Generative AI: A Novel Approach to Data Extraction for NMAs in EU JCA

HTA410

Yunchou Wu,[1] Siguroli Teitsson,[2] Cheryl Jones,[1] Bill Malcolm,[2] Nebibe Varol,[2] Emma Benbow,[1] Tim Reason,[1] Sven Klijn[3]

[1]Estima Scientific Ltd, Ruislip, LON, UK ; [2] Bristol Myers-Squibb, Uxbridge, LON, UK; [3]Bristol Myers-Squibb, Princeton, NJ, USA

Introduction

- EU HTA Regulation's Joint Clinical Assessments (JCA) aim to harmonize the clinical assessment of medical interventions across all European Union (EU) member states.[1] EU member states will be required to submit their PICO (population, intervention, comparator, outcomes) sets of interest.[2] The JCA assessors will then consolidate those PICOs, removing any duplicates, and report a final set of PICOs to health technology developers (HTDs).

- The process will likely require HTDs to conduct a large number of comparative clinical analyses and submit the results within a tight timeframe (i.e. up to 100 days or up to 60-days for accelerated procedure or variation to the terms of an existing marketing authorization). The need to complete a potentially large volume of systematic literature reviews (SLRs) and network meta-analyses (NMAs) within a short period of time has provided the impetus to investigate the extent to which those specific tasks involved in conducting clinical analyses can be automated.

- Large language models (LLMs) have previously demonstrated proficiency for extracting data from text within trial publications.[3] Previous research utilised LLMs (OpenAI's GPT-4) to extract specific outcomes data, such as, but not limited to, numbers at risk, treatment names, hazard ratios, means or medians, confidence intervals, from published trials in metastatic non-small cell lung cancer (mNSCLC) and hidradenitis suppurativa.[3] The self-consistency approach (which asked GPT-4 to extract the data multiple times and then select the most frequently occurring [mode] answer)[4], was also applied.

- Often, information that is required to populate SLR tables or NMA datasets, are not reported in the text of a publication but rather such information is embedded within tables and figures.

Aim

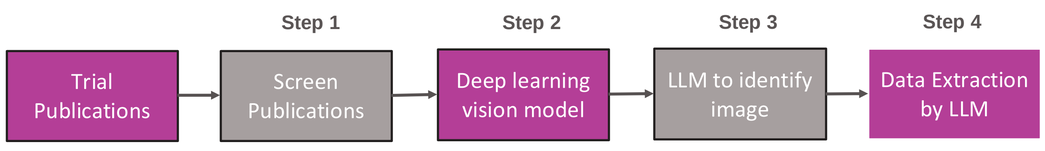
- The purpose of this research was to expand data extraction from text only and investigate the extent to which LLMs can accurately extract data from tables and figures within trial publications.

- The ability to automate and accurately extract data from text, tables, and figures from within trial publications could transform current approaches used to develop systematic reviews and network meta-analyses (NMAs). Such automation methods could help HTDs meet JCA requirements.

Methods

- **Figure 1** presents an overview of the LLM-based process for extracting data from tables and figures

Figure 1. LLM-based process for extracting data from tables and figures



Abbreviations: *LLM: Large language model*

Steps required to extract data were as follows:

1. Python was used to screen the pages within each publication and identify pages that contain a pre-defined set of key words, for example "baseline", "Kaplan Meier", and "forest plot".
2. A deep learning model, Paddle OCR,[5] published under an open-source license, was used to parse and save all tables and figures, from the pages identified in step 1, as "png" files. The deep learning model was adapted to include titles and legends within the extraction.
3. LLMs (GPT-4o [06/08/2024] and Claude 3.5 Sonnet [20/06/2024]) were used to identify: whether the image was a table or figure and what type of table or figure the image is (e.g. baseline patient characteristics tables, forest plot, Kaplan Meier plot). All irrelevant tables and figures were discarded.
4. LLMs (GPT-4o and Claude 3.5 Sonnet) were then used to label each "png" image; for example, Baseline Patient Characteristics Table, Kaplan Meier Plot, Forest Plot.

Table 1. Tables and figures used to extract data

Tables/ Figures	No. of Examples	Description	References
		Tables	
Baseline demographics and clinical characteristics	5	• Baseline characteristics such as age, gender, etc. • Clinical characteristics such as clinical performance scores (ECOG), biomarkers, previous therapies • Data related to the whole sample and subgroups	6 - 10
		Figures	
Forest Plots	2	• Subgroup information, patients (n), ORR results, number of events, total number of events, hazard ratios, 95% CIs	7, 8
KM Plots	9	• Endpoint results, 95% CIs, number of patients at risk, hazard ratios, p-values, treatment arms, number of events, subgroups • Figures containing one or multiple separate KM plots: - 1 x KM plot; n = 1 - 2 x KM plots; n = 5 - 3 x KM plots; n = 1 - 4 x KM plots; n = 1 - 6 x KM plots; n = 1	6 - 10

Abbreviations: CI: Confidence Interval; ECOG: Eastern Cooperative Oncology Group*; n: number;* OOR: objective response rate

- The method was also tested for its ability to correctly identify and extract tables and figures in situations where tables or figures were spread across more than one page within a publication, a typical scenario in many manuscripts.
- Five trial publications,[6-10] which included a variety of charts, graphs, and tables, were used to identify tables and figures, followed by data extraction. Table 1 describes the types of tables and figures that were used to test the LLM's data extraction capabilities. Note, only text-based information was being extracted from Kaplan Meier plots and should not be confused with automated digitisation.
- The tables and figures listed in Table 1 varied in terms of their format and structure, as well as the type of information reported.
- Vision capability was required to conduct this task, therefore the latest vision-LLMs, Claude 3.5 Sonnet and GPT-4o, were tested for their ability to accurately extract data from tables and figures. The latest GPT-preview-o1 model was not used as it does not have vision capabilities.
- *Open access articles were used for this research*

Acknowledgments

- This study was supported by Bristol Myers Squibb

Methods: Assessing Accuracy

- The LLMs were first assessed for their ability to accurately extract relevant tables and figures as defined by the human user
- When assessing accuracy of data extraction from tables by LLMs, it was important to test a wide range of different formatting that such tables have. For example, non-uniform formatting or structure, different colors used to illustrate columns or rows and indentations used to mark results within the same category.
- To assess data extraction from figures, it was important to test the performance of the LLMs using different examples of figures with some including single or multiple graphs. This allowed understanding of whether LLMs were able to process subfigures in isolation and extract the data accordingly, rather than extracting information across any or all of the subfigures in an illogical manner.
- The data extracted from each table or figure was assessed for inclusiveness and accuracy against the original source; the definitions of inclusivity and accuracy are listed below:
  - Inclusiveness (Tables): Extract all information from columns and rows correctly
  - Inclusiveness (Figures): Capture all essential components within the figure (e.g. axis information, numbers at risk, hazard ratios, odds ratios, number of events, etc.)
  - Accuracy (Tables and Figures): Data extracted matches the data in the publication

Results

- The deep learning model used to cut and extract images of figures and tables from publications together with GPT4o and Claude 3.5 Sonnet achieved 100% accuracy when extracting tables and figures from relevant pages in a publication and saving them as images; this was inclusive of situations where tables were split across multiple pages.
- GPT-4o and Claude 3.5 Sonnet achieved 100% inclusivity and accuracy when extracting data from tables with text size above ~6pt.
- GPT-4o and Claude 3.5 Sonnet achieved 100% inclusivity and accuracy of information extracted from images of figures with text size ~6pt, including in cases where figures comprised multiple subfigures.
- Only in instances where the text size in the image was very small (<6pts), and not easily readable by a human, were LLMs unable to extract data and became susceptible to hallucinations.
- Accuracy of data extraction from tables or figures with small/illegible to human text was significantly improved by instructing the deep learning model to re-print the text increasing the size and contrast of the text embedded in the figure. The deep learning model combined with GPT-4o or Claude 3.5 Sonnet was able to extract data with an accuracy rate of 80% and 99%, respectively, see Table 2.

Table 2. Accuracy of Data Extraction by LLM from Tables

	Data Extraction Accuracy (%)			Increased text size and contrast[a]
	Text Size ~ > 6pt	Text size ~4 - 6 pt	Text size ~≤ 4pt	Text size ~≤ 4pt
	Tables			
GPT-4o	100%	99%	90%	-
Claude 3.5 Sonnet	100%	100%	99%	-
	Figures			
GPT-4o	100%	100%	10%	80%
Claude 3.5 Sonnet	100%	100%	60%	99%

Abbreviations: pt: point

a. Deep learning model (PaddleOCR[5]) was instructed to increase the text size and contrast to help the LLM accurately read and extract the text-data embedded within the figure

- One limitation with instructing the deep learning model to re-print the text was on rare occasions the text was not printed at 100% accuracy (we found two brackets were missing; other word and numeric text were printed correctly), therefore the original image (with the small text) was also sent to the LLM to provide it with the highest probability of extracting the correct data.

Conclusion

- Using these images, LLMs have demonstrated the ability to extract data with 100% accuracy; with the exception of figures which contain very small text which is difficult even for a human to interpret. Instructing the deep learning model to re-print the information contained in figures when the text is very small helped increase accuracy of the data being extracted .
- In reality, it is not common to find text as small as 4pt in publications; however, it may occur occasionally which makes reading difficult for humans and LLMs. Therefore, this may be a signal to journals to make legibility standards mandatory
- It is important to note that LLMs by themselves are not able to extract data to a high-degree of accuracy without also using a deep learning model to extract the initial images from the publications.
- The process developed to extract relevant clinical data from tables (patient demographics and clinical characteristics) and figures (forest plots and KM plots) to inform efficacy is generalizable and can be applied to extract data from various types of tables and figures published within clinical publications and reports
- Based on results from a JCA simulation study, assuming 800 publications require data extraction (text, tables and figures) we estimate substantial time-savings with approximately 243 hours [31 working days, assuming an 8-hour workday and 5 working days per week] of human only time compared with 56 hours of human plus automated data extraction using LLMs [7 working days assuming an 8-hr workday] leading to a saving of 24 working days.
- Such automation has the potential to significantly reduce burden on HTDs preparing for JCA submissions. Further research has also demonstrated LLMs ability to automate other tasks involved with a JCA submission such as the mass extraction of PICOs from clinical abstracts[11], screening literature and assessing risk-of-bias for SLRs[12].

References

1. Regulation (EU) 2021/2282 of the European Parliament and of the Council of 15 December 2021 on Health Technology Assessment and Amending Directive 2011/24/EU. [(accessed on 16 December 2023)]. Available online: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32021R2282
2. EUnetHTA 21, Practical Guideline D4.2 Scoping Process, Version 1.0, Sep 12, 2022 (Template Version Sep 30, 2021) [(accessed on 16 December 2023)]. Available online: https://www.eunethta.eu/wp-content/uploads/2022/09/EUnetHTA-21-D4.2-practical-guideline-on-scoping-process-v1.0.pdf
3. Reason, T., Benbow, E., Langham, J. et al. Artificial Intelligence to Automate Network Meta-Analyses: Four Case Studies to Evaluate the Potential Application of Large Language Models. PharmacoEconomics Open 8, 205–220 (2024). https://doi.org/10.1007/s41669-024-00476-9
4. MSR18 Improving the Performance of Generative AI to Achieve 100% Accuracy in Data Extraction. Klijn et al. Value in Health. Volume 27, Issue 6, S262-S263
5. PaddlePaddle/PaddleOCR. [Accessed May 2024] https://github.com/PaddlePaddle/PaddleOCR
6. Dimopoulos MA, Stewart AK, Masszi T, Špička I, Oriol A, Hájek R, Rosiñol L, Siegel D, Mihaylov GG, Goranova-Marinova V, Rajnics P, Suvorov A, Niesvizky R, Jakubowiak A, San-Miguel J, Ludwig H, Ro S, Aggarwal S, Moreau P, Palumbo A. Carfilzomib-lenalidomide-dexamethasone vs lenalidomide-dexamethasone in relapsed multiple myeloma by previous treatment. Blood Cancer J. 2017 Apr 21;7(4):e554. doi: 10.1038/bcj.2017.31. PMID: 28430175; PMCID: PMC5436074.
7. Badros A, Hyjek E, Ma N, Lesokhin A, Dogan A, Rapoport AP, Kocoglu M, Lederer E, Philip S, Milliron T, Dell C, Goloubeva O, Singh Z. Pembrolizumab, pomalidomide, and low-dose dexamethasone for relapsed/refractory multiple myeloma. Blood. 2017 Sep 7;130(10):1189-1197. doi: 10.1182/blood-2017-03-775122. Epub 2017 May 1. PMID: 28461396.
8. Baz RC, Martin TG 3rd, Lin HY, Zhao X, Shain KH, Cho HJ, Wolf JL, Mahindra A, Chari A, Sullivan DM, Nardelli LA, Lau K, Alsina M, Jagannath S. Randomized multicenter phase 2 study of pomalidomide, cyclophosphamide, and dexamethasone in relapsed refractory myeloma. Blood. 2016 May 26;127(21):2561-8. doi: 10.1182/blood-2015-11-682518. Epub 2016 Mar 1. PMID: 26932802
9. Lesokhin, A.M., Tomasson, M.H., Arnulf, B. *et al.* Elranatamab in relapsed or refractory multiple myeloma: phase 2 MagnetisMM-3 trial results. *Nat Med* **29**, 2259–2267 (2023). https://doi.org/10.1038/s41591-023-02528-9
10. Motzer RJ, Escudier B, McDermott DF, George S, Hammers HJ, Srinivas S, Tykodi SS, Sosman JA, Procopio G, Plimack ER, Castellano D, Choueiri TK, Gurney H, Donskov F, Bono P, Wagstaff J, Gauler TC, Ueda T, Tomita Y, Schutz FA, Kollmannsberger C, Larkin J, Ravaud A, Simon JS, Xu LA, Waxman IM, Sharma P; CheckMate 025 Investigators. Nivolumab versus Everolimus in Advanced Renal-Cell Carcinoma. N Engl J Med. 2015 Nov 5;373(19):1803-13. doi: 10.1056/NEJMoa1510665. Epub 2015 Sep 25. PMID: 26406148; PMCID: PMC5719487.
11. Reason, T., Langham, J. & Gimblett, A. Automated Mass Extraction of Over 680,000 PICOs from Clinical Study Abstracts Using Generative AI: A Proof-of-Concept Study. *Pharm Med*(2024). https://doi.org/10.1007/s40290-024-00539-6
12. MSR80 AI-Enabled Risk of Bias Assessment of RCTs in Systematic Reviews: A Case Study. Langham, J. et al. Value in Health, Volume26, Issue 12, S408

Presented at ISPOR Europe 2024; November 17-20, 2024; Barcelona, Spain

Email: yunchou.wu@estima-sci.com