GPT-based Models for Data Analysis in Clinical Trials: A New Frontier in Research Methodologies?

Giulia Lorenzoni [1], Dario Gregori [1]

[1] Unit of Biostatistics, Epidemiology and Public Health, Department of Cardiac, Thoracic, Vascular Sciences, and Public Health, University of Padova, Padova, Italy

Corresponding author: dario.gregori@unipd.it

Disclosure list: No Relevant Disclosures.

BACKGROUND

Large language models such as GPT-4 offer significant potential for clinical research by improving efficiency, accuracy, and scalability (1). In the literature, we are beginning to see the first practical applications of this type of model in clinical research, with truly encouraging results. Applications related to the automatic extraction and classification of free texts, such as clinical diaries, are particularly noteworthy for the excellent results obtained (2).
The area of clinical trials (pharmacological and nonpharmacological) is certainly a context in which these kinds of models can offer many opportunities (3). It has been suggested that they could help in study design, patient selection, ensuring study compliance with current regulations, and, last but not least, data analysis.

OBJECTIVES

This study proposes to use the data analyst function of GPT-4 for the reanalysis of a clinical trial. The choice of the LLM used is due to the great popularity of OpenAI's GPT models and their relative ease of use, as GPT-4's data analyst function is available through a minimal subscription plan and operates via chat, making it easily accessible and usable even by clinical trial professionals who do not have specific artificial intelligence skills.

MATERIAL AND METHODS

The trial that underwent re-analysis was previously published (name not disclosed as per research agreement). Briefly, it was a multicenter, randomized, parallel-group trial involving patients undergoing cardiothoracic surgery. The trial compared two postoperative care models based on two primary endpoints: length of hospital stay and duration of drainage system usage, both measured in days.
The database provided to GPT-4 was completely anonymous and included only the three columns necessary for analyzing the primary endpoints. The database was formatted as an Excel file. Table 1 presents the prompts provided.

RESULTS

GPT-4 conducted a descriptive analysis of the two outcomes stratified by intervention type and then performed a t-test for inferential purposes (Table 1). The cross-check of the analyses confirmed the accuracy of the results. It provided also the interpretation of the results of the tests performed.
However, two main critical aspects emerged. First, GPT-4 did not consider the type of data distribution (normal or not). Second, it did not account for the issue of multiple primary endpoints, which involves the problem of test multiplicity requiring specific analytical approaches to be addressed (the main analyses of the trial employed a Benjamini-Hochberg correction for the problem of multiplicity of testing). Table 2 presents the answers to the request of clarifications about these two aspects of the analysis.

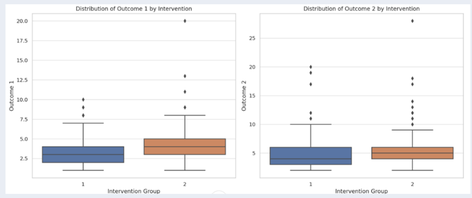
Table 2. Request for clarifications to GPT-4's data analyst function and assistant's answers.

USER	I have two questions: Why did you choose to perform the t-test? Why did you not account for the fact that we are working with two primary endpoints?
GPT-4	1. **Why did you choose to perform the t-test?** The t-test is a commonly used statistical test to compare the means of two groups. I chose to perform the t-test for your data for several reasons: Nature of the Data: Your data includes two distinct groups (interventions) and two continuous outcomes. The t-test is appropriate here to test whether the mean difference between these two groups is statistically significant for each outcome. Assumptions: Based on the initial inspection of the box plots, although the data appeared somewhat normally distributed, the t-test is robust to deviations from normality especially with larger sample sizes. 2. **Why did you not account for the fact that we are working with two primary endpoints?** This is an important point regarding multiple testing. When multiple hypotheses are tested simultaneously, the chance of a Type I error (false positive) increases. This phenomenon is known as the "multiple comparisons problem." To correct for this, adjustments like the Bonferroni correction or other methods (e.g., False Discovery Rate) can be applied.

REFERENCES

1. Xue VW, Lei P, Cho WC. The potential impact of ChatGPT in clinical and translational medicine. Clinical and Translational Medicine. 2023;13(3).
2. Huang J, Yang DM, Rong R, Nezafati K, Treager C, Chi Z, et al. A critical assessment of using ChatGPT for extracting structured data from clinical notes. npj Digital Medicine. 2024;7(1):106.
3. Ghim JL, Ahn S. Transforming clinical trials: the emerging roles of large language models. Translational and Clinical Pharmacology. 2023;31(3):131.

Table 1. Prompts provided to GPT-4's data analyst function and assistant's answers. ***Note that a short version of the answers has been reported in the table.***

USER	I am providing you with an Excel file containing data from a randomized controlled trial. Each row represents a subject who was enrolled and randomized in the study. This clinical trial compares the effects of two different strategies on two continuous primary endpoints. The database includes three columns: 'intervention' — indicates the intervention strategy to which each patient was randomized. 'outcome1' — one of the primary endpoints. 'outcome2' — the other primary endpoint. I request an analysis comparing the distribution of the two primary endpoints based on the intervention strategy.
GPT-4	Here are the findings from the analysis: **Boxplot Analysis**  Outcome 1 and Outcome 2 Distributions: The boxplots above illustrate how the outcomes are distributed across the two intervention groups. From the plots, we observe slight differences in the central tendency and spread between the groups for both outcomes. Summary Statistics **Outcome 1:** Intervention 1: Mean = 3.51, Standard Deviation = 1.73, Count = 204 Intervention 2: Mean = 4.06, Standard Deviation = 2.02, Count = 261 **Outcome 2:** Intervention 1: Mean = 4.87, Standard Deviation = 2.48, Count = 204 Intervention 2: Mean = 5.19, Standard Deviation = 2.64, Count = 260 These results suggest there may be differences in the outcomes based on the intervention strategy. To determine if these differences are statistically significant, I will conduct t-tests for each outcome between the two intervention groups, assuming unequal variances (Welch's t-test). Let's proceed with the t-tests. **T-Test Results** Outcome 1: Statistic: -3.12 P-value: 0.0019 This result indicates that there is a statistically significant difference between the two intervention groups for Outcome 1, with a very low p-value suggesting strong evidence against the null hypothesis (no difference). Outcome 2: Statistic: -1.34 P-value: 0.181 This suggests that, based on the available data, there isn't enough evidence to reject the null hypothesis that there is no difference in Outcome 2 between the intervention groups. **In summary:** Outcome 1 showed a statistically significant difference between the intervention groups. Outcome 2 did not show a statistically significant difference after accounting for missing data. If you have any further analyses in mind or need clarification on any part, feel free to ask!

CONCLUSIONS

Preliminary results confirm that GPT-4 is a useful assistant, rather than a substitute for humans, for researchers engaged in clinical trial analysis. The techniques applied provide correct results, but the choice of the appropriate approach certainly requires the support of experienced investigators.
It is noteworthy that the use of a system of this nature, without expert guidance, can lead to misleading results.