Using artificial intelligence (AI) to analyze large amounts of data: A case study

Poster 20

Brittany Friedson[a], Kate Katsaval[a], Nicole Coolbaugh[a], Julia L. Jones[a], Sara Hartwell[a], and Philip J. Sjöstedt[a]

[a]*The Medicine Group, LLC, New Hope, PA, United States*

Introduction

- Systematic literature reviews (SLR) gather and synthesize all available evidence on a specific topic or related to a particular research question.[1]
- With the amount of literature and sources rapidly growing, the traditional methods of obtaining articles for an SLR can be labor intensive, leading to lost efficiency.[2,3]
- Artificial intelligence (AI) is an emerging tool that has proven beneficial across a variety of applications, including in the processes of gathering information and writing.[2,3]

Hypothesis

- Utilizing AI will increase efficiency and aid in streamlining the SLR analysis process.

Research design and methods

- A non-traditional SLR was performed to investigate the descriptors authors used to describe a clinical trial population (i.e., “subjects”, “patients”, and/or “participants”), with a focus on migraine (results reported separately).
- The SLR followed an informal SLR process: define the inclusion/exclusion criteria, generate the search string, conduct the search, screen the titles/abstracts, extract the data, and analyze the results (**Figure 1**).

Figure 1. Research design

01 STEP

Created a search string for PubMed:

- ("Migraine"[title]) AND ("2014/01/01"[Date - Publication]: "2024/12/31"[Date - Publication]) AND "clinical trial" [Publication Type]

02 STEP

Performed a search in PubMed, applying the following filters for inclusion criteria:

- *Publication date*: ≤10 years
- *Article type*: Clinical Trial, Phase I–IV; Randomized Controlled Trial; Controlled Clinical Trial

03 STEP

PubMed2XL was used to extract the full abstract and publication dates to include in the Excel file for analysis

04 STEP

Manually excluded articles that met any of the exclusion criteria:

- *Publication type*: “review” or “comment”
- *Earliest publication year (print or electronic)*: >10 years
- *Abstract*: none

05 STEP

Developed prompts for AI platform to analyze/extract the trial population descriptor(s) used in the title/abstract:

- *Method A:* Analyze the uploaded Excel file
- *Method B:* Generate Excel formulas that can extract the desired information

06 STEP

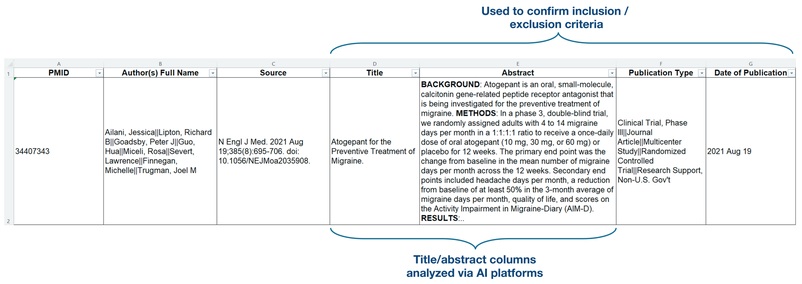
Tested instructional prompts from methods A and B using multiple AI platforms, including[4]:

- ChatGPT-4o mini
- Claude v3.0/v3.5
- Gemini v1.0/v1.5
- Llama v3/v3.1
- Microsoft CoPilot
- Mistral AI
- Perplexity AI

AI, artificial intelligence; PMID, PubMed Identifier.

- Article details from captured PubMed literature (N=850 articles) were exported to Excel through PubMed2XL for screening and analysis (**Figure 2**).

Figure 2. Snapshot of the SLR Excel spreadsheet used for screening and data extraction



This snapshot is of the Excel file exported from PubMed2XL after entering all the PMIDs captured in the literature search. AI, artificial intelligence; PMID, PubMed Identifier; SLR, systematic literature review.

- We investigated the ability of various AI platforms to execute prompts requesting analysis of the title and abstract columns in the Excel file to identify the population descriptor(s) and recorded each platform's abilities separately.
  - The AI platforms investigated were based on the LLM Overview spreadsheet developed in August 2024 by the ISMPP AI Task Force.[4]

Results

- Most AI platforms could not analyze the Excel file directly (method A) (**Table 1**).

Table 1. Comparison of AI platforms in abilities to analyze text in Excel

AI platform*	Pros	Cons	Ability to extract using Excel
ChatGPT-4o mini†	• *With a paid subscription:* – Accepts text (useful for creating formulas to enter back into Excel) – File types accepted: PDF, Excel, PowerPoint, Word, etc. – Analyzes Excel files, with unlimited prompts	• *Without a paid subscription:* – The number of prompts and ability to upload files are limited • Unable to upload multiple files at once or access external databases for analyzing large volumes of literature at once	✓
Claude v3.0 & v3.5	• Accepts text (useful for creating formulas to enter back into Excel) – File types accepted: PDF, Excel, PowerPoint, Word, etc. – Smaller size restrictions compared to other AI platforms	• Unable to upload an entire Excel file with large amounts of data	✗
Gemini v1.0 & v1.5	• Accepts text (useful for creating formulas to enter back into Excel)	• Unable to upload an Excel file or to paste contents of the Excel file • Version 1.5 allows you to use text and upload images only (e.g., screenshots)	✗
Llama v3 & v3.1	• Accepts text (useful for creating formulas to enter back into Excel)	• Does not upload other file types, including Excel	✗
Microsoft CoPilot	• Accepts text (useful for creating formulas to enter back into Excel) • Files types accepted: PDF, Excel, PowerPoint, Word, etc.	• Not as useful for analyzing an uploaded Excel file (i.e., does not easily recognize column names)	✗
Mistral AI	• Accepts text (useful for creating formulas to enter back into Excel) • File types accepted: PDF and image	• Requires an account for use • Unable to analyze an Excel file	✗
Perplexity AI	• Accepts text (useful for creating formulas to enter back into Excel) • File types accepted: PDF, Word, image, and Excel	• Can upload up to three files per day, with a maximum of four files at once • Can only ask 3 Pro searches per day (Pro Search: this feature allows Perplexity to conduct more in-depth searches with more accuracy)	✓

*All AI platforms were tested using an unpaid subscription, except for ChatGPT-4o mini. †ChatGPT 4o-mini was selected for literature analysis due to its ability to analyze extracted data from Excel and to generate formulas for easy use. Further testing also revealed that some manual analysis and trial-and-error are necessary to create specific prompts for accurate data analysis within Excel. AI, artificial intelligence.

- The most promising, a paid subscription to ChatGPT-4o mini, was also prompted to generate Excel formulas that could analyze each title/abstract (method B).
  - **Table 2** shows examples of ChatGPT-4o mini prompts for methods A and B.
- Both methods were successful; however, significant human time was spent refining the prompts to obtain more accurate results (**Table 2**).
  - ChatGPT-4o mini was able to analyze the Excel file once prompts were imputed with very specific language.
  - Human observation was necessary to identify the different contexts of each term to determine the appropriate prompt and to assess the accuracy of the results generated by each prompt.

Table 2. ChatGPT-4o mini prompts for identifying descriptors in titles/abstracts from an Excel file*

Example prompts to analyze an Excel file	Results and observations†
Help me by completing Column I in the attached file, by identifying if the text in Column D or E uses the word "patient", "participant", or "subject". If Column D or E uses multiple descriptors, please include them in separate columns.	• Did not always appropriately identify descriptor terms in titles/abstracts • Could not recognize word variations (e.g., punctuation or partial words) without specific criteria
In the "Migraine" tab in the attached Excel file, please recognize the descriptor word "participants" or "subjects" or "patients" in column D2 or E2. Please make sure the word is exact and not detected within a different larger word, UNLESS "patients" is part of the term "outpatients" or "inpatients". Please make sure the formula can still detect the descriptor word if found next to the punctuation marks ";", ",", ":", "-", ".", "(", ")", "/", or "". Make the search for the term "patients", "participants", or "subjects" case insensitive. Please give the output in the new column as "patients" if found as "patients", "outpatients", or "inpatients". In a separate column, please give the output as "participants" if the term is identified. In a separate column, please give the output as "subjects" if the term is identified. Please also make the cell blank if the word is not detected.	• Was able to identify terms within the Excel file with more accuracy
Example prompts to generate Excel formulas‡	**Results and observations†**
Please create a formula to recognize the descriptor word "participants" in column D2 or E2 in Excel. Please provide an output of the term if it is identified and make the cell blank if not detected.	• Prompt was sufficient to produce a formula that accurately identified “participants” within a title or abstract, with an output of the term in the new cell[§] • This formula was not sufficient to appropriately identify terms that can be used in different contexts (e.g., the descriptor term “men”, was still identified if it was part of the term “women”)
Please create a formula to recognize the descriptor word "men" in column D2 or E2 in Excel. Please add a space between the descriptor and the quotation so the word is exact and not detected within a different larger word. Please make sure the formula can still detect the descriptor word if found next to the punctuation marks ";", ",", ":", "-", ".", "(", ")", "/", or "". Make the search for the term "participants" case insensitive. Please also make the cell blank if the word "men" is not detected.	• Prompt was necessary to generate a formula that identified the desired term in different contexts within a title or abstract[¶] • For long and complicated formulas, human trial and error was required to modify the formula generated until the formula was recognized in Excel (i.e., manually removing extra parentheses or adding parentheses to the formula in Excel)

*For all prompts, the title was identified as Column D, and abstract was identified as Column E in the Excel file (**Figure 2**).
†To ensure accuracy, analysis results from the uploaded Excel file were reviewed with human observation.
‡The prompts were repeated for each descriptor term and each formula was pasted into a new Excel column.
§Formula output from ChatGPT-4o mini: `=IF(OR(ISNUMBER(SEARCH("participants", D2)), ISNUMBER(SEARCH("participants", E2))), "participants", "")`
¶Formula output from ChatGPT-4o mini: `=IF(OR(ISNUMBER(SEARCH(" men ", SUBSTITUTE(SUBSTITUTE(SUBSTITUTE(SUBSTITUTE(SUBSTITUTE(SUBSTITUTE(SUBSTITUTE(SUBSTITUTE(SUBSTITUTE(" "&LOWER(D2)&" ", ".", " "), "(", " "), ")", " "), "-", " "), "/", " "), ",", " "), ":", " "), ";", " "))), ISNUMBER(SEARCH(" men ", SUBSTITUTE(SUBSTITUTE(SUBSTITUTE(SUBSTITUTE(SUBSTITUTE(SUBSTITUTE(SUBSTITUTE(SUBSTITUTE(SUBSTITUTE(" "&LOWER(E2)&" ", ".", " "), "(", " "), ")", " "), "-", " "), "/", " "), ",", " "), ":", " "), ";", " "))), "men", "")`

Limitations

- There was no control in this study to quantify the validity of the hypothesis.
- As this was a novel approach, the next time performing this task will likely take less time.
- The AI tools investigated were based on a list compiled in August 2024 and were limited to free versions available in November 2024 (except for ChatGPT); thus, they do not reflect significant advances in AI technology since that time.

Conclusions

- Most AI tools investigated could not reliably analyze or extract information from Excel files, primarily due to lack of file type support.
- After extensive manual optimization of appropriate prompts, ChatGPT-4o mini assisted with the most accurate analyses of the Excel file and was also able to provide advanced Excel formulas that produced the desired output, which subsequently reduced time needed for manual analysis.
- The refinement of AI tools and precision in AI prompts may further reduce the time needed to analyze vast amounts of literature.

References

1. Pati D, et al. *HERD*. 2018;11(1):15-30.
2. Mahuli SA, et al. *Br Dent J*. 2023;235(2):90-2.
3. Fabiano N, et al. *JCPP Adv*. 2024;4(2):e12234.
4. ISMPP AI Task Force. LLM Overview. Published Aug 2024. Available from https://ismpp.memberclicks.net.

Disclosures

All authors are full-time employees of The Medicine Group LLC, New Hope, PA, United States.

T M G
The | Medicine | Group

Poster presented at the 21st Annual Meeting of ISMPP May 11–14, 2025 Washington, D.C.