Automated Data Extraction in Systematic Literature Reviews (SLRs): Assessing the Accuracy and Reliability of a Large Language Model (LLM)

MSR187

Aiswarya Shree[1], Mariana Farraia[2], Smit Pathak[1], Mahmoud Slim[3], Allie Cichewicz[4], Lalith Mittal[1], Carolina Casañas i Comabella[5]*

[1]Evidera Ltd, Bengaluru, India; [2]Evidera Ltd, Ede, Netherlands; [3]Evidera Inc., Montréal, Canada; [4]Evidera Inc., Boston, USA; [5]Evidera Ltd, London, UK.

*Presenting author

Background

- Systematic literature reviews (SLRs) are key for informing evidence-based decision-making. However, the gold-standard SLR approach to data extraction is time-consuming and demands a considerable investment of resources, leading to cost and time implications.
- Recent advancements in artificial intelligence (AI) have sparked an interest in leveraging AI-based large learning models (LLMs) to increase the efficiency of the data extraction process. LLMs can perform a wide range of text generation and comprehension tasks based on a set of instructions (i.e., prompts).
- The release of accessible LLMs, such as Generative Pre-trained Transformer-4 (GPT-4), has introduced new potential for its use in evidence synthesis, including data extraction. Despite such advancements, careful investigation is needed to ascertain their accuracy and potential time-saving benefits.

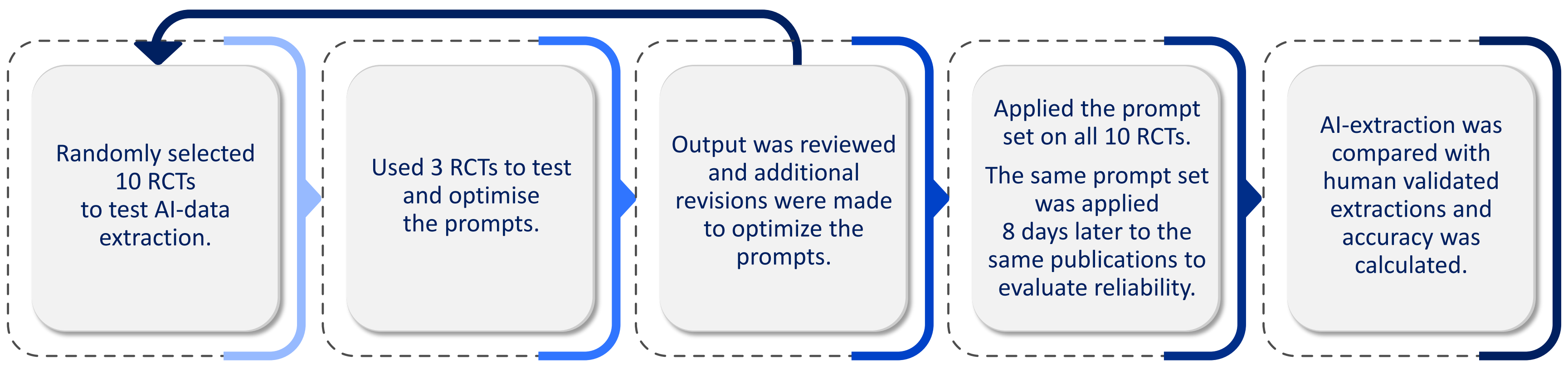
Objective

- To assess the accuracy and reliability of LLM-assisted data extraction from peer-reviewed randomized controlled trials (RCTs) by comparing the data output with that of a previously conducted traditional SLR.

Methods

- LLMs require user-provided instructions, known as prompts, to generate accurate and relevant responses. We used an iterative process to develop and refine prompts for data extraction. This involved evaluating the initial prompt effectiveness, analysing responses, and making adjustments to improve the specificity and relevance of the prompt (**Figure 1**). Prompt testing and refinement included providing adequate context, specific examples, and providing instructions for the desired response format. Specific statistical details and timepoints were also incorporated in the prompts to request specific information on complex quantitative data, for example including text instructions such as "mean change from baseline at week 16".

Figure 1. Prompt development process



Randomly selected 10 RCTs to test AI-data extraction.

Used 3 RCTs to test and optimise the prompts.

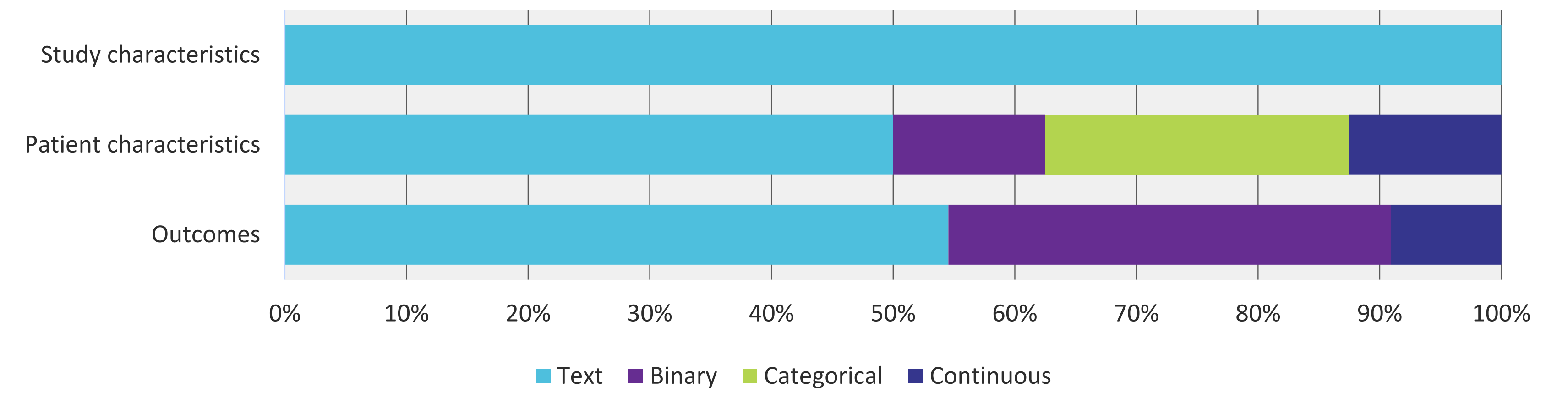
Output was reviewed and additional revisions were made to optimize the prompts.

Applied the prompt set on all 10 RCTs. The same prompt set was applied 8 days later to the same publications to evaluate reliability.

AI-extraction was compared with human validated extractions and accuracy was calculated.

- Three sets of prompts were designed to extract data on 29 variables across three main areas (**Figure 2**):
  - **Study characteristics:** Focused on extracting trial details such as name, study phase, population, interventions, and inclusion/exclusion criteria.
  - **Patient characteristics:** Designed to capture variables such as age, gender, comorbidities, and disease severity (using the IGA score).
  - **Outcomes:** Optimized for extracting clinical endpoints (EASI 75, POEM, DLQI scores), safety data, and assessment timepoints.

Figure 2. Distribution of variable types (total 29 variables)



Study characteristics: Author, year [text], trial name[text], study phase[text], population description[text], intervention/comparator[text], sample size[text], and inclusion/exclusion criteria[text].
Patient characteristics: Mean age[continuous]., proportion of males[binary], comorbidities[categorical]., and proportion patients with IGA score of 3 and 4 [categorical].
Outcomes: Assessment timepoint[text], disease-specific clinical endpoints (EASI 75 [binary], POEM [binary] and DLQI score[continuous]) and safety (discontinuation due to AE and SAEs) [binary].
Abbreviations: AD: Atopic dermatitis; AE: adverse events; DLQI: Dermatology Life Quality Index; EASI 75: Eczema Area and Severity Index (75% improvement); IGA: Investigator Global Assessment; POEM: Patient-Oriented Eczema Measure; SAE: serious adverse event.

- The data extracted by the LLM were compared with human data extraction from a previous SLR (traditional SLR). The traditional SLR included studies that evaluated the efficacy and safety of treatments in atopic dermatitis. For our study, we used a random sample of 10 recently published RCTs[1-10] (2017–2023) from the traditional SLR, which had been manually extracted by one reviewer and validated by a second reviewer, without any AI input.
- The process was replicated eight days later with the same 10 publications and identical prompts, to assess any differences in LLM performance across different days. Reliability was measured using the intraclass correlation coefficient (ICC) with the replicated data set.
- AI-generated data extraction was compared to the manual human extractions and each extracted variable was rated as: correct, incorrect, missing, or incomplete. Accuracy was determined using the following formula:

$$Accuracy = \frac{\text{Number of correctly extracted variables}}{\text{Total number of variables extracted}}$$

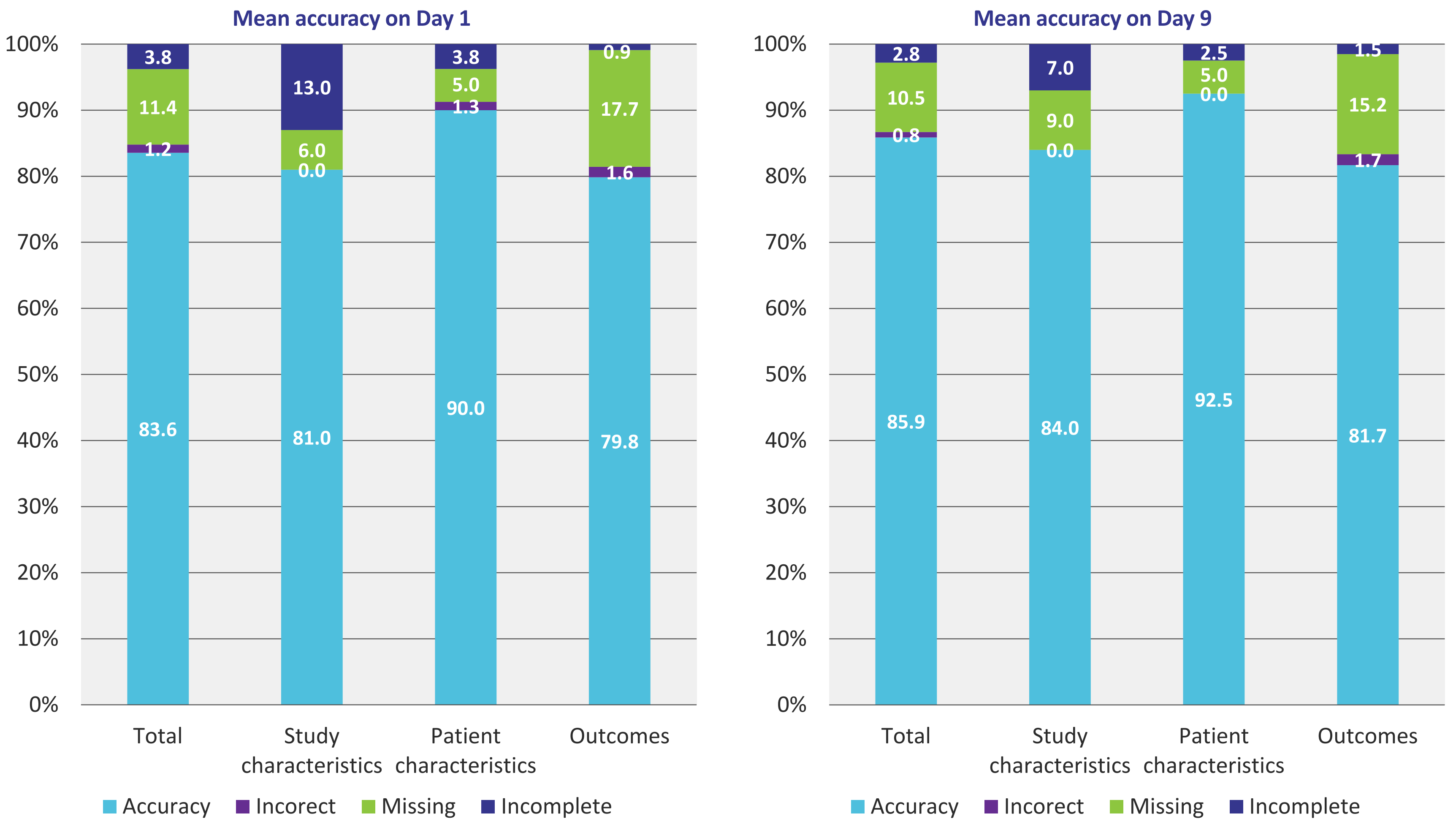
Results

- The variables extracted were in a range of formats: patient characteristics included four text, one binary, two categorical, and one continuous variable; outcomes included six text, four binary, and one continuous variable; and study characteristics consisted of 10 text variables (**Figure 2**).
- The mean overall accuracy was 84% (range: 66%–96%) across all RCTs (**Figure 3**). None of the studies achieved an overall accuracy of 100%.
- Accuracy was highest for patient characteristics, with a mean of 90% (range: 75%–100%), followed by study characteristics (81% [60%–100%]) and outcomes (80% [55%–100%]). Notably, 100% accuracy was achieved in four publications for patient characteristics, two for study characteristics, and only one for outcomes.

Results (cont.)

- Missing data ranged from 4%–28% and were more frequent for outcome variables.
- Incorrect extractions ranged from 0%–5%, observed mostly in complex quantitative variables such as the mean change from baseline at week 16 for DLQI, which AI extracted as baseline DLQI.
- Incomplete data ranged from 0%-8% and were more frequent for study characteristics, which were text variables, with a mean (range) of 13% (0-45.5%).
- When AI data extraction was replicated eight days later on the same publications using identical prompts, results varied compared with Day 1 (**Figure 3**). The ICC for patient characteristics, outcomes, and study characteristics variables were 0.95 (excellent), 0.85 (good), and 0 (poor), respectively. Most differences were observed in text variables, which AI extracted with either truncated or supplemented data across both days.

Figure 3. Mean accuracy measures



Discussion

- Based on Day 1 extractions, LLM-assisted data extraction demonstrated relatively high accuracy, with six publications achieving an overall accuracy of 85% or higher. However, none of the studies achieved a 100% accuracy overall, and performance differed by variable category.
- The LLM presented some limitations in accuracy and reliability when extracting data from RCTs:
  - Missing data were more frequent in outcome variables, primarily due to the LLM's inability to read data presented in figures. Such data would then need to be extracted manually by a human, until LLMs advance to a point where they can interpret figures.
  - Incorrect extractions were generally due to the incorrect selection of assessment timepoints and summary statistics of outcomes, particularly for complex quantitative variables.
  - Incomplete data were more commonly observed in study characteristics, suggesting a higher propensity for incompleteness for qualitative data such as text variables.
- When tested at two different times (Day 1 vs. Day 9), AI-extracted data exhibited a lack of reliability, despite using the same set of prompts on the same publications. Given that human-only extractions and validations are typically conducted over several weeks, it is crucial for AI to be reliable and reproduce the same results over time when supporting the data extraction process.

Conclusions

- LLM's rapid data processing can considerably reduce the resources and time required for data extraction in SLRs. However, this study demonstrated that substantial human input is currently necessary to achieve optimal results and avoid errors and missing data. This is particularly crucial in SLRs but may not pose a problem in rapid or scoping reviews that allow more flexibility. Hence, the decision to use AI to support evidence synthesis activities should carefully consider its advantages (e.g., time savings) and challenges (e.g., risks in accuracy and reliability).
- Future research should assess the differences in AI's capability to extract various types of data, particularly in complex indications and contexts. Additionally, the reproducibility of AI-assisted data extraction should be rigorously tested.
- Further investigation is required to optimize prompt designs to improve the accuracy and reliability of AI-assisted data extraction. This includes expanding abbreviations, incorporating synonyms, and providing definitions for scientific terms.

References

1. Merola J , et al. *Br J Dermatol*. 2023;10:10.
2. Katoh N , et al. *Dermatol Ther*. 2023;13(1):221-234.
3. Bieber T , et al. *Br J Dermatol*. 2022;187(3):338-352.
4. Gutermuth J , et al. *Br J Dermatol*. 2022;186(3):440-452.
5. Reich K, et al. *Lancet*. 2021;397(10290):2169-2181.
6. Silverberg J , et al. *J Allergy Clin Immunol*. 2020;145(1):173-182.
7. Reich K, et al. *JAMA Dermatol*. 2020;156(12):1333-1343.
8. Guttman-Yassky E , et al. *J Am Acad Dermatol*. 2019;80(4):913-921.e9.
9. Simpson E , et al. *J Am Acad Dermatol*. 2018;78(5):863-871.e11.
10. Blauvelt A , et al. *Lancet*. 2017;389(10086):2287-2303.

Disclosures/Acknowledgments

All authors are employees of Evidera, a business unit of PPD, part of Thermo Fisher Scientific. Mahmoud Slim was an employee of Evidera at the time this study was conducted. The views expressed in this study are those of the authors and not necessarily those of Evidera. The study was not funded by any organization. The authors do not have any conflicts of interest.

Presented at the ISPOR Europe Conference ● 17–20 November 2024 ● Barcelona, Spain

Funding provided by Evidera Inc., a business unit of PPD, part of Thermo Fisher Scientific