Stability of a Large Language Model for Data Extraction in Systematic Literature Reviews

MSR105

Aiswarya Shree[1], Mariana Farraia[2], Carolina Casañas i Comabella[3], Allie Cichewicz[4]

[1]Thermo Fisher Scientific, Bengaluru, India; [2]Thermo Fisher Scientific, Ede, Netherlands; [3]Thermo Fisher Scientific, London, UK; [4]Thermo Fisher Scientific, Waltham, MA, USA

Background

- Artificial Intelligence (AI) has been extensively explored in systematic literature reviews (SLRs) to save time and reduce human error, including the ability of large language models (LLMs) to perform data extraction has been tested.[1,2]
- We previously reported the high accuracy (84%, range: 66% to 96%) of an LLM for data extraction in an SLR of randomized controlled trials (RCTs).[3] However, we noted variations in responses when the same prompts were used on different days.
- Recently, the UK's National Institute for Health and Care Excellence (NICE) released a position statement on the use of AI for evidence generation, raising concerns about the reproducibility of AI, particularly regarding automated data extraction.[4]
- To our knowledge, there is limited evidence on the reproducibility and reliability of LLMs for data extraction, particularly given that some response variability is expected with these models and the impact on the trustworthiness of LLM-extracted data has not yet been characterized.

Objectives

- **This study aimed to evaluate the reproducibility and reliability of LLM-extracted data when considering variations with the same user or different users from two geographic locations. Additionally, the accuracy of LLM-extracted data was compared with manual human extractions from a previously conducted traditional SLR.**

Methods

- Three previously developed[3] one-shot prompts were used to extract 29 variables from five RCTs on atopic dermatitis[5-9] (**Table 1**).

Table 1. Variables for data extraction

Category	Variable Type	Variables
Study characteristics	Free-text fields	Author, year, trial name, phase, population description, intervention, comparator, inclusion criteria, exclusion criteria, and overall sample size
Patient characteristics	Free-text fields	Author, year, treatment arm
	Numeric fields	Sample size, mean age, male sex (%), comorbidities (%), disease severity (%)
Outcomes	Free-text fields	Author, year, treatment arm, analysis population, time point
	Numeric fields	Sample size, mean CFB in DLQI score, EASI 75 (%), POEM (%), treatment discontinuation (%), serious adverse events (%)

Numeric fields included binary, categorical, and continuous variables.
Abbreviations: CFB = change from baseline; DLQI = Dermatology Life Quality Index; EASI = Eczema Area and Severity Index; POEM = Patient-Oriented Eczema Measure

- Two reviewers (AS and MF) used the same LLM prompts, with creativity set to 0, and the same publications to test data extraction by the LLM for reproducibility and reliability between responses (**Figure 1**) and accuracy compared with human extraction.
  - The first set of LLM responses obtained by AS in the reproducibility test served as the reference and was compared with the second LLM extraction by AS (reproducibility) and MF (reliability). Reproducibility and reliability were calculated as the proportion of variables where LLM-extracted content was the same between responses, considering both content and formatting.
  - Accuracy was calculated as the proportion of correctly extracted variables compared with the validated human extractions (conducted previously).[3]

Methods (cont.)

Figure 1. Methods to assess reproducibility, reliability, and accuracy

- ***Reproducibility*** *refers to the consistency of results when the same prompts were deployed by the same user on the same day, ensuring that the data extraction process using LLMs can be replicated under identical conditions.*
- *The same user (AS) deployed the extraction prompts twice on the same publications within the same day.*
- ***Reliability*** *pertains to the consistency of the extracted data when different users applied the same prompts, highlighting the robustness of the process across different individuals.*
- *Two different users located in different geographic areas (AS in India and MF in the Netherlands) deployed the extraction prompts on the same publications within the same day.*
- ***Accuracy*** *refers to the correctness of the LLM responses for each variable out of the total number of variables when compared with a fully validated human extraction.*
- *The first responses by AS and MF in the reproducibility and reliability set, respectively, were compared with manual human extractions.*

Abbreviations: AS = Aiswarya Shree; LLM = large-language model; MF = Mariana Farraia

Results

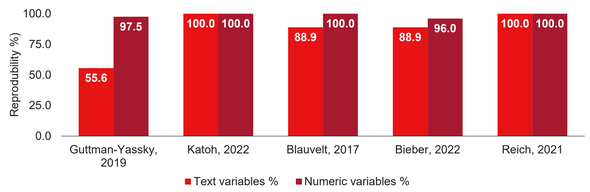
Reproducibility

- Reproducibility of LLM responses to the data extraction prompts ranged from 80.6% to 100% (**Figure 3**).
  - Reproducibility of responses for patient characteristics and outcome variables ranged from 88.5% to 100% and 72.7% to 100%, respectively.
  - Reproducibility of responses for study characteristics varied from 80% to 100%.
- Reproducibility of responses for text vs. numeric variables are displayed in **Figure 4**.
  - Reproducibility of responses for text variables was >85% in four studies, except for one study with 55.6%. Reproducibility for numeric variables was >95% in all studies.

Figure 3. Overall reproducibility and reliability; and by categories

Results (cont.)

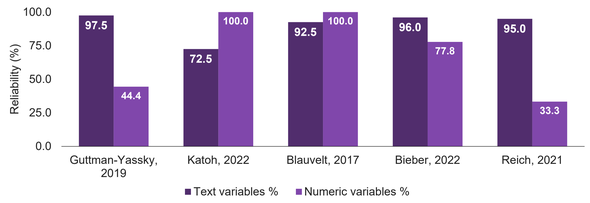
Figure 4. Reproducibility for text vs. numeric (binary, categorical, and continuous) variables



Reliability

- Reliability was lower than reproducibility, ranging from 65.7%–95.5% (**Figure 3**).
  - Reliability for patient characteristics and outcome variables ranged from 62.5% to 100% and 63.7% to 100%, respectively.
  - Reliability for study characteristics variables was the lowest, ranging from 50% to 90%.
- Reliability in text vs numeric variables.
  - None of the extractions were 100% reliable regarding text variables (range: 72.5% to 97.5%). Reliability of numeric variables varied considerably: responses for two studies had <50% reliability, while two others had 100% (**Figure 5**).

Figure 5. Reliability for text vs. numeric (binary, categorical, and continuous) variables



Accuracy

- Compared with validated human extractions, the LLM did not achieve an overall extraction accuracy of 100% for any publications in either test. The accuracy of LLM-extracted data was slightly higher with the reproducibility set of responses (79.1% to 98.5%) compared with the reliability set (74.6% to 97.7%).
  - Accuracy with text variables: Overall accuracy remained consistent between response sets as the LLM captured the same underlying information for text variables. Despite slight variations in syntax, style, or length of response, this did not impact the accuracy, as the content of the LLM-extracted data aligned with the manual human reference extraction.
  - Accuracy with numeric variables: Accuracy was adversely affected by the extraction of numeric fields, where discrepancies were observed between LLM responses and the human reference extraction.

Discussion

- This study highlights both the benefits and potential challenges of using LLMs for data extraction from RCTs. Our findings indicate high reproducibility rates, ranging from 80.6% to 100%, suggesting that LLMs can consistently replicate extraction of data under the same user conditions. However, reliability, which evaluates consistency between different users, was lower despite using identical prompts (range: 65.7% to 95.5%).

Discussion (cont.)

- Overall, the accuracy of LLMs for data extraction was high.
  - The observed variations in text responses did not negatively affect the overall accuracy of data extraction.
  - However, in addition to formatting inconsistencies, the extraction of numeric fields created discrepancies that led to errors that affected accuracy.
- SLRs require meticulous data extraction processes, often involving multiple extractors, followed by thorough data validation. This process may take several weeks (depending on study volume) and is susceptible to human error (up to 50%).[10,11] Using an LLM-based extraction approach may result in faster, more consistent and reliable results, particularly over time and across different users, and potentially reduce human error and increasing extraction quality.
- Additional research is needed to understand and mitigate the factors contributing to variation in LLM-extracted data, including developing techniques to improve the consistency of numeric data extraction and to reduce the impact of stochastic elements in LLMs.

Limitations

- Our findings may not be generalizable to other types of studies beyond RCTs, as different study designs and reporting of outcomes may introduce additional challenges.
- This study tested a small sample of publications. Larger volumes of data might exhibit higher variations in LLM responses, particularly if variables are reported more heterogeneously across publications.
- One-shot prompting was used; a more iterative prompting approach might achieve better accuracy, but could also introduce further variations between users, potentially impacting reproducibility and reliability.
- Testing was conducted at a specific time point (November 2024). Given how quickly LLMs are evolving, future testing may demonstrate different or improved results.
- Only one LLM was tested. Other LLMs may yield different results, and further research is needed to compare the performance of various models in data extraction tasks.

Conclusions

- **When using LLMs for data extraction in SLRs, reproducibility was generally high, but reliability was affected by user interaction, with variations between different users leading to discrepancies, particularly for numeric data. These findings highlight the need for human validation of LLM-extracted data to ensure data quality.**
- **While variations in text responses are expected with LLMs and the impact on overall accuracy was minimal, numerical discrepancies highlight the need for human oversight such that AI-driven extractions must still undergo human validation to ensure data accuracy.**
- **Transparent reporting of AI-assisted methods in SLRs is crucial to contextualize results and maintain scientific rigor. Our findings underscore the importance of addressing the concerns raised by NICE's AI position statement and are important for informing future Cochrane guidance on the applications of AI in SLRs.**

References

1. Konet et al., *Res Synth Methods*. 2024;15(5):818-824; **2.** Gartlehner et al., *Res Synth Methods*. 2024;15(4):576-589. **3.** Shree et al., Elsevier Science Inc., 2024; **4.** NICE. https://www.nice.org.uk/about/what-we-do/our-research-work/use-of-ai-in-evidence-generation--nice-position-statement 2025; **5.** Bieber et al., *Br J Dermatol*, 2022;187(3):338-352; **6.** Blauvelt et al., *Lancet*. 2017;389(10086):2287-2303; **7.** Reich et al., *Lancet*. 2021;397(10290):2169-2181; **8.** Guttman et al., *J Am Acad Dermatol*. 2019;80(4):913-921. e9; **9.** Katoh et al., *Dermatol Ther*. 2023; 13(1):221-234; **10.** Tang et al., *Postgrad Med J*. 2025:qgae195; **11.** Mathes et al., *BMC Med Res Methodol*. 2017;17:1-8.

Disclosures

AS, **MF**, **CCC**, and **AC** are employees of PPD™ Evidera™ Health Economics & Market Access, Thermo Fisher Scientific.

Acknowledgments

Editorial and graphic design support were provided by Sean Smith and Kawthar Nakayima of Thermo Fisher Scientific.

Presented at ISPOR—The Professional Society for Health Economics and Outcomes Research | 13–16 May 2025 | Montreal, Quebec, Canada