Improving the Performance of Generative AI to Achieve 100% Accuracy in Data Extraction

S L Klijn,[1] S Teitsson,[2] T Reason,[3] B Malcolm,[2] N Hill,[1] E Benbow [3]

[1]Bristol Myers Squibb, Princeton, NJ, USA; [2]Bristol Myers Squibb, Uxbridge, UK ; [3]Estima Scientific, London, UK

Key Messages

What's already known on this topic

As part of a systematic review or NMA, there is a need to extract data from relevant publications. This is a time-consuming and error-prone part of this process and usually requires input from a minimum of three people[1-8].

It has previously been demonstrated that there is potential to use LLMs, such as GPT-4, to automate data extraction for NMA[9]. However, the stochastic nature of LLMs can affect the extraction success rate if the LLM is only asked to extract the data once. For LLM data extraction to become acceptable for use in practice, it should achieve human-like performance, or better.

What this study adds

This study has demonstrated that very near-perfect data extraction can be achieved when implementing a simple modal approach to using LLMs to extract data from trial publications.

How this study might affect research, practice, or policy

The ability to use LLMs to automate data extraction and analysis could result in significant time savings, increased concordance and reduced human error in the systematic review and NMA process. There is a need to test LLM-based processes developed across a greater number of disease areas and outcome types to demonstrate the generalizability of the approach and the general level of performance that can be achieved implementing a modal approach.

Introduction

- The emergence of artificial intelligence (AI), capable of human-level performance on some tasks[10,11], presents an opportunity to revolutionize development of systematic reviews and network meta-analyses (NMAs).
- We have previously demonstrated that there is potential to use large language models (LLMs), such as GPT-4, to automate data extraction for NMA[9]. Whilst data extraction accuracy of over 97% was achieved, there is scope to explore how output from LLM models can yield data extraction with performance and reliability of 100%, to provide guidance for best practices for full implementation in health economics outcome research (HEOR).
- Reproducing results with LLMs can be difficult because of their stochastic nature[12], i.e. LLMs do not produce deterministic results. Therefore, in the previous work we asked GPT-4 to extract the data 20 independent times, to capture the variation in performance[9]. Whilst the performance was not perfect, we noticed that, in the vast majority of iterations, GPT-4 extracted the correct data.
- Extending this prior work, an a priori defined modal algorithm was therefore postulated, developed, and tested.

Aims

- Using four case studies, the aim was to assess whether asking GPT-4 to extract data multiple times, and then calculating the most frequently occurring (mode) of these answers, would improve the already good data extraction rate seen and whether perfection could be achieved (i.e., 100% data extraction accuracy).

Methods

- In previous work, we tested GPT-4's ability to extract data using four case studies. These four studies required GPT-4 to review publications and extract the data required to conduct an NMA from the text. The case studies covered three different outcomes (two time-to-event [TTE] and one binary) and two disease areas (metastatic non-small-cell lung cancer and moderate-to-severe hidradenitis suppurativa).
- For each case study, we wrote a Python script that sent a prompt via an API call to the LLM for each publication in the NMA. The prompt included text from the publication and a request to extract all relevant data from the supplied publication text (Figure 1). Detail of the case studies and the process used is provided in Reason et al., 2024[9].

Figure 1. LLM-based process for extracting data required for NMA

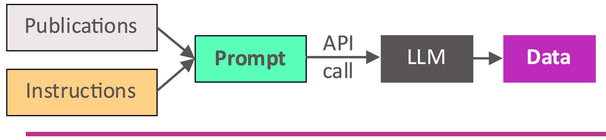
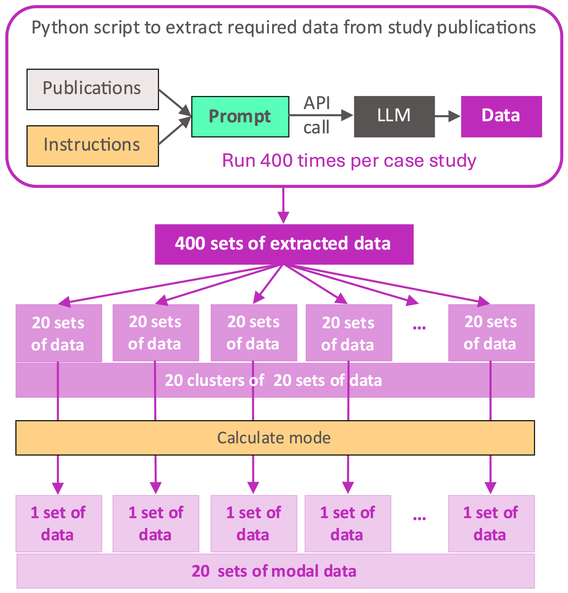


Figure 2. Modal approach to data extraction



- We used the same four case studies as used for the previous research and ran the Python script for each 400 times. Thus, for each case study, we obtained 400 sets of GPT-4's attempts to extract data per publication (Figure 2).
- The 400 sets of data were then sequentially divided into 20 groups of data sets (Figure 2), and the "mode" calculated (most commonly occurring value) for each data point, within the Python script. This resulted in 20 sets of modal data. 20 groups of 20 datasets was chosen for consistency with the previous work[9].
- Results, for both the 400 sets of data, and the 20 sets of modal data, were then compared with the results of the data extraction conducted (and checked) by (human) systematic literature review and NMA experts.

Results

Results of 400 individual runs (Table 1)

- For case study 1, forty individual items of data needed to be extracted from the publications, in order to inform an NMA for this outcome and patient population. When considering individual runs and individual data items, GPT-4 extracted the correct value for between 318 (clinical response, PIONEER 1 trial[9]) and 400 of the runs (e.g., clinical response, SUNSHINE trial[9]).
- For Case Study 2, forty-one individual items of data needed to be extracted from the publications, in order inform an NMA for OS in patients with mNSCLC. When considering individual runs and individual data items, GPT-4 extracted the correct value for between 397 (number at risk, CheckMate017 trial[9]) and 400 of the runs (e.g., hazard ratio, OAK trial[9]).
- For Case Study 3, eighty-six individual items of data needed to be extracted from the publications, in order to inform a sensitivity analysis NMA for OS in patients with mNSCLC. When considering individual runs and individual data items, GPT-4 extracted the correct value for between 304 (hazard ratio, H3E_MC_JMID trial[9]) and 400 of the runs (e.g., hazard ratio, JMEI trial[9]).
- For Case Study 4, forty-one individual items of data needed to be extracted from the publications, in order inform an NMA for PFS in patients with mNSCLC. When considering individual runs and individual data items, GPT-4 extracted the correct value for between 154 (confidence level on the hazard ratio, OAK trial[9]) and 400 of the runs (e.g., hazard ratio, OAK trial[9]).

Table 1. GPT-4 Extraction Accuracy

Type of data item	For all 400 data sets			For all 20 modal data sets		
	Min	Max	Mean	Min	Max	Mean
Binary outcome, hidradenitis suppurativa						
Treatment names	95.5%	100%	98.3%	100%	100%	100%
Numbers at risk	85.3%	99.8%	96.0%	100%	100%	100%
Clinical responses	79.5%	100%	94.3%	100%	100%	100%
Time-to-event outcome, mNSCLC 1						
Treatment comparisons	99.8%	100%	99.96%	100%	100%	100%
Numbers at risk	99.3%	100%	99.9%	100%	100%	100%
Hazard ratios	100%	100%	100%	100%	100%	100%
Confidence interval limits	100%	100%	100%	100%	100%	100%
Confidence interval levels	100%	100%	100%	100%	100%	100%
Time-to-event outcome, mNSCLC 2						
Treatment comparisons	96.8%	100%	99.6%	100%	100%	100%
Numbers at risk	79.5%	100%	97.7%	100%	100%	100%
Hazard ratios	76.0%	100%	97.9%	100%	100%	100%
Confidence interval limits	76.0%	100%	97.8%	100%	100%	100%
Confidence interval levels	97.5%	100%	99.8%	100%	100%	100%
Time-to-event outcome, mNSCLC 3						
Treatment comparisons	98.3%	100%	99.3%	100%	100%	100%
Numbers at risk	95.3%	100%	98.1%	100%	100%	100%
Hazard ratios	98.3%	100%	99.4%	100%	100%	100%
Confidence interval limits	98.0%	100%	99.4%	100%	100%	100%
Confidence interval levels	38.5%	100%	84.9%	30.0%	100%	88.3%

Results of 20 modal datasets (Table 1)

- Whilst individual runs were not always perfect, when the modal value was calculated from 20 runs, the correct data was accurately extracted 20 out of 20 times for Case Studies 1 to 3 i.e., using a modal approach resulted in perfect data extraction every time for these case studies.
- For Case Study 4, when the modal value was calculated from 20 runs, the correct data was accurately extracted 20 out of 20 times for almost all data items i.e., using a modal approach resulted in perfect data extraction every time. The one exception to this was the level of the confidence interval on the hazard ratio for the OAK study, where GPT-4 failed to extract a value more times than it extracted the required value (of 95%). Thus, the modal confidence interval was only correct for 6 of the 20 modal values and the rest of the time, GPT-4 provided "Not reported" as its response.

Discussion

- There is some up-front time and investment required to develop prompts with which to instruct the LLM to extract data from publication text. However, once this has been done for one disease area and type of outcome, it should be easily adapted to others: We believe that it is possible to use the methods developed herein for any disease area and any outcome, and the prompts used to communicate with the LLM are not specific for use only with GPT-4 but can be used with other LLMs.
- The level of performance demonstrated by GPT-4 suggests that, if using a modal approach, LLMs can offer the same, or an even better level of accuracy for data extraction from text than is generally achieved by humans[1-8], taking a fraction of the time and costing significantly less (full data extraction by a human can take up to 45 minutes per publication, whereas an LLM takes <5 minutes to do the same task, costing $1 or less). Concomitant data extraction from multiple publications is also possible when using an LLM, which could result in even greater time saving compared to human extraction.
- Whilst a modal approach requires the LLM to extract data several times (e.g., 20 times, as in this study), these extractions can be conducted concomitantly, so the time taken per publication could still be < 5 minutes.
- In the one case, where taking the mode of the LLM's responses did not result in perfect data extraction (confidence level for hazard ratio for one study), the LLM did still extract the correct data for several of the runs contained in the modal calculation. By ignoring "Not reported" responses, and calculating the mode of the remaining responses, the correct data would have been obtained. In this case, it would be essential to report a level of confidence in the response (e.g., percentage of runs reporting the modal value), perhaps implementing a threshold below which human intervention is required.
- Agent-based approaches, potentially leveraging multiple LLMs, may further improve performance and would warrant further research.

Conclusions

- Whilst GPT-4 generally extracts data perfectly for the case studies considered, there are some rare occasions when it fails to extract all required data from a publication.
- We have demonstrated a useful method that improves the accuracy, repeatability and reliability of data extraction compared to single-pass data extraction.
- In three of the four case studies considered, we observed perfect extraction by GPT-4. For the fourth case study, almost perfect data extraction was observed, with just one data item not being consistently extracted.
- Work to determine the optimal number of datasets from which to calculate a "mode" is underway, along with testing data extraction in other disease areas and for other outcomes.

References

1. Higgins J, Thomas J, Chandler J, et al. Cochrane Handbook for Systematic Reviews of Interventions Version 6.4. Cochrane; 2023. www.training.cochrane.org/handbook
2. Mathes T, Klaßen P, Pieper D. Frequency of data extraction errors and methods to increase data extraction quality: a methodological review. *BMC Med Res Methodol.* 2017;17(1):152. doi:10.1186/s12874-017-0431-4
3. Saldanha IJ, Schmid CH, Lau J, et al. Evaluating Data Abstraction Assistant, a novel software application for data abstraction during systematic reviews: protocol for a randomized controlled trial. *Syst Rev.* 2016;5(1):196. doi:10.1186/s13643-016-0373-7
4. Carroll C, Scope A, Kaltenthaler E. A case study of binary outcome data extraction across three systematic reviews of hip arthroplasty: errors and differences of selection. *BMC Res Notes.* 2013;6(1):539. doi:10.1186/1756-0500-6-539
5. Horton J, Vandermeer B, Hartling L, Tjosvold L, Klassen TP, Buscemi N. Systematic review data extraction: cross-sectional study showed that experience did not increase accuracy. *J Clin Epidemiol.* 2010;63(3):289-298. doi:10.1016/j.jclinepi.2009.04.007
6. Gøtzsche PC, Hróbjartsson A, Marić K, Tendal B. Data Extraction Errors in Meta-analyses That Use Standardized Mean Differences. *JAMA.* 2007;298(4). doi:10.1001/jama.298.4.430
7. Buscemi N, Hartling L, Vandermeer B, Tjosvold L, Klassen TP. Single data extraction generated more errors than double data extraction in systematic reviews. *J Clin Epidemiol.* 2006;59(7):697-703. doi:10.1016/j.jclinepi.2005.11.010
8. Jones AP, Remmington T, Williamson PR, Ashby D, Smyth RL. High prevalence but low impact of data extraction and reporting errors were found in Cochrane systematic reviews. *J Clin Epidemiol.* 2005;58(7):741-742. doi:10.1016/j.jclinepi.2004.11.024
9. Reason T, Benbow E, Langham J, Gimblett A, Klijn SL, Malcolm B. Artificial Intelligence to Automate Network Meta-Analyses: Four Case Studies to Evaluate the Potential Application of Large Language Models. *PharmacoEconomics - Open.* 2024;8(2):205-220. doi:10.1007/s41669-024-00476-9
10. Wang A, Pruksachatkun Y, Nangia N, et al. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. Published online 2019. doi:10.48550/ARXIV.1905.00537
11. Adiwardana D, Luong MT, So DR, et al. Towards a Human-like Open-Domain Chatbot. Published online 2020. doi:10.48550/ARXIV.2001.09977
12. Edwards B. As ChatGPT gets "lazy," people test "winter break hypothesis" as the cause. *ARS Technica.* Published December 11, 2023. Accessed December 12, 2023. https://arstechnica.com/information-technology/2023/12/is-chatgpt-becoming-lazier-because-its-december-people-run-tests-to-find-out/

Email: Sven.Klijn@bms.com