

Generative Artificial Intelligence in sample size estimation - challenges, pitfalls, and conclusions

Igor Goldfarb, Accenture Life Sciences R&D
Vikas Sharma, Catalyst Clinical Research

ABSTRACT

The last couple of years are characterized by storming expansion of Generative Artificial Intelligence (GenAI) and leading pharma companies plan to invest billions of dollars in development of GenAI in upcoming years.

The goal of this work is to increase awareness of investigators, scientists and statisticians about challenges and risks in application of GenAI to calculate the sample size for prospective clinical trial.

The authors assessed the performance, benefits, and risks of using GenAI for sample size calculations in clinical study design. One of the popular GenAI products - Chat Generative Pretrained Transformer (ChatGPT) – was chosen as a working tool. The sample size given in the completed study (the data were taken from ClinicalTrials.gov) was replicated using commercial software nQuery and ChatGPT. The authors found that ChatGPT failed to produce proper outcome (nQuery returned the correct results) while demonstrating a spectrum of errors. ChatGPT also did not meet reproducibility test – the tool returned various results in response to the same request asked multiple times.

The authors discuss the potential reasons for these observations and assume that the fundamental properties of the Large Language Models (LLM) lying in the core of GenAI may be responsible for these outcomes.

While the use of GenAI is very promising, there are many challenges and limitations in its application at its current stage. (e.g., bias, hallucinations, complexity, replicability, etc.). In general, scientists and statisticians should exercise caution in using GenAI. The companies planning the clinical study are recommended to hire experienced biostatisticians for sample size estimation.

INTRODUCTION

The last couple of years are characterized by storming expansion of Generative Artificial Intelligence (GenAI) in almost every area of human life and pharmaceutical industry is not an exclusion. Leading pharma companies plan to invest billions of dollars in development and application of GenAI in upcoming years.

One of the remarkable advancements in this realm is the advent of virtual chatbots and conversational agents. Thanks to their capacity to mimic natural human responses to text inputs, they have surged in popularity in thousands of professional applications and millions of regular computer users. Such a sharp rise is greatly attributed to the recent progress in the development of large language modules (LLM), that allowed researchers to significantly enhance their performance.

In our days there is a whole family of very popular GenAI-based tools, but one can surely say that ChatGPT is a top one in this unofficial rating. ChatGPT was developed and is marketed by Open AI and it uses deep learning techniques and advanced LLM to generate human-like textual responses in conversational manner. The tools based on GenAI technologies have found wide use in various domains including customer support, virtual assistants, content creation, finance, healthcare, etc. With the increasing adoption of AI-powered chat systems in real-world applications, there arises a crucial need to evaluate their reliability and performance. ChatGPT has been tested in a variety of situations and has

been able to achieve passing scores in many professional exams. To learn more about it and to check its potential usefulness in clinical trials domain we tried to replicate sample size calculations using ChatGPT.

Any person involved in pharmaceutical business study knows how time-consuming and expensive clinical trials are. Each study is carefully and thoroughly planned from its very early stage. The developed plan typically contains multiple items (e.g., main objectives of the trial, primary and secondary endpoints, optimal sample size, method of collecting the data, etc.) and represents a natural part of the Protocol of prospective study. Estimation of a sample size (accompanied by its scientific justification) - a number of subjects to be enrolled in the trial – symbolizes one of the key aspects of the plan. Sample size must be carefully evaluated to ensure that research time, personnel effort, and cost are not wasted. It is not uncommon for a study to fail to detect even large treatment effects due to insufficient sample size. There are many formulas and statistical tests available, which can make it difficult for clinical investigators to decide which method to use. Normally sample size is calculated by an experienced statistician who would use commercial or public software.

The authors tried to assess the performance and to identify the benefits and risks of using GenAI for sample size calculations in the prospective clinical trial. One of the popular GenAI products - Chat Generative Pretrained Transformer (ChatGPT) – was chosen as a working tool. At the first step the sample size given in the study (the data were taken from the government website ClinicalTrials.gov) was replicated using commercial software nQuery and one of free online calculators by entering appropriate information (e.g., type of test, study design, size effect, power, level of significance, etc.). Next, this information in the textual form (close to as it appeared in the Protocol) was provided to ChatGPT with request to calculate sample size. The authors found that ChatGPT failed to produce proper outcome while demonstrating a wide spectrum of errors, while nQuery and free calculators returned the correct results. Also, ChatGPT did not pass reproducibility test – the tool returned various results in response to the same request asked multiple times within the same chat or in another chat. The authors discuss the potential reasons of these observations and assume that the fundamental properties of the Large Language Models (LLM) lying in the core of GenAI may be responsible for these outcomes.

The goal of this work is to increase awareness of investigators, medical writers and statisticians who work on a prospective clinical trial about challenges and risks in application of GenAI to the calculation of the sample size for this trial. While the use of GenAI is very promising, at its current stage there are many challenges and limitations in its application (e.g., bias, hallucinations, complexity, replicability, etc.). The companies planning the study are recommended to use experienced statisticians for sample size estimation and scientists and statisticians should exercise caution in using GenAI.

BACKGROUND – BASIC STATISTICAL NOTIONS AND TERMS

The current Section contains a brief description of the statistical terms and notions and incomplete review of the published articles devoted to the issue of sample size estimation and its reporting in the study protocols.

TERMS AND NOTIONS

The calculation of the correct sample size is one of the first and most important steps in study design and development of its protocol. Before starting an actual presentation of the results let's remind – very shortly - basic statistical terms and notions one needs to understand to compute sample size for prospective study.

Study Hypotheses – Null H_0 and Alternative H_1

Consider an arbitrary pharmaceutical company testing its investigational product against control treatment

(already existing on the market) to evaluate if the investigational product is better in curing subject's suffering from some specific disease as compared to control treatment. Let's formulate two basic statistical hypotheses around this hypothetical research that are typically under consideration:

- Null Hypothesis (H_0): investigational product is equal to control treatment.
- Alternate Hypothesis (H_1): investigational product is better than control treatment.

Type I and II Errors – Significance and Power

Let's assume that based on the results of the study, it is concluded that investigational product is better than control treatment. In other words, one rejects the null hypothesis H_0 and accepts alternative one H_1 . Knowing (assumption just to illustrate the example) that the investigational product in reality is equivalent (=not better) to the control treatment (H_0), one is making an error here by rejecting H_0 . This is called a Type I error. Statistically it is defined as Type I error = [(Reject H_0)/(H_0 is true)]. The probability of Type I error is called a level of significance and is denoted as α .

To demonstrate an opposite case let's assume that based on the results of the study it is concluded that the investigational product is not better (=equivalent) than the control treatment. In other words, one rejects the alternative hypothesis H_1 and accepts the null one H_0 . Knowing (once again - assumption just to illustrate the example) that the investigational product in reality is better than the control treatment (H_1), one is making an error here by rejecting H_1 . This is called a Type II error. Statistically it is defined as Type II error = [(Reject H_1)/(H_1 is true)]. Probability of Type II error is denoted as β , but the term "power" (=1- β) is used more frequently than "type II error β ".

Normally these types of errors are presented in the form of table (Table 1) for better convenience.

	The null hypothesis is true: there is no true difference (H_0 is true, H_1 is false)	The null hypothesis is false: there is a true difference (H_0 is false, H_1 is true)
Study accepts the null hypothesis H_0	Correct conclusion: specificity (1- α)	Type II error (β)
Study rejects the null hypothesis H_0	Type I error (α)	Correct conclusion: power (1- β)

Table 1. Type I and II Errors

Sample Size Calculation – Required Items

Using the standard approach for determining the appropriate sample size, the following items need to be specified to make the computation feasible:

- A primary variable(s)
- The test statistic(s)
- The null hypothesis
- The alternative (working) hypothesis at the chosen dose(s)
- Effect size (Cohen's d)
- The probability of erroneously rejecting the null hypothesis (the type I error - α)
- The probability of erroneously failing to reject the null hypothesis (the type II error - β)
- Adjustments for treatment withdrawals and protocol violations

Sample Size Calculation – Other Terms and Notations

Using the standard approach for determining the appropriate sample size, the following items need to be specified to make the computation feasible:

- α - expected level of significance of the planned study
- β - the probability of erroneously failing to reject the null hypothesis (the type II error - β)
- $(1 - \beta)$ - expected power of the planned study
- (n, n_1, n_2) – number of subjects in each group (equal or not)
- Z - Z-score
- r - a ratio of sizes of two groups ($n_1 = rn_2$)
- d – expected difference between means of two groups
- $(\sigma, \sigma_1, \sigma_2)$ – standard deviations (common, per group)
- (μ, μ_1, μ_2) – means (common, per group)

Other Abbreviations and Terms

The following items abbreviations will be frequently used in the present paper:

- LLM – large language model
- LMM – large multi-modal model
- LHS/RHS – left/right hand side of equation
- WHO – World Health Organization
- GenAI – Generative Artificial Intelligence
- GAI – General Artificial Intelligence
- NLP – Natural Language Processing
- PI – principal investigator
- ChatGPT - Chat Generative Pre-Trained Transformer

LLM vs LMM

Let us very shortly draw a line to distinguish between these two large groups of models - LLM and LMM. According to Wikipedia LLM is a language model notable for its ability to achieve general-purpose language generation and other natural language processing tasks such as classification. LLMs acquire these abilities by learning statistical relationships from text documents during a computationally intensive self-supervised and semi-supervised training process. LLMs can be used for text generation, a form of generative AI, by taking an input text and repeatedly predicting the next token or word.

LMM, on the other hand, relies on a combination of various modalities of data, often arising in real-world applications (e.g., text with imaging data consisting of pixels and annotation tags). These modalities have fundamentally different statistical properties that require development of appropriate modelling strategies and algorithms.

LLM lies behind technologies that allow computer to read, process, understand, analyze, and reproduce the natural human language (papers, books, articles, documents, etc.). ChatGPT is probably the most well-known (at the moment) representative of GenAI having LLM as its basic models.

LMM, as it was mentioned above, is not restricted by the textual input/output. It can receive, process and output information in the form of the text, but is also able to take input and provide output in the form of photos, pictures, songs, videos, etc. It can take a drawing and describe in detail what is depicted there. Or, vice versa, it can produce a picture according to its textual description. Google Gemini is a good example of the technologies built on LMM.

Generative AI vs General AI

This Subsection includes a very concise explanation of a difference between two very similar terms – Generative Artificial Intelligence and General Artificial Intelligence.

Generative Artificial Intelligence, mostly known as GenAI, represents an umbrella name for multiple programs that leverage machine learning techniques and can create new products in the form of text, imaginary, video and even computer code based on the instructions given. GenAI tools rely on LLM models and are trained by feeding in billions of various documents. The backside is that GenAI is limited by the data it is fed and everything it produces is to some extent a derivative of the input data. As was mentioned above ChatGPT is probably the most well-known representative of the GenAI.

General Artificial Intelligence, sometimes referred to as Artificial General Intelligence (AGI), represents a term describing a much broader vision of machine's abilities – very close if not identical to human way of thinking. AGI presumes that the program possesses cognitive abilities akin to human intelligence, i.e. it can be expert in many specific things, it can learn on their own, it can act autonomously in any imaginable environment, etc. Unfortunately, the only place where we can encounter these systems today – in science fiction books.

LITERATURE OVERVIEW

It would be worthwhile to start the overview of existing publications on the subject of discussion from some kind of disclaimer. In the last couple of years, we have all witnessed an enormous increase in a number of publications devoted to research and investigation of various characteristics of GenAI in general and different tools using this technology in particular. The authors could read and review only a small fraction (relevant to the subject) of myriad of papers, preprints, blogs and similar sources of information. Therefore, the list of references should not be considered as complete in any way. Moreover, the authors will be very thankful for any referral to a publication relevant to the subject of this work.

The famous paper written by Alan Turing back in 1950 starts with the sentence “I propose to consider the question, ‘Can machines think?’”. In the next sentence he suggests defining the terms “machine” and “think”.

Interestingly, Turing (1950) mentioned that the idea of digital computers an old (!) one. He referred to Charles Babbage, Lucasian Professor of Mathematics at Cambridge from 1828 to 1839. Babbage (according to Turing) planned such a machine, called the Analytical Engine, but it was never completed. Although Babbage had all the essential ideas, his machine was not at that time such a very attractive prospect. Note here, that Because Babbage never published a detailed description of the Analytical Engine, Ada's translation (Lovelace Countess, 1842) of Menabrea's paper, with its lengthy explanatory notes, represents the most complete contemporary account in English of the intended design and operation of the first programmable digital computer.

Coming back to our days one has to note that popularity and massive use of GenAI technology broke every record when ChatGPT emerged and became available to an arbitrary user having access to Internet. Immediately millions of professionals, engineers, computer scientists and simply amateurs started to research and investigate its capabilities.

Both Borhi (2023) and Azaria (2022) mentioned that ChatGPT is limited in its capability to handle mathematical expressions and calculate even relatively simple tasks. They noted that like most LLM, ChatGPT struggles with tasks such as multiplying large numbers, finding roots, computing powers (especially with fractions), and adding or subtracting from irrational numbers.

Spitale *et al.*, (2023) tried to evaluate whether recruited individuals can distinguish disinformation from accurate information, structured in the form of tweets, and determine whether a tweet is organic or synthetic, i.e., whether it has been written by a Twitter user or by the GenAI model GPT-3. The results of their preregistered study (697 participants) showed that GPT-3 is a double-edged sword: in comparison with humans, it can produce accurate information that is easier to understand, but it can also produce more

compelling disinformation. The authors (Spitale *et al.*, 2023) also showed that humans cannot distinguish between tweets generated by GPT-3 and written by real Twitter users. In their paper they reflected on the dangers of GenAI for disinformation and on how information campaigns can be improved to benefit global health.

Chen *et al.* (2023) evaluated how the GenAI models perform over time. They conducted this evaluation in March 2023 and June 2023 versions of GPT-3.5 and GPT-4 on several diverse tasks: 1) math problems, 2) sensitive/dangerous questions, 3) opinion surveys, 4) multi-hop knowledge-intensive questions, 5) generating code, 6) US Medical License tests, and 7) visual reasoning. The authors found that the performance and behavior of both GPT-3.5 and GPT-4 can vary greatly over time. For example, GPT-4 (March 2023) was reasonable at identifying prime vs. composite numbers (84% accuracy) but GPT-4 (June 2023) was poor on these same questions (51% accuracy). This is partly explained by a drop in GPT-4's amenity to follow chain-of-thought prompting. Interestingly, GPT-3.5 was much better in June than in March in this task. GPT-4 became less willing to answer sensitive questions and opinion survey questions in June than in March. GPT-4 performed better at multi-hop questions in June than in March, while GPT-3.5's performance dropped on this task. Both GPT-4 and GPT-3.5 had more formatting mistakes in code generation in June than in March. The authors provide evidence that GPT-4's ability to follow user instructions has decreased over time, which is one common factor behind the many behavior drifts. Overall, their findings show that the behavior of the "same" LLM service can change substantially in a relatively short amount of time, highlighting the need for continuous monitoring of LLMs.

Azaria *et al.* (2023) investigated the capabilities of ChatGPT as an automated assistant in diverse domains, including scientific writing, mathematics, education, programming, and healthcare. The authors explored the potential of ChatGPT to enhance productivity, streamline problem-solving processes, and improve writing style. Furthermore, they highlighted the potential risks associated with excessive reliance on ChatGPT in these fields. These limitations encompass factors like incorrect and fictitious responses, inaccuracies in code, limited logical reasoning abilities, overconfidence, and critical ethical concerns of copyrights and privacy violation. The authors outlined areas and objectives where ChatGPT proves beneficial, applications where it should be used judiciously, and scenarios where its reliability may be limited. In light of observed limitations and given that the tool's fundamental errors may pose a special challenge for non-experts, ChatGPT should be used with a strategic methodology. By drawing from comprehensive experimental studies, they offered methods and flow charts for effectively using ChatGPT. Their recommendations emphasized iterative interaction with ChatGPT and independent verification of its outputs. Considering the importance of utilizing ChatGPT judiciously and with expertise, they recommend its usage for experts who are well-versed in their respective domains.

Methani *et al.* (2023) investigated the accuracy of CHatGPT in calculating sample size for sport-sciences and sports-medicine research studies. They conducted an analysis on 4 published papers (i.e., examples 1-4) encompassing various study designs and approaches for calculating sample size in 3 sport-science and -medicine journals, including 3 randomized controlled trials and 1 survey paper. The authors provided ChatGPT with all necessary data such as mean, percentage SD, normal deviates ($Z_{\alpha/2}$ and $Z_{1-\beta}$), and study design. Prompting from 1 example has subsequently been reused to gain insights into the reproducibility of the ChatGPT response. According to the authors ChatGPT correctly calculated the sample size for 1 randomized controlled trial but failed in the remaining 3 examples, including the incorrect identification of the formula in one example of a survey paper. After interaction with ChatGPT, the correct sample size was obtained for the survey paper. Intriguingly, when the prompt from Example 3 was reused, ChatGPT provided a completely different sample size than its initial response. Methani *et al.* (2023) concluded that while the use of artificial-intelligence tools holds great promise, it should be noted that it might lead to errors and inconsistencies in sample-size calculations even when the tool is fed with the necessary correct information. As artificial-intelligence technology continues to advance and learn from human feedback, there is hope for improvement in sample-size calculation and other research tasks. However, it is important for scientists to exercise caution in utilizing these tools.

Ignjatovic and Stevanovic (2023) investigated the performance of ChatGPT (GPT-3.5 and GPT-4) as a study tool in solving biostatistical problems and tried to identify any potential drawbacks that might arise from using ChatGPT in medical education, particularly in solving practical biostatistical problems. ChatGPT was tested to evaluate its ability to solve biostatistical problems from the Handbook of Medical Statistics by

Peacock and Peacock in this descriptive study. Tables from the problems were transformed into textual questions. Ten biostatistical problems were randomly chosen and used as text-based input for conversation with ChatGPT (versions 3.5 and 4). Authors found that GPT-3.5 solved 5 practical problems in the first attempt, related to categorical data, cross-sectional study, measuring reliability, probability properties, and the t-test. GPT-3.5 failed to provide correct answers regarding analysis of variance, the chi-square test, and sample size within 3 attempts. GPT-4 also solved a task related to the confidence interval in the first attempt and solved all questions within 3 attempts, with precise guidance and monitoring. They concluded that both versions of ChatGPT performance in 10 biostatistical problems revealed that GPT-3.5 and 4's performance was below average, with correct response rates of 5 and 6 out of 10 on the first attempt. GPT-4 succeeded in providing all correct answers within 3 attempts. These findings indicate that students must be aware that this tool, even when providing and calculating different statistical analyses, can be wrong, and they should be aware of ChatGPT's limitations and be careful when incorporating this model into medical education.

Yanagita *et al.* (2023) evaluate whether ChatGPT can provide accurate diagnoses and medical knowledge for Japanese input. They used questions (292 text questions out of total 400) from the National Medical Licensing Examination (NMLE) in Japan, administered by the Japanese Ministry of Health, Labour and Welfare in 2022. The authors found that the correct response rate for GPT-4 was 81.5% (237/292), which was significantly higher than the rate for GPT-3.5, 42.8% (125/292). Moreover, GPT-4 surpassed the passing standard (>72%) for the NMLE, indicating its potential as a diagnostic and therapeutic decision aid for physicians. Yanagita *et al.* (2023) concluded that GPT-4 reached the passing standard for the NMLE in Japan, entered in Japanese, although it is limited to written questions. As the accelerated progress in the past few months has shown, the performance of the AI will improve as the large language model continues to learn more, and it may well become a decision support system for medical professionals by providing more accurate information.

Finally, let us cite the developers of ChatGPT (Open AI, 2023). According to one of the texts placed on the system card for LMM (GPT-4): "We found that GPT-4-early and GPT-4-launch exhibit many of the same limitations as earlier language models, such as producing biased and unreliable content".

One can easily conclude from the incomplete literature review above that the issue of reliability of GenAI tools in the field of exact sciences (sample size estimation using ChatGPT represents a small stone in a mosaic) is very hot topic and a subject for discussions on different forums, in various formats and within unparallel professional communities.

CHATGPT ESTIMATES SAMPLE SIZE

This Section contains a number of examples demonstrating how a popular GenAI tool estimates sample size in various situations (when different tests and approaches are supposed to be used) and how inconsistent ChatGPT's responses are.

T-TEST (DIFFERENCE IN MEANS OF TWO GROUPS)

Let's begin our analysis from an application of the classical t-test (difference in means for two groups) for estimation of the sample size.

Consider the study "A Randomized Controlled Trial Comparing the Effect of Topical Dorzolamide-Timolol versus Placebo Combined with Intravitreal Anti-Vascular Endothelial Growth Factor (VEGF) Injections in Patients with Neovascular Age-Related Macular Degeneration Who Are Incomplete Anti-VEGF Responders" (ClinicalTrials.gov Identifier: NCT03034772). Its design represents a standard two groups parallel design. The primary efficacy endpoint is a change in mean central subfield thickness (CST) on spectral domain optical coherence tomography (SD-OCT) from baseline to the final visit. The study

protocol contains following statement describing the assumptions regarding the sample size calculation: “Assuming a power of 0.9 and an alpha of 0.05, approximately 44 patients will be required to demonstrate a CST improvement of 400 to 350 microns with a standard deviation of 50 microns.” In other words, according to the Protocol the minimal number of subjects required to reach the announced goal is 22 per each of 2 groups, i.e. 44 subjects in total.



You

Estimate sample size assuming a power of 0.9 and an alpha of 0.05 to demonstrate a CST improvement of 400 to 350 microns with a standard deviation of 50 microns.



ChatGPT

To estimate the sample size for comparing means between two groups, you can use the formula for sample size in a two-sample t-test. The formula is given by:

$$n = \left(\frac{Z_{\alpha/2} + Z_{\beta}}{\text{Effect Size}} \right)^2 \times \left(\frac{\text{Common Standard Deviation}^2}{n_1 + n_2} \right)$$

Where:

- $Z_{\alpha/2}$ is the Z-score corresponding to the desired level of significance (α).
- Z_{β} is the Z-score corresponding to the desired power ($1 - \beta$).
- Effect Size is the difference in means divided by the common standard deviation.
- n_1 and n_2 are the sample sizes in each group.

For your case:

- Significance level (α) = 0.05
- Power ($1 - \beta$) = 0.9
- Effect Size = $\frac{\text{Difference in means}}{\text{Standard Deviation}}$



Message ChatGPT...



ChatGPT can make mistakes. Consider checking important information.

Figure 1. ChatGPT response (part 1) to the request to estimate sample size for the prospective study using the assumptions identical to ones in the Protocol NCT03034772.

An elementary calculation using one of many existing free web-based statistical calculators (e.g., <https://www.statskingdom.com/>) provides one with the same outcome – 44 subjects in total (22 per group). Commercial (licensed) software nQuery gives very close value – 23 subjects per group. Note here, that Julious (2009) displays result of 23 subjects per group for in a parallel group study for 90% Power, two-sided type I Error of 5% and effect size 1 (Julious 2009, Table 3.2).

Let's now consider ChatGPT response to authors' request to calculate the sample size for hypothetical study using the same conditions that were formulated in the Protocol (Figure 1). The formula displayed by GenAI hits the eye immediately – both sides of the expression contain an unknown number of the subjects (per group) for the hypothetical study (the left side (LHS) of the equation contains variable n while the right side (RHS) of the same equation contains variables n_1 and n_2). In addition to the evident

fact that the formula is presented in the way when the dependent variable (unknown n , that authors are trying to estimate) appears in both LHS and RHS of the expression, the presentation of the dependent variable is different – notion n (LHS) assumes that the sizes of the two parallel groups are equal to each other, notions n_1 and n_2 (RHS) assumes that two groups of various sizes are under discussion. Therefore, the expression presented by ChatGPT causes serious doubts in its correctness even to an arbitrary researcher, not mentioning experienced statistician. The correct version of the formula (Cohen 1988, Julious 2009) reads as follows (two-sided t-test)

$$n_1 = \frac{(1+r) \left(Z_{1-\beta} + Z_{1-\frac{\alpha}{2}} \right)^2}{rd^2} \sigma^2 \quad (1)$$

Note also, that notations in the ChatGPT response are slightly different from what is normally used in the statistical literature (of course it is a matter of convention) when an estimation of the sample size is considered ($Z_{\alpha/2}$ vs $Z_{1-\alpha/2}$ and Z_β vs $Z_{1-\beta}$), while explanation of these notations is correct.

- Difference in means = 400 - 350 = 50 microns
- Standard Deviation = 50 microns

Let's calculate the Z-scores:

$$Z_{\alpha/2} = Z_{0.025}$$

$$Z_{0.025} \approx 1.96 \text{ (from standard normal distribution tables)}$$

$$Z_\beta = Z_{0.10}$$

$$Z_{0.10} \approx 1.28 \text{ (from standard normal distribution tables)}$$

Now, plug these values into the formula:

$$n = \left(\frac{1.96+1.28}{50/50} \right)^2 \times \left(\frac{50^2}{n_1+n_2} \right)$$

$$n = (3.24)^2 \times \left(\frac{2500}{n_1+n_2} \right)$$

Assuming an equal sample size in both groups ($n_1 = n_2$), you can simplify this:

$$n = (3.24)^2 \times \left(\frac{2500}{2n} \right)$$

Now, solve for n :

$$n = \frac{(3.24)^2 \times 2500}{2 \times n}$$

$$n = \frac{10.4976 \times 2500}{2}$$

$$n \approx 13122$$

So, you would need approximately 13122 samples in each group to achieve a power of 0.9 with a significance level of 0.05 and detect a difference of 50 microns with a standard deviation of 50 microns.

Figure 2. ChatGPT response (part 2) to the request to estimate sample size for the prospective study using the assumptions identical to ones in the Protocol NCT03034772.

Now let's ignore for some time the obvious errors in the formula itself and move to the next step, where GenAI demonstrates the actual calculation of the expected sample size (Figure 2). Specifically, consider three lines after the words "Now, solve for n ". One can easily see that there is an essential difference between the first two lines – the unknown variable n disappeared from RHS of the expression, but it was not transferred to its LHS as one could expected – n in the LHS still remains to be in the first power. It looks as if ChatGPT simply removed the unknown variable from RHS and continued its calculations as usual. The final outcome is shocking – according to GenAI one needs to recruit more than 13 thousand (!) of subjects for hypothetical study, which is three orders of magnitude higher than the number of subjects (44) enrolled in the actual study NCT03034772. Remind, different sources mentioned above provided the authors with 22/23 subjects per group.

To analyze the level of reproducibility of ChatGPT's responses let's open a new window and run the exactly the same question about hypothetical study we investigate. From the first glance it becomes obvious that the two results of two runs of ChatGPT in response to the same query are totally different. First of all, the displayed analytical expression for the sample size calculation looks similar to the formula from the textbook. Second, the final result (21 subjects per group) lies very close to the correct one (and is very different from 13000 from previous run).

Let's analyze the obtained value more closely. According to ChatGPT (Figure 3) $Z_{\alpha/2}=1.96$ (let's leave aside the difference between $Z_{\alpha/2}$ and Z_{α} - it can be investigated separately, inconsistency in displaying the key variables is not the primary issue here). It is well known that the value 1.96 approximates a more exact value that can be displayed as 1.959964 (six decimal digits are enough for the declared purpose). Similar, value 1.28 was obtained after rounding more exact number 1.281552. Once the calculation is performed (according to the same formula presented in Figure 3) using the exact values (not rounded) the final result reads as 21.01 and according to the general approach this value should be rounded up, giving 22 subjects per group (correct result). This example proves the very well-established rule that any rounding must be applied at the very final stage of any numeric calculation. Deviation from this approach leads to inevitable errors like in our example (compare Figure 2 vs Figure 3). In the considered case the difference is not large, but it is essential to understand the root cause of its presence.

Additional runs of ChatGPT with demands to respond to exactly the same query provides with other various responses that allows the authors to conclude that the results of sample size calculation (based on the standard t-test) for the study NCT03034772 cannot be reproduced with the help of this version of ChatGPT.



You

Estimate sample size assuming a power of 0.9 and an alpha of 0.05 to demonstrate a CST improvement of 400 to 350 microns with a standard deviation of 50 microns.



ChatGPT

To estimate the sample size required for a study comparing central corneal thickness (CST) before and after an intervention with a power of 0.9, significance level (alpha) of 0.05, and assuming a true improvement of 400 to 350 microns with a standard deviation of 50 microns, you can use a formula for the sample size in a two-sample t-test.

The formula for sample size in a two-sample t-test is given by:

$$n = \frac{2 \cdot (\sigma^2) \cdot (Z_\alpha + Z_\beta)^2}{(\text{mean difference})^2}$$

where:

- σ is the standard deviation
- Z_α is the Z-score corresponding to the chosen alpha level (0.05 for a 95% confidence interval is approximately 1.96)
- Z_β is the Z-score corresponding to the chosen power level (0.9 for a power of 0.9 is approximately 1.28)
- (mean difference) is the expected difference in means



Given your values:

- $\sigma = 50$ microns
- $Z_\alpha = 1.96$
- $Z_\beta = 1.28$
- (mean difference) = $400 - 350 = 50$ microns

Substitute these values into the formula:

$$n = \frac{2 \cdot (50^2) \cdot (1.96 + 1.28)^2}{50^2}$$

Calculate the result:

$$n = \frac{2 \cdot 2500 \cdot (3.24)^2}{2500}$$

$$n = \frac{2 \cdot 2500 \cdot 10.4976}{2500}$$

$$n = \frac{52488}{2500}$$

$$n \approx 20.9952$$

Therefore, you would need approximately 21 samples per group to detect a mean difference of 50 microns with a power of 0.9 and a significance level of 0.05, assuming a standard deviation of 50 microns. Keep in mind that you might want to round up to the nearest whole number since you



Figure 3. ChatGPT response (part 3) to the request to estimate sample size for the prospective study using the assumptions identical to ones in the Protocol NCT03034772.

CHI-SQUARE TEST (TWO SAMPLE CHI-SQUARE TEST)

Let's now change the main statistical test used in the corresponding study from t-test to chi-square test (two samples) and continue our analysis for estimation of the sample size.

Consider the study devoted to Bovine viral diarrhea virus in beef heifers in commercial herds in Australia. The study itself and its results are presented in the paper written by Morton *et al.* (2013). The main purpose of the study was to describe management group (mob)-level seroprevalences and incidences of seroconversion to bovine viral diarrhea virus (BVDV), and to determine the efficacy of a vaccine against BVDV, in beef heifers in commercial herds in Australia.

The study was performed as a standard double-blind vaccine efficacy trial. The paper contains the following statement describing the assumptions regarding the sample size calculation: *"We planned to enroll 2500 heifers (1250 each of Pestigard®- and placebo-treated heifers). Statistical power calculations indicated that the trial would have had power of 0.72, 0.73 and 0.91 for detecting a significant difference at the 0.05 level (two-sided) in proportions of heifers that were pregnant by 6 weeks between Pestigard® and placebo treated heifers, if the proportions in the placebo-treated heifers were 40%, 60% and 80%, respectively, and if Pestigard® treatment increases these proportions by 0.05 (i.e. 5 percentage points). If the increase is 0.10 (i.e. 10 percentage points), statistical power would have been 1.00."*

As we did in the previous example (t-test) before asking ChatGPT to replicate this outcome, let's check the authors' results using independent calculation. Free web-based statistical calculators (e.g., <https://www.statskingdom.com/>) provides one with the same outcome – 1250 subjects in per group (2500 in total). Commercial (licensed) software nQuery gives exactly the same value - 1250 subjects per group. Therefore, one can be sure that this value is the correct one.

Response of ChatGPT to authors' request to calculate the sample size for hypothetical study using the same conditions (Morton *et al.*, 2013) is displayed in the Figure 4. Similar to the previous example the formula displayed by GenAI does not look trustable – it is evident that the value of β has no role in it. For every experienced statistician it is a red flag – sample size naturally depends on the power (in other language – on the probability of erroneously failing to reject the null hypothesis). Without any further investigation one can conclude that the expression presented by ChatGPT is incorrect one.

The next step is verification of the conducted calculation (ignore at this stage the extent of the original formula's relevance to the hypothetical study). The first of the two lines marked by red (the Figure 4) reads as

$$n_1 = n_2 = \frac{2 * 3.8416 * 0.425 * 0.575}{0.0025} \quad (2)$$

And this is a result of correct transformation of the previous line. But the multiplication $3.8416 * 0.425 * 0.575$ does not result in 2.20294 as ChatGPT displays (second line in the red oval, the Figure 4). The right outcome is 0.938791, that in turn gives us the final value of 751.0328. Therefore, if ChatGPT would perform all the calculations in the appropriate way it would provide the authors with the outcome of 752 subjects per group. The actual output obtained by ChatGPT is 1762. No need to say that neither of the two are correct.

Reproducibility of ChatGPT's responses in this case can be verified in a way similar to one used for t-test. Let's open a new window and run the exactly the same question about hypothetical study we investigate (Figure 5). It is evident that the results of these two runs of ChatGPT in response to the same query are totally different. First, the displayed analytical expression for the sample size calculation looks similar to the formula from the textbook (Equation 1). Second, one needs to pay attention that the final value of calculations 1233.38 is explicitly rounded up to 1234 by ChatGPT, that is incorrect mathematically, but makes perfect sense from the statistical point of view (number of subjects must be a whole number). But the next, third step is totally unexpected - ChatGPT applies the same argument (number of subjects has to be a whole number) another time and changes the final outcome from 1234 to 1235 (orange oval in Figure 5) and this time it is totally inappropriate. This result (1235 subjects per group according to ChatGPT) lies not far from the correct one (1250), but not close enough. The last update does not change

much from the statistical point of view (1234 and 1235 subjects per group will generate almost negligible difference in outcome), but it witnesses that ChatGPT “does not understand” what it is doing.

$$n = \left(Z_{1-\beta} + Z_{1-\frac{\alpha}{2}} \right)^2 \frac{(p_1(1-p_1) + p_2(1-p_2))}{(p_1 - p_2)} \quad (3)$$

Regarding elementary mathematical operations one can conclude that they were performed much better this time, but still not 100% correctly. The final value (assuming all the values as they displayed in red oval, Figure 5) is expected to be 1236.7560. It is very close to what ChatGPT depicts (1233.38), but still inaccurate. Moreover, a value of Z_β is not obtained correctly – for the expected power 71.54% one can expect $Z_\beta = 0.57$. Using this value in the calculation according to the expressions one obtains 1248 subjects per group – much closer to the actual value of 1250 subjects as was used in the study (Morton *et al.*, 2013).

No need to say that the subsequent runs of ChatGPT with the same query brought us other versions of its responses.

VI

compute sample size using two sample chi square test for equal proportion assuming a power of 71.54%, significance level of 0.05, group 1 proportion is 0.40, group 2 proportion is 0.45.



To compute the sample size for a two-sample chi-square test with equal proportions, we'll use the formula:

$$n1 = n2 = \frac{2 \cdot Z_{1-\frac{\alpha}{2}}^2 \cdot p \cdot (1-p)}{\text{Effect Size}^2}$$

where:

- $Z_{1-\frac{\alpha}{2}}$ is the critical value of the standard normal distribution corresponding to the significance level ($\alpha/2$).
- p is the common proportion in both groups ($p1 = p2 = p$).
- Effect Size is the practical difference or effect size you want to detect between the groups.

Given the values:

- Power of the test ($1 - \beta$) = 71.54% = 0.7154
- Significance level (α) = 0.05
- Proportion in both groups ($p1 = p2 = p$) = $(0.40 + 0.45) / 2 = 0.425$ (assuming equal

Regenerate

Default (GPT-3.5)

- Proportion in both groups ($p1 = p2 = p$) = $(0.40 + 0.45) / 2 = 0.425$ (assuming equal proportions).
- Effect Size = $|p1 - p2| = |0.40 - 0.45| = 0.05$

Next, find the critical value $Z_{1-\frac{\alpha}{2}}$ for a significance level of 0.05. $Z_{1-\frac{\alpha}{2}} = Z_{0.975} \approx 1.96$ (rounded to two decimal places).

Now, we can plug these values into the formula:

$$n1 = n2 = \frac{2 \cdot 1.96^2 \cdot 0.425 \cdot (1-0.425)}{0.05^2}$$

$$n1 = n2 = \frac{2 \cdot 3.8416 \cdot 0.425 \cdot 0.575}{0.0025}$$

$$n1 = n2 = \frac{2 \cdot 2.20294}{0.0025}$$

$$n1 = n2 \approx 1761.24$$

Since we cannot have a fractional sample size, we should round up to the nearest whole number. Therefore, the required sample size for each group is approximately 1762.

Figure 4. ChatGPT response (part 1) to the request to estimate sample size for the prospective study using the assumptions presented in the paper by Morton *et al.* (2013).



Anonymous

compute sample size for two sample chi square test assuming a power of 71.54%, significance level of 0.05, group 1 proportion is 0.40, group 2 proportion is 0.45.



ChatGPT

To compute the sample size for a two-sample chi-square test, you can use the formula for sample size calculation in the context of hypothesis testing for proportions. The formula is as follows:

$$n = \frac{(Z_{\alpha/2} + Z_{\beta})^2 \cdot (p_1(1-p_1) + p_2(1-p_2))}{(p_1 - p_2)^2}$$

Where:

- n is the required sample size for each group.
- $Z_{\alpha/2}$ is the critical value for the desired significance level (α). For a significance level of 0.05 (two-tailed), $Z_{\alpha/2}$ is the critical value for 0.025 in each tail, which is approximately 1.96.
- Z_{β} is the critical value for the desired power ($1 - \beta$). Since you mentioned a power of 71.54%, Z_{β} is the critical value for the 28.46% quantile, which is approximately 0.5584 (you can find this value from a standard normal distribution table or calculator).
- p_1 is the proportion in group 1, which is 0.40.
- p_2 is the proportion in group 2, which is 0.45.
- $(p_1 - p_2)^2$ is the squared difference in proportions between the two groups.

Now, plug these values into the formula:

$$n = \frac{(1.96 + 0.5584)^2 \cdot (0.40(1-0.40) + 0.45(1-0.45))}{(0.40 - 0.45)^2}$$

Calculate the sample size:

$$n = \frac{(2.5184)^2 \cdot (0.24 + 0.2475)}{(-0.05)^2}$$

$$n = \frac{6.34201216 \cdot 0.4875}{0.0025}$$

$$n \approx 1233.38$$

You will need approximately 1234 observations in each group to achieve a power of 71.54% with a significance level of 0.05, assuming group 1 proportion is 0.40 and group 2 proportion is 0.45 in a two-sample chi-square test. Since sample sizes must be whole numbers, you may round up to 1235 if you

Figure 5. ChatGPT response (part 2) to the request to estimate sample size for the prospective study using the assumptions presented in the paper by Morton *et al.* (2013).

SIMON TWO-STAGE PHASE II STUDY (OPTIMAL DESIGN)

Consider now the study “A Phase I/II study of TRC105 in combination with Sorafenib in Hepatocellular Carcinoma (HCC)” (ClinicalTrials.gov Identifier: NCT01306058). This trial was performed in two stages – one was Phase I study, another one – Phase II trial. Our attention will be concentrated on the Phase II part. It was a standard Phase II trial conducted in accordance with Simon two-stage optimal design. Previous studies established sorafenib as a standard consideration in hepatocellular carcinoma (HCC), which is the fifth most common malignancy worldwide with a median survival of 6-9 months. The primary goal of Phase 2 part of the trial was to estimate response rate according to RECIST criteria for the combination of TRC105 with sorafenib in HCC.

The sample size and interim stopping rule were determined using a Simon optimal two-stage design (Simon, 1989). A median 5.5-month time to progression, translating into a 6 month progression free survival rate of 47%, would be considered not promising ($p_0=0.47$), while a 67% 6-month PFS rate (corresponding to a median time to progression of 10.5 months) would be considered promising ($p_1=0.67$). It was assumed that $\alpha=0.10$ and $\beta=0.10$.

According to the Protocol 21 patients were supposed to be accrued to the study in the first stage. If 10 or fewer patients demonstrate lack of progression on imaging after 6 months (three scans) on-study, the accrual to the study will be terminated early and the trial declared to have a negative result. A pause in the accrual may be necessary to ensure that enrollment may continue to the second stage. If 11 or more patients are progression-free at 6 months, enrollment will be extended to accrue a total of 45 patients.

As we did in the previous examples before asking ChatGPT to replicate this outcome let's check the values presented in the Protocol using independent calculation. Free web-based statistical calculators (e.g., <http://cancer.unc.edu/biostatistics/program/ivanova/SimonsTwoStageDesign>) provides one with the same outcome – 21 patients in the first stage, 10/11 is a critical value to continue/terminate the trial, 45 patients in total including the second stage. Commercial (licensed) software nQuery gives exactly the same values – 21/45 for the first/total stage, 10/11 is a criterion. Therefore, one can be sure that this set of values is the correct one.

Response of ChatGPT to authors' request to calculate the sample size for hypothetical study using the same conditions (ClinicalTrials.gov Identifier: NCT01306058) is displayed in Figure 6 and Figure 7. Similar to the previous example the formula displayed by GenAI does not look trustable – calculations lying in the foundation of Simon two stage design are based on binomial distribution. But let's leave aside the questionable correctness of the formulae and concentrate on more transparent things.

Let's verify now calculations displayed by ChatGPT (as earlier ignore at this stage the extent of the original formula's relevance to the hypothetical study). The first of the two lines marked by red rectangle (Figure 7) reads as

$$n_1 = \frac{1.6449^2 * 0.47 * 0. (1 - 0.47)}{(0.67 - 0.47)^2} \quad (4)$$

but the outcome essentially differs from one presented by ChatGPT – an arbitrary reader can easily confirm that the sequence $1.6449^2 * 0.47 * 0. (1 - 0.47)/0.2^2$ equals to 16.8497, and not to 34.93 as in Figure 7. Note, both values (16.85 and 34.93) are different from correct one that was used in the study (21).

Let now analyze the recommendations for conduction of the study prepared by ChatGPT (yellow and orange rectangles in Figure 7). On one hand, according to ChatGPT (yellow rectangle) all subjects in the first stage of the study (n_1) must be responders (meet the criteria) to provide investigators with a green light to start the second stage. In the reality of the aforementioned trial all 21 patients were supposed to be the responders, which is fundamentally incorrect. On the other hand (orange rectangle), ChatGPT instructs PI to proceed to the second stage only if a number of responders is 15 or more. Remember, CHatGPT did not share with the user where and how the criterion 14/15 was derived. Evidently, these two sets of instructions contradict each other and neither of them is correct one.



You

Calculate sample size and interim stopping rule using a Simon optimal two-stage design for the following clinical study:

A median 5.5 month time to progression, translating into a 6 month progression free survival rate (PFS) of 47%, would be considered not promising ($p_0=0.47$), while a 67% 6-month progression-free survival (PFS) rate (corresponding to a median time to progression of 10.5 months) would be considered promising ($p_1=0.67$);
alpha=0.10 and beta=0.10



ChatGPT

To calculate the sample size and interim stopping rule for a Simon optimal two-stage design, we'll follow these steps:

1. Determine the parameters:
 - p_0 (null hypothesis response rate)
 - p_1 (alternative hypothesis response rate)
 - alpha (Type I error rate)
 - beta (Type II error rate)
2. Calculate the sample size for each stage.
3. Set the interim stopping rules based on the outcomes of the first stage.
4. Apply the Simon's two-stage design formulae.

Let's calculate each step:

1. **Parameters:**
 - p_0 (null hypothesis response rate) = 0.47
 - p_1 (alternative hypothesis response rate) = 0.67
 - alpha (Type I error rate) = 0.10
 - beta (Type II error rate) = 0.10
2. **Calculate Sample Size for Each Stage:**
 - For the Simon's two-stage design, the sample size in the first stage is calculated using the formula:

$$n_1 = \frac{Z_{\alpha}^2 \times p_0 \times (1 - p_0)}{(p_1 - p_0)^2}$$

Figure 6. ChatGPT response (part 1) to the request to estimate sample size for the prospective study using the assumptions identical to ones in the Protocol NCT01306058.

where:

- Z_α is the z-value for the given alpha level (0.10)
- p_0 is the null hypothesis response rate (0.47)
- p_1 is the alternative hypothesis response rate (0.67)
- Calculate Z_α (z-value for alpha = 0.10)
 - From standard normal distribution tables, $Z_\alpha = 1.6449$

- Calculate sample size for first stage:

$$n_1 = \frac{1.6449^2 \times 0.47 \times (1 - 0.47)}{(0.67 - 0.47)^2}$$

$$n_1 \approx 34.93$$

Round up to the nearest whole number: $n_1 = 35$

- For the second stage, use the formula:

$$n_2 = \frac{Z_\beta + Z_\alpha}{p_1 - p_0} + n_1$$

- Z_β is the z-value for the given beta level (0.10)
- $Z_\beta = 1.2816$ (from standard normal distribution tables)

- Calculate sample size for second stage:

$$n_2 = \frac{1.2816 + 1.6449}{0.67 - 0.47} + 35$$

$$n_2 \approx 36.28$$

Round up to the nearest whole number: $n_2 = 37$

3. Interim Stopping Rule:

- If fewer than n_1 patients respond positively (meet the criteria), stop and conclude that the treatment is not promising.
- If n_1 or more patients respond positively, proceed to the second stage.
- If the total number of positive responses by the end of the study is $n_1 + n_2$, conclude that the treatment is promising.

So, in this Simon optimal two-stage design:

- Stage 1: Enroll 35 patients.
- If 14 or fewer patients meet the criteria, stop the trial.
- If 15 or more patients meet the criteria, proceed to Stage 2.
- Stage 2: Enroll additional 37 patients, making a total of 72 patients.
- If the total number of positive responses (meeting the criteria) reaches 51 or more by the end of the study, conclude that the treatment is promising.

Figure 7. ChatGPT response (part 2) to the request to estimate sample size for the prospective study using the assumptions identical to ones in the Protocol NCT01306058.

2. Calculate Sample Size for Each Stage:

For the Simon's two-stage design, the sample size in the first stage is calculated using the formula:

$$n_1 = \frac{Z_{\alpha}^2 \times p_0 \times (1-p_0)}{(p_1-p_0)^2}$$

where:

- Z_{α} is the z-value for the given alpha level (0.10)
- p_0 is the null hypothesis response rate (0.47)
- p_1 is the alternative hypothesis response rate (0.67)

Calculate Z_{α} (z-value for alpha = 0.10):

- From standard normal distribution tables, $Z_{\alpha} = 1.6449$

Calculate sample size for the first stage:

$$n_1 = \frac{1.6449^2 \times 0.47 \times (1-0.47)}{(0.67-0.47)^2}$$
$$n_1 \approx 33.63$$

Round up to the nearest whole number: $n_1 = 34$

For the second stage, use the formula:

$$n_2 = \frac{Z_{\beta}^2 \times p_1 \times (1-p_1) + Z_{\alpha} \times p_0 + Z_{\beta} \times p_0}{(p_1-p_0)^2}$$

Calculate Z_{β} (z-value for beta = 0.10):

- From standard normal distribution tables, $Z_{\beta} = 1.2816$

Calculate sample size for the second stage:

$$n_2 = \frac{(1.2816^2 \times 0.67 \times (1-0.67)) + (1.6449 \times 0.47) + (1.2816 \times 0.47)}{(0.67-0.47)^2}$$
$$n_2 \approx 22.87$$

Round up to the nearest whole number: $n_2 = 23$

Figure 8. ChatGPT response (part 3) to the request to estimate sample size for the prospective study using the assumptions identical to ones in the Protocol NCT01306058.

Repeating the same request to ChatGPT in the same chat provides us with response that is slightly different from the first version (Figure 6 and Figure 7). To save the space Figure 8 displays the most interesting part of the rerun. One can see that the expression for n_1 (blue rectangle) is identical to the formula from the first run, but the numeric result is different (33.63 vs 34.93). Even more intriguing facts can be found in the purple rectangle – it looks as ChatGPT suggests another formula for n_2 and the numeric value varies from the previous one (22.87 vs 36.82). Therefore, one can conclude that ChatGPT is not able to reproduce its response to the request (remind, no letter was changed in the text of the request while running it second, third, etc. time).

It would be worthwhile to note also that the difference between the responses provided by ChatGPT is changed dramatically when one opens a new chat and asks the same question that was submitted to ChatGPT in another chat. All the responses displayed in the Figure 6, Figure 7, Figure 8 were taken from the same chat (the identical requests were submitted a number of times). On the contrary, Figure 9 contains the response that was provided by ChatGPT in another chat. One can see that the content is

totally different – it does not display any formula or number. The response consists of a general description of what ChatGPT is going to do, but no action or result is given. Actually, ChatGPT did not respond to the submitted request, because the user asked the tool to estimate the required sample size, not to present the high-level instruction how it can be performed.



ChatGPT

In a Simon optimal two-stage design, the sample size and interim stopping rule are determined based on the desired response rates (p_0 and p_1), as well as the significance level (α) and power ($1 - \beta$) of the study.

In the first stage, a certain number of patients are enrolled, and if a pre-specified number of responses are observed, the study proceeds to the second stage. If the response rate is not met in the first stage, the study is stopped for futility.

To estimate the sample size and interim stopping rule, we can follow these steps:

1. Determine the parameters:
 - p_0 : The 6-month progression-free survival rate considered not promising (0.47).
 - p_1 : The 6-month progression-free survival rate considered promising (0.67).
 - α : The significance level (0.10).
 - β : The desired power (0.10).
2. Calculate the critical values for each stage:
 - For the first stage, the critical value is determined by the lower bound of the one-sided exact binomial test with a significance level of α .
 - For the second stage, the critical value is determined by the upper bound of the one-sided exact binomial test with a significance level of α .
3. Calculate the sample sizes for each stage:
 - For the first stage, calculate the sample size required to achieve the critical value for the first stage.
 - For the second stage, calculate the sample size required to achieve the critical value for the second stage, assuming the observed response rate from the first stage.
4. Determine the interim stopping rule:
 - If the observed response rate in the first stage falls below the critical value, the study is stopped for futility.
 - If the observed response rate in the first stage meets or exceeds the critical value, the study proceeds to the second stage.

Please note that I will now perform the calculations based on the provided parameters.

Figure 9. ChatGPT response (part 4) to the request to estimate sample size for the prospective study using the assumptions identical to ones in the Protocol NCT01306058.

FISHER EXACT TEST

Consider now the study “Prevention of overt hypothyroidism following radioactive iodine therapy for Graves’ disease; a novel protocol with implications for future studies” (ClinicalTrials.gov Identifier: NCT# 01950260). The trial was devoted to Graves’ disease (GD). The most common method of treating GD in the US relies on rendering the patient hypothyroid with the use of radioactive iodine (RAI) ablation of the thyroid gland followed by lifelong replacement with thyroid hormone (levothyroxine). The main goal of the study was to test the hypothesis that early initiation of levothyroxine following a precise protocol using small doses of the medication can prevent development of overt hypothyroidism without an increase in rate of symptomatic hyperthyroidism. The primary efficacy endpoint was an evaluation of difference in the incidence rates of hypothyroidism at 8 weeks and it was supposed to be tested using Fisher’s exact test.

The sample size for this study was based on the incidence of hypothyroidism primary endpoint using an $\alpha=0.10$ (two-sided), $\beta=0.10$. The proportions of participants that develop hypothyroidism in the placebo and experimental groups were assumed to be 60% and 20%, respectively. According to the Protocol a sample size of 30 per group was expected to provide 90% power to detect a difference between the group proportions of 40% using the two-sided Fisher’s Exact test.

Like what we did in the previous examples let’s check the values presented in the Protocol using independent calculation before asking ChatGPT to replicate this outcome. Free web-based statistical calculators (e.g., <https://homepage.univie.ac.at/robin.risti/samplesize.php?test=fishertest>) provides one with the exactly the same outcome – 30 patients per group. Commercial (licensed) software nQuery confirms this verification – 30 patients per group under given conditions. Therefore, one can be sure that this set of values is the correct one.

Responses of ChatGPT to authors’ requests (remind, all the requests are identical to each other) to calculate the sample size for hypothetical study using the same conditions (ClinicalTrials.gov Identifier: 01950260) are displayed in Figure 10 and Figure 11. Similar to the previous example the formula displayed by GenAI does not look trustable – calculations lying in the foundation of Simon two stage design are based on binomial distribution. But let’s leave aside a questionable correctness of the formulae and concentrate on more transparent things.

Let us analyze first Figure 10 and start with an evident observation – ChatGPT provides the user with the explicit formula for calculation of the sample size (like one that was used, for example for t-test). It is well known, however, that there is no explicit formula for Fisher exact test. The test assesses the null hypothesis of independence applying hypergeometric distribution of the numbers in the cells of the table; therefore, its use requires numeric calculations. Many free software packages provide the results of Fisher’s exact test for 2×2 contingency tables (but not for bigger contingency tables with more rows or columns).

Next, one can easily find clear similarities between this response (Figure 10) and ones that we have analyzed earlier (where other statistical tests were applied). Namely, GenAI makes up the results in some cases instead of really calculating them even if it is simple multiplication. The first line in the orange rectangle (Figure 10) contains 2.927^2 and the next line there displays the outcome of this calculation – 8.584. But simple calculator says that the right answer is 8.567329 that cannot be rounded to the displayed value by any means. In addition, the denominator of the first line in the first line in the orange rectangle contains 0.16^2 , but the next line shows the value 0.04, what is wrong (remind, contains $0.16^2 = 0.0256$). Finally, the estimated sample size is 86 subjects per group which by far more than was enlisted in reality (in actual study NCT# 01950260) and was confirmed by independent calculations using both free and commercial software.

Let’s now proceed to the next important item in the same algorithm of our investigation – reproducibility of ChatGPT responses. Repeating the same request to ChatGPT in the same chat provides us with absolutely different response (refer to Figure 11). To save the space, the initial request is omitted here because it is identical to one shown in the previous figure (Figure 10). One can see that the symbols for n, n_1, n_2 (blue rectangle) appear now in both LHS and RHS of the expression. To resolve the equation and to get final value of n ChatGPT correctly assumes that the two groups are equal in size and suggests applying iterative calculations (Figure 11, Step 5) to the formula. It makes an impression that ChatGPT “knows” about iterative method of calculations, but clearly its practical implementation requires some

improvement – what about a convergence of the obtained sequence of numbers? Why did ChatGPT stop after second iteration? And many other questions that remained unanswered. No need to say that the outcome of this attempt is not even close to the right value – ChatGPT thinks that one needs to enlist 178 subjects per group based on the conditions identical to ones used in the actual study NCT# 01950260 (remind, correct answer is 30 subjects per group).



Anonymous

This study is based on the incidence of hypothyroidism primary endpoint using an $\alpha=0.10$ (two-sided), $\beta=0.10$. The proportions of participants that develop hypothyroidism in the placebo and experimental groups are assumed to be 60% and 20%, respectively using the two-sided Fisher's Exact test. Estimate the sample size using above conditions.



ChatGPT

To estimate the sample size required for comparing the proportions of participants developing hypothyroidism in the placebo and experimental groups using the two-sided Fisher's Exact test, we can use the formula:

$$n = (Z_{\alpha/2} + Z_{\beta})^2 * (p_1 * (1 - p_1) + p_2 * (1 - p_2)) / (p_1 - p_2)^2$$

Where:

- n is the required sample size per group
- $Z_{\alpha/2}$ is the critical value corresponding to the desired significance level α ($0.10/2 = 0.05$ for a two-sided test)
- Z_{β} is the critical value corresponding to the desired power ($1 - \beta = 0.90$, so $\beta = 0.10$)
- p_1 is the proportion of participants developing hypothyroidism in the placebo group (0.60)
- p_2 is the proportion of participants developing hypothyroidism in the experimental group (0.20)

Let's calculate the sample size:

$Z_{\alpha/2} = 1.645$ (lookup from the standard normal distribution table for a two-sided test with $\alpha = 0.05$)

$Z_{\beta} = 1.282$ (lookup from the standard normal distribution table for a desired power of 0.90)

$$n = (1.645 + 1.282)^2 * ((0.60 * (1 - 0.60)) + (0.20 * (1 - 0.20))) / (0.60 - 0.20)^2$$

$$n = 2.927^2 * (0.24 + 0.16) / 0.16^2$$

$$n = 8.584 * 0.4 / 0.04$$

$$n = 85.84$$

Rounding up to the nearest whole number, the estimated sample size required per group is approximately 86 participants.

Figure 10. ChatGPT response (part 1) to the request to estimate sample size for the prospective study using the assumptions identical to ones in the Protocol NCT01950260.



ChatGPT

To estimate the sample size for comparing the proportions of participants developing hypothyroidism in the placebo and experimental groups using the two-sided Fisher's Exact test, we can use the following approach:

1. Determine the effect size, which is the difference between the proportions of participants with hypothyroidism in the two groups:
 $p1 - p2 = 0.60 - 0.20 = 0.40$
2. Calculate the standard deviation (SD) of the difference in proportions:
 $SD = \sqrt{[(p1 * (1 - p1))/n1 + (p2 * (1 - p2))/n2]}$
 Since we don't have specific sample sizes yet, we can assume equal sample sizes for simplicity:
 $SD = \sqrt{[(0.60 * (1 - 0.60))/n + (0.20 * (1 - 0.20))/n]}$
 $SD = \sqrt{[(0.24 + 0.16)/n]}$
 $SD = \sqrt{0.40/n}$
3. Determine the critical value ($Z_{\alpha/2}$) corresponding to the desired significance level α ($0.10/2 = 0.05$ for a two-sided test).
 $Z_{\alpha/2} = 1.96$ (lookup from the standard normal distribution table for a two-sided test with $\alpha = 0.05$)
4. Determine the critical value (Z_{β}) corresponding to the desired power ($1 - \beta = 0.90$, so $\beta = 0.10$).
 $Z_{\beta} = 1.28$ (lookup from the standard normal distribution table for a desired power of 0.90)
5. Calculate the required sample size (n) per group:
 $n = [(Z_{\alpha/2} + Z_{\beta}) * SD / (p1 - p2)]^2$
 $n = [(1.96 + 1.28) * \sqrt{0.40/n} / 0.40]^2$

To solve for n , we need to perform an iterative calculation. Let's start with an initial guess of $n = 100$:

$$n = [(1.96 + 1.28) * \sqrt{0.40/100} / 0.40]^2$$

$$n = [(3.24) * (0.0632) / 0.40]^2$$

$$n = 0.0515^2$$

$$n \approx 0.0027$$

The result is not a reasonable sample size. Let's continue the iteration:

$$n = [(1.96 + 1.28) * \sqrt{0.40/0.0027} / 0.40]^2$$

$$n = [(3.24) * (8.7189) / 0.40]^2$$

$$n \approx 177.36$$

Rounding up to the nearest whole number, the estimated sample size required per group is approximately 178 participants.

Figure 11. ChatGPT response (part 2) to the request to estimate sample size for the prospective study using the assumptions identical to ones in the Protocol NCT01950260.

DISCUSSION

The present research represents a natural continuation of the previous work of the authors (Goldfarb *et al.*, 2022) and allows us to conclude that unsupervised application of GenAI for such specific statistical task as sample size estimation faces multiple challenges and pitfalls. Aforementioned examples clearly demonstrate that there are a few disadvantages that must be overseen and addressed appropriately while working with ChatGPT on similar subjects. This Chapter summarizes our experience.

ChatGPT “does not know” the appropriate formulae for sample size and even “does not realize” that in some cases the explicit formula is not available. The authors analyzed completed clinical trials where various statistical tests were (t-test, chi-square, Simon two-stage, etc.) used, and it was shown very clearly that this GenAI tool is not able to operate with correct formulae. It can be understood to some extent if to remember that GenAI tools with LMM in their foundation are not exactly the library where one can find an exact answer to the question after some time spent on search. The underlying models were trained on billions of sources and documents, but when you ask for an exact formula, quote or citation – most likely the system makes it up relying on the specific statistical considerations of appropriate machine learning model. The issue is complicated even more because of the very convincing way of the response that ChatGPT provides its customer. A natural human desire to rely on computer-based technology is heated by magic words “artificial intelligence” and it adds additional bias to our inability to critically consider the outcomes of application of modern technologies.

The authors showed that ChatGPT cannot be trusted when it presents the results of elementary mathematical operations. One could see in numerous examples that plugging in the numbers displayed by this GenAI tool does not allow us to obtain the same result that ChatGPT presents on the screen. This disadvantage was mentioned by multiple authors (refer, for example, to Azaria, 2022, Chen, 2023, Borhi, 2023, etc.)

Finally, ChatGPT failed the test for reproducibility. Remember, reproducibility means obtaining consistent computational results using the same input data, computational steps, methods, code, and conditions of analysis. As we can see the obtained results are not reproduceable – every run of the same requests generates different response – and it relates to both formulae and numeric values. And the situation became even worse when one starts a new chat and copies there the request that was used in another chat (opened by the same or another person). The response in the new chat can be (as the authors showed above) dramatically different from what was responded earlier in the original chat.

The issue of reproducibility is potentially less important while generating human-like answers to a question of the type “What was the weather in London yesterday?”. Two versions (presumably obtained in two runs) “It was relatively warm, a little bit rainy day” and “There was rain for half a day yesterday” are probably good enough for human ear. Moreover, both probably correctly describe yesterday’s conditions in London. But this situation is unacceptable when we are handling stuff related to exact sciences. Any reasonable statistician or programmer would expect that the computer-based tool will produce identical responses to the same request. As one could see it is not the case when we are working with ChatGPT.

The goal of this paper is to attract attention of the professionals working in the pharmaceutical industry to an absolute necessity of extremely careful utilization of the current GenAI tools using LLM and LMM as their foundation. GenAI development rapidly changes the way how the humanity lives and works, but its development is in the constant process. Therefore, to correctly evaluate a sample size for the prospective study and to avoid undesired consequences of a wrong estimation an arbitrary pharmaceutical company must have onboard an experienced statistician. One of his tasks should be to oversee the results of GenAI application in case these tools were used for the declared purposes. All this is because GenAI may generate inaccurate information presented with conviction, making it extremely dangerous for an inexperienced statistician or programmer. Therefore, the leadership of experienced statistician is of vital importance for successful conduction of clinical trials in our age of GenAI surrounding us.

CONCLUSIONS

To recap the discussion of the situation with computing and reporting of the sample size in RCT it would be worthwhile to summarize the performed analysis:

1. Literature review and the research conducted by the authors highlighted the potential risks associated with excessive reliance on ChatGPT in sample size estimation. The limitations include factors like incorrect and fictitious responses, inaccuracies in code, limited logical reasoning abilities, overconfidence, etc.
2. ChatGPT demonstrates inability to perform correctly elementary arithmetical calculations.
3. Outputs produced by ChatGPT contain hallucinations. The GenAI tool provides the user with incorrect (never existed) formulae.
4. Responses generated by ChatGPT are not reproduceable. Two subsequent runs of the same request may produce two different answers to the same question under consideration.
5. Estimation of a sample size for a prospective study and consequent development of the statistical section of a protocol should be performed by an experienced statistician equipped with appropriate knowledge and tools.

REFERENCES

- Azaria, A. 2022. *ChatGPT Usage and Limitations*. hal-03913837.
- Azaria, A., Azoulay, R., Reches, S. June 2, 2023. ChatGPT is a Remarkable Tool - For Experts. arXiv:2306.03102.
- Chen, L., Zaharia M., Zou, J. October 31, 2023. How Is ChatGPT's Behavior Changing over Time? arXiv:2307.09009.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences, (2nd edn.)*. Hillsdale, NJ, USA: Routledge.
- Countess of Lovelace, "Translator's notes to an article on Babbage's Analytical Engiro". Scientific Memoirs (ed. by R.Taylor), vol.3(1842), 691–731.
- Goldfarb, I., Karwal, R., Shu, X. 2022. "Analysis of sample size calculations in clinical trial – errors, pitfalls, and conclusions". *Proceedings of PharmaSUG-2022*, Paper SA-142.
- Ignjatović A, Stevanović L. "Efficacy and limitations of ChatGPT as a biostatistical problem-solving tool in medical education in Serbia: a descriptive study." *Journal of Educational Evaluation for Health Professions*. 2023;20:28. doi: 10.3352/jeehp.2023.20.28. Epub 2023 Oct 16.
- Julious, S. A. (2009) *Sample sizes for clinical trials*. Chapman & Hall/CRC, FL, USA. Boca Raton.
- "Large Language Model" Available at https://en.wikipedia.org/wiki/Large_language_model.
- Methani, J., Latiri, I., Dergaa, I., Chamari, K. "ChatGPT for sample-size calculation in sports medicine and exercise sciences: a cautionary note." *International Journal of Sport Physiology and Performance*, 2023:1-5. Available at <https://api.semanticscholar.org/CorpusID:260485789>
- Morton, J.M., Philips, N.J., Taylor, L.F., McGowan, M.R. 2013, "Bovine viral diarrhoea virus in beef heifers in commercial herds in Australia: mob-level seroprevalences and incidences of seroconversion, and vaccine efficacy". *Australian Veterinarian Journal*. Volume no.91:517–524

Open AI, "GPT-4 system card". Accessed 28 May 2023. <https://cdn.openai.com/papers/gpt-4-system-card.pdf>

Simon. R. (1989). Optimal Two-Stage Designs for Phase II Clinical Trials. *Controlled Clinical Trials* 10: 1-10.

Spitale G, Biller-Andorno N, Germani F. *AI model GPT-3 (dis)informs us better than humans*. *Sci Adv*. 2023;9(26):eadh1850. doi:10.1126/sciadv.adh1850.

Study "A Randomized Controlled Trial Comparing the Effect of Topical Dorzolamide-Timolol versus Placebo Combined with Intravitreal Anti-Vascular Endothelial Growth Factor (VEGF) Injections in Patients with Neovascular Age-Related Macular Degeneration Who Are Incomplete Anti-VEGF Responders" (ClinicalTrials.gov Identifier: NCT03034772).

Study "A Phase I/II study of TRC105 in combination with Sorafenib in Hepatocellular Carcinoma (HCC)" (ClinicalTrials.gov Identifier: NCT01306058).

Study "Prevention of Overt Hypothyroidism Following Radioactive Iodine Therapy for Graves' Disease" (ClinicalTrials.gov Identifier: NCT01950260).

Turing, A. October 1950. Computing Machinery and Intelligence. *Mind*, Volume LIX, Issue 236: Pages 433–460.

World Health Organization; 2024. "Ethics and governance of artificial intelligence for health. Guidance on large multi-modal models." Geneva: License: CC BY-NC-SA 3.0 IGO.

Yanagita Y, Yokokawa D, Uchida S, Tawara J, Ikusaka M. "Accuracy of ChatGPT on Medical Questions in the National Medical Licensing Examination in Japan: Evaluation Study". *JMIR Form Res*. 2023 Oct 13;7:e48023. doi: 10.2196/48023. PMID: 37831496; PMCID: PMC10612006.

ACKNOWLEDGMENTS

The authors are very thankful to upper management of Accenture Life Sciences for their constant support of this work.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Please feel free to contact anyone of the authors at:

Igor Goldfarb
Accenture Life Sciences R&D
Igor.goldfarb@accenture.com

Vikas Sharma
Catalyst Clinical Research
Vikas.Sharma@catalystcr.com