Using a Large Language Model (LLM) for Data Extraction of Studies: Learnings from a Targeted Literature Review (TLR) in Non-small Cell Lung Cancer (NSCLC)

SA101

Mariana Farraia,[1] Anuja Pandey,[2] Eugenia Priedane,[3] Allie Cichewicz,[4] Caroline von Wilamowitz-Moellendorff[2]

[1]Thermo Fisher Scientific, Ede, Netherlands; [2]Thermo Fisher Scientific, London, UK; [3]BeOne Medicines (UK), Ltd., London, UK; [4]Thermo Fisher Scientific, Waltham, MA, USA

Background

- Large Language Models (LLMs) streamline literature reviews by reliably extracting relevant information, thus reducing manual effort.
- We previously reported:
  - High accuracy (84%, range: 66%–96%) using zero-shot prompts for extraction of disease-specific clinical outcomes from a systematic literature review (SLR) of randomized controlled trials (RCTs) in atopic dermatitis.[1]
  - High reproducibility rates (80.6%–100%) under the same human operator and conditions, but lower reliability (65.7%–95.5%) between different operators using identical prompts.[2]
- While extraction has been tested on a broad range of therapeutic areas using RCT data,[3,4] real-world evidence (RWE) studies and complex research questions, such as those involving specific subpopulations, are not fully understood.

Objectives

- **This study aimed to evaluate a zero-shot GPT-4-assisted extraction approach for clinical outcomes from RWE studies in a specific cancer subpopulation to address the following objectives:**
  - **Identify and document challenges during data extraction, especially in RWE designs and studies with mixed populations, focusing on subsets of interest.**
  - **Provide recommendations for using LLMs in these instances.**

Methods

- A targeted literature review (TLR) assessed the comparative effectiveness and safety of immunotherapy treatments for non-small cell lung cancer (NSCLC) patients with programmed death-ligand 1 (PD-L1) expression ≥50% in RWE studies. Sixteen publications covering ten observational primary studies were included.
- Zero-shot prompts were developed, tested, and optimized for a proprietary GPT-4-based LLM using one included publication. Once satisfactory output was achieved (i.e., all predefined fields extracted with correct formatting and no critical omissions or errors after human validation), data were then extracted by the LLM for each remaining publication individually.
  - Zero-shot prompt: input given to a LLM to perform a task without any prior specific training or examples for that task; the model relies solely on the prompt and the provided documents
- Data extracted by the LLM were validated by a human investigator, with the main challenges noted during the process.
- Related publications were identified and categorized manually by a human reviewer.

Results

- The main challenges identified during the extraction and validation process were:
  1. Difficulties in isolating data for the target population (PD-L1 ≥50%) when reported as a subset of a broader NSCLC population with diverse PD-L1 expression statuses.
  2. Incorrect or missing data extracted by the LLM for subgroups (e.g., sex, age, smoking status) and fields related to study characteristics (e.g., follow-up duration, data source).
  3. Multiple files for a single study caused discrepancies from data presentation (tables, text, figures), multiple data types and timepoints, variations between main text and supplemental materials, and related publications.
- As a result of these challenges, additional extraction and re-validation of subgroup data, along with correction of formatting issues, resulted in time expenditure equal to or greater than validating manual extractions.

Limitations and recommendations

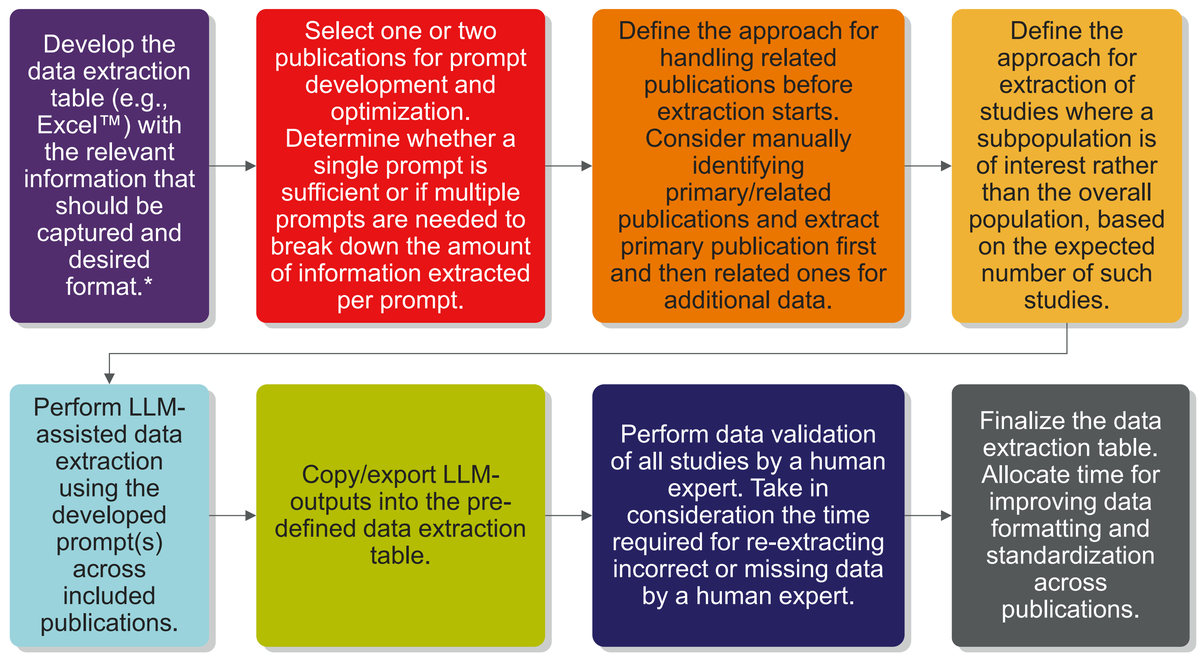
- For each identified challenge, the limitations with the extraction approach are presented in **Table 1**, along with a suggested workflow and considerations for future projects using LLMs for data extraction (**Figure 1**).

Table 1. Challenges and corresponding recommendations for future use

Data extraction challenge	Limitations	Recommendations
Mixed population studies: difficulties identifying data from the target population (PD-L1 ≥50%) within a broader population (NSCLC)	• LLM often extracts data for the overall population instead of the subpopulation of interest.	• Develop specific prompts for studies with subpopulations of interest: – Include detailed context and clearly indicate that only data for the target subpopulation are of interest.
Incorrect or missing data for subgroups	• LLM may extract incorrect/incomplete subgroup data. • The LLM used for this study did not recognize data in figures.	• Enhance prompt specificity for subgroup data. • Include detailed instructions and examples in the prompt.
Lack of recognition of related publications	• The prompt was not designed to enable the LLM to identify related publications, resulting in some redundant extractions.	• Consider two approaches: 1. Identify related publications manually and design specific prompts to extract only new data. 2. Identify related publications, compare data manually, and use LLM for extraction only if new data points are present, refining the prompt as needed.
Supplementary materials	• Multiple documents (such as the main publication and its supplement) complicate LLM processing, especially in mixed population studies with extensive subgroup data.	• Enhance prompt instructions for handling multiple documents. • Specify prioritized information for extraction. • Note relevant subpopulation data may be in supplementary material.
Formatting issues	• Inconsistent formatting in extracted data requires additional processing and standardization.	• Standardize format requirements in prompts. • Provide clear examples.
Generic platform vs platform integrated within literature review software	• Using a generic LLM interface may increase time requirements due to manual steps needed to use it for literature reviews (e.g., identifying related publications, providing examples in prompts).	• LLMs integrated into a literature-review platform are trained on structured research data extraction variables and outcomes with higher fidelity for TLRs. • Built-in linkage across related publications reduces redundant data extraction from overlapping or duplicate studies.

Note: Limitations may not be generalizable across various LLM architectures, implementations, or platform integrations.
Abbreviations: LLM = large language model; NSCLC = non-small cell lung cancer; PD-L1 = programmed death-ligand 1

Figure 1. Suggested workflow diagram and considerations for future projects using LLM for data extraction



Some LLMs now support reading Excel™ files. If available, uploading the Excel extraction table for direct population may streamline the workflow. At the time of this work, this functionality was not available
Abbreviations: LLM = large language model

Conclusions

- **LLMs can extract data from RWE studies but face significant challenges with clinical outcomes, especially for subgroups and mixed populations. Heterogeneity and limited standardization in reporting of observational studies likely contributed to errors in LLM-assisted extraction with zero-shot prompts.**
- **Further research into prompt engineering (i.e., design and execution) is needed to improve efficiency and accuracy in LLM-assisted data extraction for RWE study designs and complex research questions, particularly across multiple files per study and various types of data. Additional considerations are needed for generic LLMs versus. SLR-specific tools.**
- **Future projects should consider and evaluate the time required for human re-extraction, re-validation, and prompt development/optimization to accurately assess potential time savings.**

References

1. Shree A, et al. *Value Health*. 2024;27(12):S475.
2. Shree A, et al. *Value Health*. 2025;28(6):S295.
3. Gartlehner G, et al. *Res Synth Methods*. 2024;15(4):576-89.
4. Schmidt L, et al. *F1000Res*. 2021;10:401.

Disclosures

MF, AP, CvWM are employees of PPD™ Evidera™ Health Economics & Market Access, Thermo Fisher Scientific. **AC** was employed by Thermo Fisher Scientific at the time this study was conducted. This poster was funded by Thermo Fisher Scientific.

Presented at ISPOR —The Professional Society for Health Economics and Outcomes Research—Europe | 9–12 November 2025 | Glasgow, Scotland, UK