Accelerating Data Extraction for Systematic Reviews with a Large Language Model

Orion Kooistra, Postdoc Fellow Biomedical Informatics, OHSU

Ilya Ivlev, MD, PhD, MBI, OHSU Faculty Advisor and Mentor

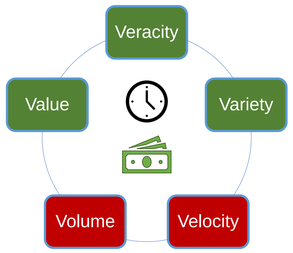
School of MEDICINE

Background

Why Systematic Reviews Matter:

Systematic reviews of Randomized Controlled Trials (RCTs) are considered the highest quality evidence in healthcare.

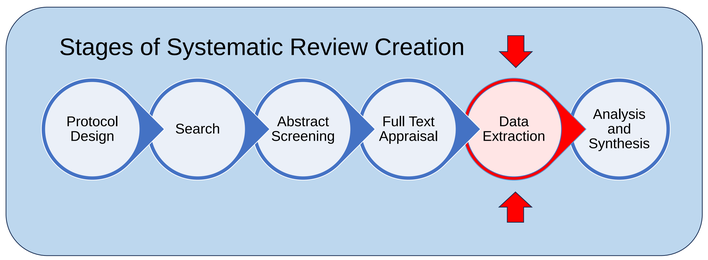
They are essential in synthesizing a huge **variety** of information, determining **veracity** through rigorous appraisal, and increasing the **value** of evidence through synthesis.



Traditional methods of systematic review creation are challenged by the increasing **volume** and **velocity** of information, adding to costs in time and money.

- Systematic reviews registered with PROSPERO take, on average, 67 weeks from protocol registration to publication.[1]
- One source estimates and average cost of $141,194.80[2], though there is a lot of variance based on size and complexity of the review.
- It has been reported that the time taken to produce a systematic review has not changed in 30 years.[2]

The Problem: Though there are many artificial intelligence (AI) tools available for the Search and Abstract Screening stages, there are relatively few data extraction tools that have been tested and evaluated.



- Manually extracting patient characteristics from RCT reports is time-consuming and labor intensive, slowing the completion of systematic reviews. Much of the patient characteristic data is found in tables and in the text of the report.
- Recent advances in large language models (LLMs), such as Claude 3.5 Sonnet, offer the potential to automate this process and reduce human workload.

Purpose: To develop a concept and software to evaluate whether Claude 3.5 Sonnet can accurately extract patient characteristics from RCTs, using human-abstracted data from an ongoing systematic review on behavioral interventions for children as the reference standard.

Methods

Software Development

- We created a public Github repository: https://github.com/Orion907/research_data_extractor
- We created a virtual environment and installed Python packages:
  - Core Python packages:
    - os – For file and directory operations
    - sys – For system-specific parameters and functions
    - json – For working with JSON data
    - logging – For logging functionality
    - datetime – For working with dates and times
    - unittest – For unit testing
  - PDF, CSV, data handling, llm integration packages:
    - PyPDF2 – For text extraction from PDF files
    - csv – For working with csv files
    - Pandas – For data manipulation and analytics
    - openai, anthropic, google APIs
    - Streamlit – For web-based GUI
- The IDE used for this project is VScode with Github Copilot set to Claude Sonnet 3.5 for code completion, evaluation and debugging.

Data Extraction Pipeline

We created a data extraction pipeline, written in Python with the GUI shown below.

The pipeline functions as follows:

PDF Processing
- .pdf file is uploaded through web-based GUI pictured above
- .pdf converted to .txt format

Text Chunking
- .txt file is split into chunks to increase LLM comprehension.
- sentence structure preserved; paragraph structure preserved if possible.

API Integration
- The LLM is prompted via API call
- Anthropic API was used for the purpose of this project

Output
- LLM output is captured as a JSON object
- JSON object is converted to .csv for storage and analysis

Anticipated Results

Performance Metrics

- We will evaluate recall, precision, f1 score, accuracy and estimated time saved.
- We will incorporate a qualitative evaluation of extraction results based on expert feedback

Possible Successes:

Efficiency
- A study reported LLM data extraction time around 90s per RCT vs. around 87 min for manual data extraction.[3]

High Accuracy
- Claude 2 was reported to achieve 96.3% accuracy on a similar data extraction task.[4]

High Recall
- Claude 2 was reported to achieve 96.2% recall, extracting 151/157 data items.[4]

Possible Challenges:

False Positives
- GPT 4 Turbo was reported to have a 19% false positive rate over three test runs.[5]

Opacity and Confabulation
- Current LLMs are black boxes and tend to make things up.

Stochasticity
- Even with minimal temperature settings there is a degree of randomness.

Future Directions

- Develop internal validation of extracted data with robust error handling for LLM responses.
- Parallel processing to handle batches of PDFs.
- A command line interface for increased functionality and integration with other software
- Improve error handling and reporting
- Results management to store, organize, and analyze extraction results
- Metrics collection to track processing performance
- Enhance prompt versioning so user can modify prompts through GUI
- Implement techniques such as Retrieval Augmented Generation and Test-Time Training to enhance accuracy.
- Documentation and user guides

References

1. Marshall IJ, Marshall R, Wallace BC, Brassey J, Thomas J. Rapid reviews may produce different results to systematic reviews: a meta-epidemiological study. Journal of Clinical Epidemiology. 2019 May;109:30–41.
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. Contemporary Clinical Trials Communications. 2019 Dec;16:100443.
3. Lai H, Liu J, Bai C, Liu H, Pan B, Luo X, et al. Language models for data extraction and risk of bias assessment in complementary medicine. npj Digit Med. 2025 Jan 31;8(1):74.
4. 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 Jul;15(4):576–89.
5. Spillias S, Ollerhead K, Andreotta M, Annand-Jones R, Boschetti F, Duggan J, et al. Evaluating Generative AI to Extract Qualitative Data from Peer-Reviewed Documents [Internet]. In Review; 2024 [cited 2025 Apr 29]. Available from: https://www.researchsquare.com/article/rs-4922498/v1