SWAR will be embedded (Implementation strategies for preventive mental health interventions in children) will use Claude 2 but all other reviews will switch to Claude 3 (which was released on 4 March 2024.

Interventions and comparators
Intervention 1: Team 1: Human-led data extraction led by one investigator followed by validation of completeness and accuracy by another investigator. This team will extract data into standardized data extraction sheets with initial data extraction done by a single investigator. A second investigator will review extracted data for completeness and accuracy against the PDF of the study report. Discrepancies will be resolved by discussion and a table will be prepared with the final data extraction.
Intervention 2: Team 2: Semi-automated data extraction by Claude 2 followed by validation of completeness and accuracy by a human investigator. This team will receive a pool of prompts which were successfully used previously for data extraction with Claude 2. They will have access to a data scientist with expertise in prompt engineering if support is needed and will (1) create, test, and refine prompts based on two study publications used during piloting; (2) data extract using Claude 2; (3) verify the completeness and accuracy of this data extraction through human investigator checks; (4) resolve discrepancies between Claude 2 and the human investigator; and (5) prepare a table with the final data extraction. When Team 2 has devised efficient prompts for each data element, they will apply them to the selected study reports and use the browser version of Claude Pro (which is the paid tier of Claude 2), to upload the PDF of each study report and prompt the model for each data element. They will collect data extracted by Claude Pro in Excel spreadsheets. A human investigator will review the extracted data for completeness and accuracy, following the same process used by Team 1.

Index Type: Data extraction, ,

Method for allocating to intervention or comparator
Non-Random

Outcome measures
Primary: Concordance of extracted data between the two data extraction processes and the time required for data extraction tasks (defined as the proportion of extracted data items that are factually congruent, even if there are variations in style and presentation and the total time taken for all the tasks necessary to complete the extraction and verification, including prompt engineering for semi-automated data extraction.
Secondary: Accuracy and types of error made by each data extraction process. To determine accuracy, results of the human-led data extraction will be the reference standard. When discrepancies between the two extraction processes occur, we will check the respective full-text PDF to validate the accuracy of the reference standard. Because human-led data extraction is an imperfect reference standard, we will follow guidance by the Agency for Healthcare Research and Quality [12] and make necessary corrections if there are errors in the reference standard. To categorize the types of errors made and their impact, we will use the following classification system with one investigator classifying the types of errors and a second investigator reviewing these classifications for correctness.

Types of errors
Missed or omitted data: Data available in the source PDF but overlooked or omitted by the data extraction process.
Fabricated data: Data that were not available in the source PDF but were either inaccurately completed by human data extractors or erroneously generated (hallucinated) by the LLM.
Misallocated data: Data in the source PDF that were allocated to the wrong data extraction field.
Incorrect calculations: Mathematically incorrect calculations of data elements based on information in the source PDF, including rounding errors.
Difference in level of detail: The level of detail is the only difference.
Other: Optional “other” field to be used if none of the above apply.

Impact of errors

Major error: This error significantly compromises the accuracy of the data, and, if uncorrected, could lead to erroneous conclusions in the evidence synthesis; for example, grossly incorrect calculations, misallocated data that results in a different interpretation, or hallucinations of the LLM that result in a new or different interpretation of the data.
Minor error: This error is less severe than a major error and may or may not impact interpretation of the existing data. For example, small calculation errors or rounding errors that do not critically affect the data's overall utility, additional or alternative language describing study inclusion/exclusion criteria or the intervention that don't inherently alter the meaning.
Inconsequential error: This difference most likely would not impact the interpretation of the data.

Analysis plans

Comparison of results of the extracted data

Blinded investigators who did not take part in the data extraction will compare the results of human-only and semi-automated data extraction. In instances of discrepancies, the investigators will review the PDF of the study report and make a final determination regarding the accuracy of data extraction.

Sample size calculation

Our proof-of-concept study rendered a concordance between human data extraction and LLM-assisted data extraction of 83%.[10] For this SWAR, we aim to estimate the proportion of concordant data elements between the two processes, with a confidence interval (CI) of approximately +/- 3.5 percentage points for each category of data element (study characteristics, participant characteristics, interventions and participant flow, and results; as shown in Table 1). To achieve this, we will consider 500 data elements to be extracted per category from all publications collectively, assuming an observed proportion of 83%. Using the Clopper-Pearson method to calculate two-sided 95% CI, we estimate the range to be 79.4% to 86.2%, assuming that the data follows a binomial distribution with parameters: n=500 (number of data elements) and p=0.83 (observed proportion).

Data analysis

In general, the unit of analysis is the data element. In some instances, the data element might be split into more than one unit (e.g., if a data element requires more complex information such as proportions and an effect estimate). The exact unit of analysis will be determined on a case-by-case basis for each included review.

We will calculate 95% Clopper-Pearson confidence limits for the proportion of concordant data elements, separately by category of data element. We might also conduct exploratory subgroup analyses, which may encompass various factors, including but not limited to the included review (different topics and scopes) and (2) design of included studies (randomized versus non-randomized)

If the data allow for it, we can estimate the proportion of concordance and calculate confidence limits for a specific category using a Generalized Estimating Equation (GEE) model. This takes into consideration the potential correlation between residuals from the same publication. The summary measure from this model will offer a marginal interpretation, providing insights into the average data element. In addition, if the data permit, we may further explore the data through a generalized linear mixed model approach. To assess time required for data extraction, investigators will track their time using a time tracking app while working the data extraction tasks (including prompt engineering).

For the remaining endpoints, we will conduct descriptive exploratory analyses, with no adjustments for multiplicity in the analyses. This allows us to explore the data comprehensively and generate valuable insights without imposing strict adjustments for multiple comparisons.

Data Management

We will use publicly available published study reports and store data electronically in Excel. All investigators will have access to the data. The final prompts will be saved and shared as part of dissemination and we will make the final data available upon request.

Possible problems in implementing this SWAR

No problems expected.

References

1. Nussbaumer-Streit B, Ellen M, Klerings I, et al. Resource use during systematic review production varies widely: a scoping review. Journal of Clinical Epidemiology 2021;139:287-96.
2. Li T, Saldanha IJ, Jap J, Smith BT, 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.
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;17:152.
4. Yi Z, Xi Y, Suhail ARD, et al. Effects of double data extraction on errors in evidence synthesis: a crossover, multicenter, investigator-blinded, randomized controlled trial. medRxiv 2023:2023.10.16.23297056.
5. Jonnalagadda SR, Goyal P, Huffman MD. Automating data extraction in systematic reviews: a systematic review. Systematic Reviews 2015;4:78.
6. Marshall IJ, Wallace BC. Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. Systematic Reviews. 2019;8:163.
7. Blaizot A, Veettil SK, Saidoung P, et al. Using artificial intelligence methods for systematic review in health sciences: A systematic review. Research Synthesis Methods. 2022;13(3):353-62.
8. Schmidt L, Olorisade B, McGuinness L, et al. Data extraction methods for systematic review (semi)automation: update of a living systematic review [version 2; peer review: 3 approved]. F1000Research 2021;10:401.
9. Anthropic. Claude 2: https://www.anthropic.com/news/claude-2 (accessed 9 February 2024).
10. Gartlehner G, Kahwati L, Hilscher R, et al. Data Extraction for Evidence Synthesis Using a Large Language Model: A Proof-of-Concept Study. medRxiv 2023:2023.10. 02.23296415.
11. Devane D, Burke NN, Treweek S, et al. Study within a review (SWAR). Journal of Evidence-Based Medicine 2022;15(4):328-32.
12. Trikalinos TA, Balion CM. Chapter 9: options for summarizing medical test performance in the absence of a "gold standard". Journal of General Internal Medicine 2012;27(Suppl 1):S67-75.

Publications or presentations of this SWAR design

Examples of the implementation of this SWAR

People to show as the source of this idea: Gerald Gartlehner
Contact email address: gerald.gartlehner@donau-uni.ac.at
Date of idea: 11/FEB/2023
Revisions made by:
Date of revisions: