Automated Non-interventional Research Protocol Generation: Case Studies in Melanoma and Renal Cell Carcinoma

[RWD137]

Langham J[1], Benbow E[1], Reason T[1], Malcolm B[2], Gimblett A[1], Hill N.R.[3]
[1]Estima Scientific Ltd, London, UK, [2]Bristol Myers Squibb, Middlesex, London, UK, [3]Bristol Myers Squibb Company, Princeton, NJ, USA

Key Messages

What is already known on this topic

It is critical that non-interventional research protocols align with good practice guidelines to ensure the study design and conduct are clear, operationalizable and minimizes bias, to ensure results are valid and reliable. Protocol development is a costly and resource-intensive process.

What this study adds

The advancement of foundation models, including large language models like GPT-4, offers new opportunities for automating aspects of research.[1-4]

This study investigated the feasibility of protocol automation using GPT-4. We were able to automatically generate sections of a protocol comparable in quality to human-generated text and that adhered to established guidelines.

How this study might affect research, practice or policy

Our case study demonstrates that GPT-4 has the potential to quickly and efficiently assist in producing non-interventional research protocols, reducing human labor, and potentially enhancing efficiency and consistency across studies.

Background

- The development of non-interventional research (NIR) study protocols, vital for ensuring the integrity and replicability of studies used for regulatory decisions and the advancement of treatments, is traditionally a resource-intensive task.
- As the landscape of artificial intelligence (AI) evolves, large language models (LLMs) like GPT-4 offer the potential to automate traditionally human-intensive processes.
- With established guidelines promoting protocol standardization[5,6] there is now an opportunity to explore the practicality of applying LLMs to enhance protocol development efficiency in NIR studies.

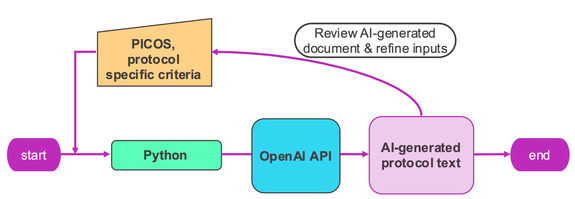
Objectives

- This study investigated the feasibility of NIR protocol automation by developing a method utilizing artificial intelligence (LLMs), to automate the development and the writing of protocols, to improve the efficiency of protocol development and the ability to conduct research.
- This pilot study aimed to automate the writing of the methods sections (the Study Design and Statistical Analysis plan) in four case studies and develop a generalizable method to further apply to a wider set of protocols

Methods

- We selected GPT-4 (Generative Pre-trained Transformer 4, developed by OpenAI)[5] for protocol generation for this study.
- A Python application programming interface (API) was used to send instruction “prompts” to GPT-4 to generate text for the Study Design and Statistical Analysis sections of NIR protocols.
- Prompts included detailed, protocol-specific instructions, including PICO-formulated research questions.
- Our approach used an iterative development process, aligning the AI-generated content with the quality and criteria of human-generated protocols, according to the standards set by STaRT-RWE template[6] and the HARmonized protocol.[7]
- **Figure 1** shows the development process for protocol generation.
- We tested the methods by generating four new (previously unseen by GPT-4) protocols.

Figure 1. The prompt development process



- Prompts for the protocols included the PICO information for NIR studies, specifically non-comparative retrospective cohort studies using electronic health records, in oncology (melanoma and renal cell carcinoma [RCC]), and descriptive outcomes (**Table 1**).
- Table 1 also shows the information we were asking GPT-4 to generate in this case study.
- The accuracy and completeness of GPT-4’s outputs were qualitatively assessed against the criteria outlined in the guidelines,[6,7] and against the original human-produced protocol content, focusing on the identification of critical points, and noting any omissions or inaccuracies.
- The two RCC protocols that had been auto-generated, alongside the original human-generated protocols, were also entered to BMS’ internal processes for study approval and the reviewer was blinded to the source of the text. (i.e., AI or human).

Table 1. Prompts and inputs, and the information generated by AI for each protocol section

Protocol Section	Prompt content and inputs to model	AI-generated information
Study Design	Specification of study design e.g., retrospective observational multi-site physician survey and medical chart review, example text about observational NIR studies.	Overview of Study Design Rationale for study design choice.
Population	Study Population (Inclusion Criteria, Exclusion Criteria), example text. Oncology (The study population included advanced melanoma [locally advanced or metastatic] patients or patients diagnosed with advanced renal cell cancer [aRCC]).	The population, defined in terms of persons, place, time period and selection criteria.
Objective 1	Describe the baseline demographic and clinical characteristics.	-
Objective 2	Characterize treatment patterns and sequence.	-
Objective 3	Clinical outcomes e.g., Estimate overall survival (OS) , Estimate progression free survival (PFS).	-
Objective 4	Healthcare Resource Utilization and Costs.	-
Data Sources	Data Source Data collection methods Data linkage Study timelines, recruitment period, start and end date, index date etc., example text.	Text generated using template text specific for each data source including data specification, geographical coverage, the source population from which the study population will be selected, reason for selection, strengths and limitations of data source, etc.
Statistical analysis plan	Specific aims and objectives, Predefined outcomes and covariates, example text.	Summary Methods. Analysis Plan for Objective(s).

Results

- Study design and statistical analysis sections for four NIR protocols were autogenerated, two in melanoma and two in RCC patients.
- We found that GPT-4 was able to generate sections of an NIR protocol in line with the pre-specified criteria, and that were comparable to human-produced text.
- GPT-4 was particularly good at producing high quality statistical analysis plans for each of the objectives included in this study.
- The protocols that were blind-reviewed were approved with minor comments, something that is rarely observed with human-produced text.
- Results are shown in Table 2.

Conclusions

- Detailed prompts were required to ensure GPT-4 was able to undertake this task. Further prompt refinement and fine-tuning with GPT-4 would increase the accuracy, particularly for the more complex study designs.
- Once the prompts and instructions had been developed for a particular study design, they were easily adapted for different clinical indications and different datasets, which makes methods generalizable with time-saving potential.
- This study demonstrates that given the right information, generative AI can help automate production of NIR protocols.
- This study focused on the main methods sections, i.e. Study design, and statistical analysis plan. Work is currently underway to extend the program to encompass the remaining sections (e.g., Introduction, sample size) and to include a more varied study design type, which will help demonstrate generalizability.

Table 2. Analysis of GPT-4 text against content guidelines and compared with human-generated protocols

	Strengths of GPT-4	Limitations
Study design and population	• Generated aspects of the process for study design and population inclusion and exclusion criteria. • Provided the rationale for the study design and data source that were missing in human generated text.	• Sometimes lacked specific details, particularly related to the study design such as dates of the study, the index date and the length of follow-up, unless it was specified in the template or PICOS text provided. • In one of the four protocols, GPT-4 did not correctly use the study period (Jan-2008 to Dec-2021) given in the input text and instead included a statement “*The study period spans from January 1, 2021, to December 31, 2022*” that had taken from one of the examples given. This was the only example of incorrect text in the GPT-4 generated text.
Data source	• GPT-4 provided more elaborate details on data collection settings, coding standards for diagnoses and procedures, and the rationale for the study design e.g., patient inclusion criteria, the geographical and temporal scope of the study, compared with the human generated protocol.	• Information required included specific details of how patients would be recruited e.g., how the 'analytic dataset of interest' will be created (defining index date, follow-up period, process of linkages between files), as well as the variables that will be extracted. As could be expected, GPT-4 was unable to provide this without guidance.
Statistical Analysis plan	• There was a high standard of content for the description of statistical methods, with GPT-4 following the overall guidelines and providing a clear methodology for analysis for each objective. • The GPT-4-generated text often contained more detail regarding the statistical techniques used than the human generated protocol.	• GPT-4 required detailed information about potential subgroup analysis, and time points and definition of outcomes. It did not try to invent this information. However, much of this information may be generalizable across studies measuring the same outcomes in the same populations.

Discussion

- Considerable effort and time to ensure accurate and generalisable prompt engineering were required for this study. This has now been achieved and was transferable to other NIR protocols with only minor adaptation.
- We believe that it is feasible to extend these methods to produce the remaining sections of the protocol, for example, writing the Introduction section by using Retrieval-Augmented Generation (RAG).
- Applying this method to a wider variety of NIR study designs will require further development and testing for accuracy.
- The LLM used was GPT-4 but the same prompts could be used with an alternative LLM.
- Protocol development is a resource-intensive process, Implementing AI into workflows may help focus the question, improve the quality and consistency of protocols produced and enable more time to be spent on conducting studies to answer the key clinical questions.

References

1. Reason T, et al. Artificial Intelligence to Automate Network Meta-Analyses *PharmacoEconomics - Open*. Published online February 10, 2024. doi:10.1007/s41669-024-00476-9
2. Langham J, et al . MSR80 AI-Enabled Risk of Bias Assessment of RCTs in Systematic Reviews *Value Health*. 2023;26(12):S408. doi:10.1016/j.jval.2023.09.2139
3. Reason T , et al .. MSR46 Breaking Through Limitations: Enhanced Systematic Literature Reviews With Large Language Models. Value Health. 2023;26(12):S402. doi:10.1016/j.jval.2023.09.2105
4. Reason, T , et al . Artificial Intelligence to Automate Health Economic Modelling *PharmacoEconomics Open* 8, 191–203 (2024). doi:10.1007/s41669-024-00477-8
5. OpenAI. GPT-4 is OpenAI’s most advanced system, producing safer and more useful responses. Accessed January 29, 2024. https://openai.com/gpt-4
6. Wang SV, et al. STaRT-RWE:*BMJ*. Published online January 12, 2021:m4856. doi:10.1136/bmj.m4856
7. Wang SV, et al. HARmonized Protocol Template to Enhance Reproducibility of hypothesis evaluating real-world evidence studies on treatment effects. *Pharmacoepidemiol Drug Saf*. 2023;32(1):44-55. doi:10.1002/pds.5507
8. The International Society for Pharmacoepidemiology (ISPE). Guidelines for Good Pharmacoepidemiology Practices (GPP). Accessed January 29, 2024. https://www.pharmacoepi.org/resources/policies/guidelines-08027/

Presented at ISPOR 2024; May 5-8, 2024; Atlanta, GA

Email: [julia.langham@estima-sci.com]