Validating the use of Large Language Models to generate Individual Patient-Level data for use in Health Technology Assessment

MSR218

Emily Foreman,[1] Oliver Pople,[1] Bryony Langford,[1] Laura Sawyer[1]

[1]Symmetron Limited, London, England • Poster inquiries: eforeman@symmetron.net • www.symmetron.net • Presented at ISPOR EU 2025 Glasgow Annual Meeting

Background and Objectives

- The use and potential of Large Language Models (LLMs) continues to grow within health economics and outcome research.
- Individual Patient-Level data (IPD) are a valuable, but often difficult to obtain commodity in healthcare research. For this reason, it might be of interest to utilise LLMs to generate realistic IPD for use in research projects.
- The objective of this research was to explore the strengths and limitations of using LLMs to generate IPD.

Figure 1: LLM prompts to obtain IPD, using incrementally more information

(1) "Generate individual patient-level data (IPD) based on the following summary of patient characteristics."

(2) "Regenerate the IPD, this time accounting for a negative correlation between age and health utility (i.e. accounting for the idea that older patients are generally expected to have worse health utility than younger patients)."

(3) "Regenerate IPD again, this time adding correlation that you would expect between the clinical characteristic variables (type 1/2 diabetes, hypertension, kidney disease, cardiovascular disease and ICU admission), as well as their possible impact on health utility. Please provide an outline of the assumptions made."

Figure 2: Prompt for the LLM to create IPD based on its own assumptions

(4) "Generate individual patient-level data (IPD) based on the following summary of patient characteristics. Please make and apply sensible assumptions, for example any correlations that you would expect to see between the included variables, as well as restrictions on sensible ranges for each variable. Please outline the assumptions that have been made."

Abbreviations: IPD, individual patient-level data; LLM, large language model; ICU, intensive care unit.

Method

1. **Simulating IPD in R (v4.5.0)[1]:**

- A simulated dataset (SIM-1) was generated for 1000 patients, including health utility and the following list of demographic and clinical characteristics:
  - Sex, age, marital status, educational level, employment status, residence, type 1 diabetes, type 2 diabetes, hypertension, kidney disease, cardiovascular disease and ICU admission.
- Realistic data characteristics were incorporated into the data; for example, a negative correlation between the age and health utility variables.

2. **Creating LLM-generated IPD that approximate the SIM-1 data:**

- The SIM-1 data was summarised (using mean and SD for continuous variables and counts and percentage for binary variables), and these summary statistics were entered into an LLM (ChatGPT 5.0) to obtain an LLM-generated approximation of the SIM-1 data (LLM-1), following prompt (1) in **Figure 1.**[2]
- Further LLM-generated data (LLM-2 and LLM-3) were obtained by giving incrementally more information to the LLM via prompts (2) and (3) in **Figure 1,** to establish whether a better approximation to SIM-1 could be made.

3. **The LLM-generated datasets were compared to SIM-1 in R using the following methods:**

- Comparison of summary statistics via standardised mean difference.
- Plots used to visualise correlation of variables.
- Propensity score[3]:
  - Defined as the probability of dataset assignment conditional on all observed characteristics; used to determine how well each LLM dataset emulated SIM-1.
  - For sufficiently similar datasets, you would expect to not be able to predict the source of each row of data (i.e. whether it came from SIM-1 or from an LLM-generated dataset), from the characteristic values. This scenario would therefore yield a propensity score close to 0.5 for each row of data.
  - Using ROC (receiver operating characteristics) analysis, the area under the curve (AUC) was also calculated as an overall performance indicator; where AUC close to 0.5 suggests high similarity between datasets.
  - Propensity score was visualised via a mirrored histogram plot of propensity scores.
- Prompt (4) was entered into the LLM independently; requiring the LLM to make its own assumptions about the data based on the summary statistics and return an LLM-generated dataset using these assumptions (LLM-4). The aim of this approach was to assess the ability of an LLM to make clinically reasonable assumptions (e.g. about correlations between variables).

Table 1. Summary statistics for continuous variables in each dataset

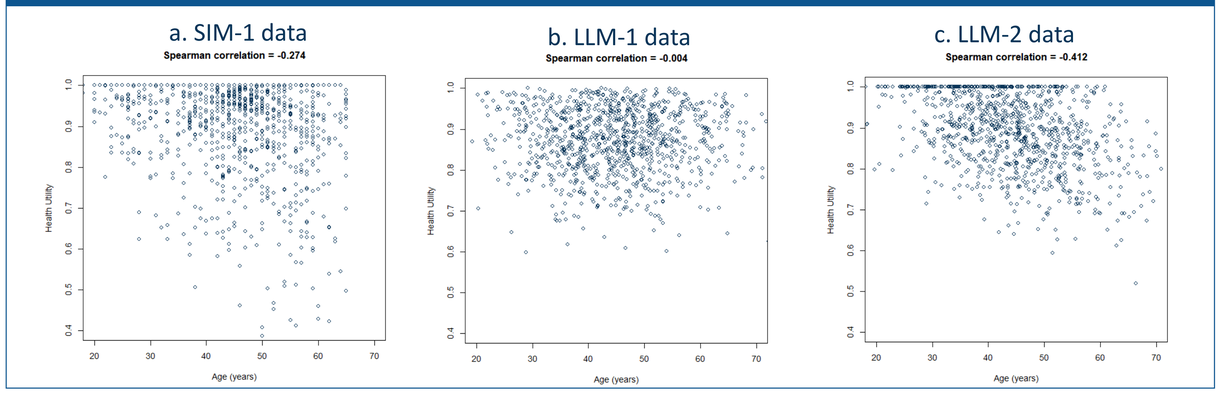
	SIM-1 data (N =1000)	LLM1 data (N=1000)		LLM2 data (N=1000)		LLM3 data (N=1000)		LLM4 data (N=1000)	
	Mean (SD)	**Mean (SD)**	**SMD**	**Mean (SD)**	**SMD**	**Mean (SD)**	**SMD**	**Mean (SD)**	**SMD**
Age	44.27 (10.95)	44.69 (10.37)	0.039	44.15 (10.25)	0.011	43.91 (10.34)	0.034	44.14 (10.53)	0.012
Utility Score	0.90 (0.12)	0.87 (0.08)	0.294	0.89 (0.09)	0.109	0.88 (0.09)	0.218	0.90 (0.10)	0.023

Abbreviations: SD, standard deviation; SMD, standardized mean difference.
SMD is calculated for each LLM-generated dataset versus the corresponding variable in the SIM-1 data.

Results

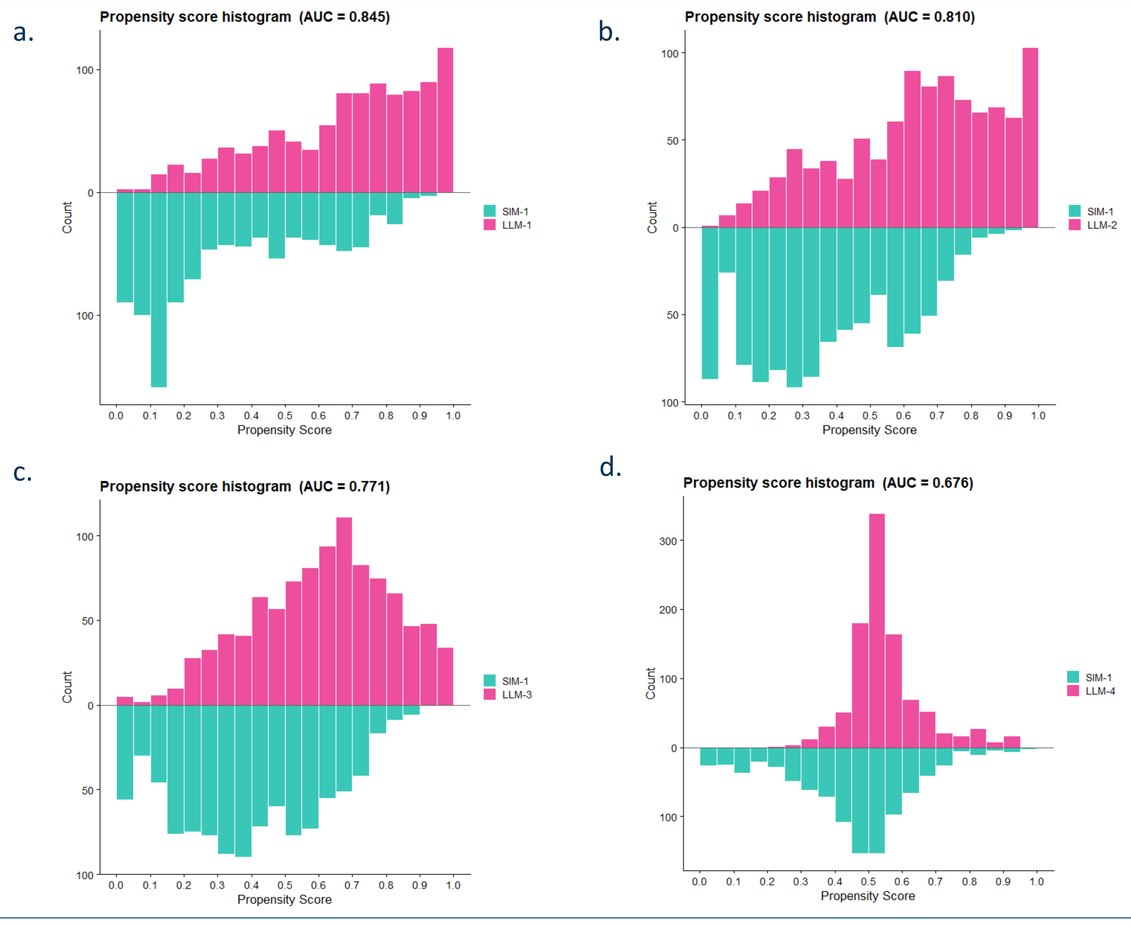
- For all prompts, the LLM was able to generate IPD with an identical distribution for categorical variables, but with minor deviations for continuous variables. This is seen in **Table 1** with instances of SMD >0.1, which can be a sign of covariate imbalance. Statistical software, such as the modgo package in R, is more efficient in this regard; able to match the distribution of all variables exactly.[3,4]
- Given prompt (1), the LLM would make minimal assumptions on its own:
  - The output from the LLM stated that age was assumed to be normally distributed on an appropriate range [18,80], and utility scores were restricted to a probability scale [0,1].
  - However, it did not specify whether assumptions were made about correlation between variables and, upon inspection, it was clear that it had assumed independence between all variables. For example, **Figure 3a** shows a negative correlation between age and health utility in the SIM-1 data which is not present in the LLM-1 data **(Figure 3b).**
- The LLM could account for dependence between variables once explicitly asked to do so, as done in prompt (2). **Figure 3c** shows a negative correlation between age and health utility, more closely resembling that in SIM-1/**Figure 3a.**

Figure 3: Scatterplots of Age versus Health Utility



- The propensity scores showed minor improvement in LLM performance between the LLM-1 and LLM-2 data, with AUC reduced from 0.845 to 0.810 (Figure 4a/b).
- Prompt (3) provided the LLM with some flexibility to make its own assumptions about the relationships of clinical variables in the data. This had a positive impact on performance, with the AUC reducing from 0.810 to 0.771 between prompts (2) and (3) **(Figure 4b/c).**
- Prompt (4) provided the LLM with full flexibility to make its own assumptions about what it would expect to see in data containing the included variables.
  - The prompt provided a summary of the key assumptions used to generate the IPD, which seemed sensible, and aligned with the data characteristics seen in LLM-4.
  - This approach produced the best approximation to the SIM-1 data, with an AUC of 0.676 and a propensity score histogram roughly centred around 0.5.

Figure 4: Histograms of propensity scores for each LLM-generated dataset versus the SIM-1 data



Abbreviations: AUC, area under the curve; LLM, large language model.

Conclusions

Table 2. Summary of the strengths and weaknesses of LLMs for IPD generation

Strengths	Weaknesses
✓ Able to approximate IPD where actual data are unavailable.	× Under normal circumstances, validating LLM-generated IPD against 'real' IPD, as done in this research, would not be feasible since 'real' IPD would not be available. Data produced by an LLM would therefore need to be sense checked before use (e.g. checking variable relationships) since, as demonstrated in this research, LLM output should not always be taken at face value.
✓ LLMs can reconstruct IPD distributions that are fairly consistent with the original, given summary statistics.	× Lack of reproducibility: LLM reconstruction processes can be opaque, posing a challenge for regulatory acceptance. Research also showed variability in LLM response to identical prompts when repeated.
✓ LLMs can accelerate the process of reconstruction, and do not rely on programming knowledge.	× Lack of transparency on the data that the LLM model has been trained on. May cause issues if model is trained on biased medical literature, for example, leading to the need to consult with clinicians to validate the assumptions of the LLM.
✓ Reduced cost compared to manual simulation, or gaining access to original data.	× Lacks statistical rigour of traditional IPD reconstruction methods, which can provide quantifiable uncertainty and distributional control.
✓ Can allow LLM to make clinical assumptions and simulate inter-variable dependencies based on learnings from literature, given a specific indication or population type.	

- This research demonstrated the ability of an LLM to generate IPD when given summary statistics but showed that the IPD will not always be constructed realistically (e.g. considering inter-variable dependencies), unless explicitly asked to do so.
- LLMs can make and apply clinical assumptions and variable correlations when asked to do so, providing the ability to reconstruct IPD without consulting clinical experts.
- The issue, however, remains how to validate LLM-generated IPD without access corresponding 'real' data and/or input from clinicians.
- Given the possible limitations of LLMs to generate IPD, analyses using LLM-generated IPD is unlikely to be accepted by healthcare regulators for decision-making and may therefore be limited to use in exploratory/sensitivity analyses until further research is conducted and formal guidance issued.

References: *(1) R Core Team (2025). _R: A Language and Environment for Statistical Computing_. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>. (2) OpenAI. (2025). ChatGPT (GPT-5) [Large language model]. OpenAI. https://chat.openai.com/ (3) Austin PC. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. Multivariate Behav Res. 2011 May;46(3):399-424. doi: 10.1080/00273171.2011.568786. Epub 2011 Jun 8. PMID: 21818162; PMCID: PMC3144483. (4) Koliopanos G, Ojeda F, Ziegler A. A Simple-to-Use R Package for Mimicking Study Data by Simulations. Methods Inf Med. 2023 Sep;62(3-04):119-129. doi: 10.1055/a-2048-7692. Epub 2023 Mar 7. PMID: 36882158; PMCID: PMC10462429.*

Acknowledgements: *Thanks to Corinne LeReun for additional advice and statistical support.*
Declaration of funding: *This project has been funded in full by Symmetron Limited.*