Przegląd Statystyczny. Statistical Review, 2022, vol. 69, 4, 20–28 https://doi.org/10.59139/ps.2022.04.2
Received: 4.01.2023, revised: 31.03.2023, accepted: 15.04.2023

Sample size in clinical trials – challenges and approaches

Andrzej Tomski,[a] Barbara Gorzawska[b]

Abstract. Sample size estimation is a necessary and crucial step in clinical trial research. Statistical requirements, limited patient availability and high financial risk of a clinical trial necessitate the proper calculation of this measure. The aim of this paper is to discuss the reasons why the estimation of the sample size is important and, based on the obtained results, to show how this process may be completed in selected cases. Stochastic simulations based on the Monte Carlo methods approach are applied. Therefore, new challenges facing this area of research are mentioned.

Keywords: sample size, clinical trial, Monte Carlo methods, stochastic simulations

JEL: C13, C15, C18

1. Introduction

Estimating the sample size is an important issue, relating in the recent years particularly to statistics for clinical trials. Clinical trials are prospective biomedical research studies on humans designed to answer crucial questions about biomedical interventions, including new treatments (National Institutes of Health, n.d.). While scientists in some disciplines have easy access to all of the data representing their research topic, it is not the case in the field of clinical trials. In biology, scientists need to obtain a sufficient portion of research material, while in medicine, they have to select patients who are suitable for therapy and only then begin the research on the treatment. Clinical trials are even more problematic, as they are expensive, with a high level of formal requirements and involve a complex patient recruitment process. Sample size estimation is a problematic procedure, as it is costly, time-consuming, the number of subjects fit for the process tends to be limited, while the statistical requirements are very specific. On the other hand, estimating the sample size is mandatory in these studies before any patients are even recruited.

At the same time, with the increasing number of instances requiring the estimation of a sample size before a given study is initiated, no data relating to it is available. Therefore, a problem arises, for example, when a researcher wants to apply

[a] University of Silesia in Katowice, Institute of Mathematics, ul. Bankowa 12, 40-007 Katowice, e-mail: andrzej.tomski@us.edu.pl, ORCID: https://orcid.org/0000-0001-6944-0600.

[b] Parexel FSP, Parexel International, Warsaw, Poland, e-mail: gorzawska.barbara@gmail.com, ORCID: https://orcid.org/0000-0002-4263-8489.

for a grant and needs to provide the precise costs and details of a study, in particular the number of samples needed. Another issue is the fact that researchers tend not to consider such concepts as the test power, errors or the effect size. Researchers pay great attention mainly to the concept of the p-value, while ignoring other statistical indicators (Amrhein et al., 2019). Consequently, the results of many studies are concluded mainly on the basis of low p-value results considered as significant. On the other hand, the non-random components (Szreder, 2022) of the total survey error do not have to decrease as the sample size increases. Non-random errors such as coverage errors or measurement errors generate a bias which does not depend on the sample size (Chin, 2012).

Relying only on the p-value has been widely criticised in the recent years (Platek and Särndal, 2001) and researchers have trouble distinguishing between statistical significance and practical significance. It should be noted that statistical significance does not mean significance in general. For example, the term 'clinical relevance' refers to the practical significance of a treatment effect. Researchers focusing solely on statistical significance or the lack of it may report results that are not significant in practice. This can occur especially when a large sample size is considered. A large sample size is crucial when determining statistical significance and in such a case confusion relates to its interpretation. However, an article published by the American Statistical Association (ASA) presents a formal statement explaining several commonly agreed upon principles underlying the correct use and interpretation of p-values (Wasserstein & Lazar, 2016).

In this paper, we will briefly describe the requirements and standards of clinical trials to show the importance of sample size estimation. Researchers of other disciplines interested in this issue may also consider adopting a similar approach in their field of study.

2. Clinical trial guidelines

Medicine and pharmacy are disciplines most interested in sample size estimation, although the group of people involved in those fields also greatly benefit from sample size estimation as it saves them time, money and effort. This is definitely important, especially when gathering data requires significantly more work from the researcher than just downloading them from the internet.

Recent years have shown that clinical trials are a very dynamically developing field of study, thus the demand for its corresponding statistical methods is constantly growing. In the initial phase of a study, a clinical study protocol[1] and a statistical

[1] A protocol is a document that describes how a clinical trial will be conducted and ensures both the safety of the subjects and the integrity of the collected data.

analysis plan[2] for the research are developed. Both require the sample size of the patients recruited for the study. The principles for conducting such surveys do not clearly indicate how the sample size should be estimated. However, there are some requirements in terms of the statistical and medical demand. Guidelines for statisticians are provided by the European Medicines Agency (EMA, n.d.) and the U.S. Food and Drug Administration (n.d.). Statistical analysis, therefore, does not start with the receipt of data and the selection of specific tests. In practice, the researcher first formulates the objectives of the study and determines the acceptable level of significance α. According to the EMA guidelines, it is set at the level of 5%. The Sponsor[3] expects a test to be constructed so that its power is as high as possible for the fixed α level. This is where the interest of the Sponsor (who wants the most powerful test) and the patients (who just want effective treatment) may clash. As a result, the statistician recommends to both the Sponsor and the Investigator[4] that the minimum sample size for the test power exceeds the acceptable level to at least 80% in accordance with the EMA guidelines (EMA, n.d). Thus, we assume that

$$\alpha = 5\% \, and \, 1-\beta \geq 80\%, \tag{1}$$

where β is a type II error rate.

The next step involves gathering a sufficiently large sample size by the Sponsor and Investigator by inviting the required number of subjects to participate in the study until its completion. After that, the main statistical analysis can start. In practice, interim analyses[5] (Hayes and Patterson, 1921) are performed in such cases, i.e. research is conducted after only a part of the recommended sample is collected. Performing interim analyses ensures the safety of the study so that the risk of complications, in particular serious adverse events[6] is minimised. Additionally, if the study shows that the administered treatment is ineffective, it may be discontinued, thus saving any further effort. While the study can be stopped at any moment due to patient safety concerns or unsatisfactory results indicated by the interim analyses, their status cannot be in any way considered as confirmatory. Therefore, in order to pronounce the treatment effective, the study has to be continued. This means that it is not possible to perform a statistical analysis by collecting a smaller sample to

[2] A statistical analysis plan (SAP) outlines the analytical approach of the data collected in a clinical trial.

[3] Legal person who funds the research.

[4] It is a person who is involved in running a clinical trial.

[5] It is an analysis of data that is conducted before data collection is completed.

[6] An adverse event is any undesirable experience associated with the use of a medical product on a patient. The event is considered serious when its outcome is life-threatening or it leads to the patient's death, hospitalisation or permanent health impairment.

confirm the investigated hypotheses, namely the effectiveness of the drug. The issue of sample size estimation must also take into account the results of the recently introduced non-inferiority tests, which assume a certain margin of error compared to reference objects, i.e. existing drugs.

3. Sample size calculation – selected computational tools

Sample size estimation is a topic that is growing in popularity along with the big data sector, the increase of the number of clinical trials (Delgado et al., 2018) and technological advances. Numerous reference books (Chow et al., 2007) and software tools for estimating sample size in the simplest cases are widely available. These tools are in most cases very user-friendly and publicly provided, nevertheless, they do have certain disadvantages.

Many sample size calculators are available on the Internet, although here we will provide examples of two of them to describe the nature of their work.

The G*Power (Faul et al., 2009) is a tool used to perform statistical power analyses for many different tests, including t-tests, F tests, χ^2 tests, z-tests and some exact tests. The G*Power can also be used to compute effect sizes and to depict the results of power analyses. In order to calculate the sample size, it only requires the user to select a test from a list, provide a measure of the effect size and enter the test power along with its significance level. However, the program has some disadvantages: the test selection is limited to a list, there is no possibility to specify the exact form of the statistical model or link the results to confidence intervals. Moreover, the lack of access to its source code raises questions as to how these estimates were obtained.

Similarly, the Sample Size Calculator (Raosoft, 2004) does not refer to the type of the investigated variable. It does not even offer a choice of the test. This tool requires from the user to enter the α level, the upper limit of the sample size and a confidence level without any explanation. Its overall applicability seems to be quite limited and thus it may not be able to estimate sample size in certain models accurately enough. What is more, the graphical interface of the application shows a lot of additional windows, which may surprise and confuse the user.

Online sample size estimation tools tend to offer a narrow range of possibilities. It is sometimes difficult to clearly identify which models they refer to or they are intended for a very simple and one specific statistical model. Another serious disadvantage of these models is that, in general, they do not refer to scientific results obtained in papers presenting this type of research. In conclusion, a more universal method needs to be devised offering an efficient approach for a wide variety of statistical models, but which would also refer to the results provided in the literature.

As in the case of the numerous online tools, there are many books and research articles extensively discussing the issue of sample size estimation.

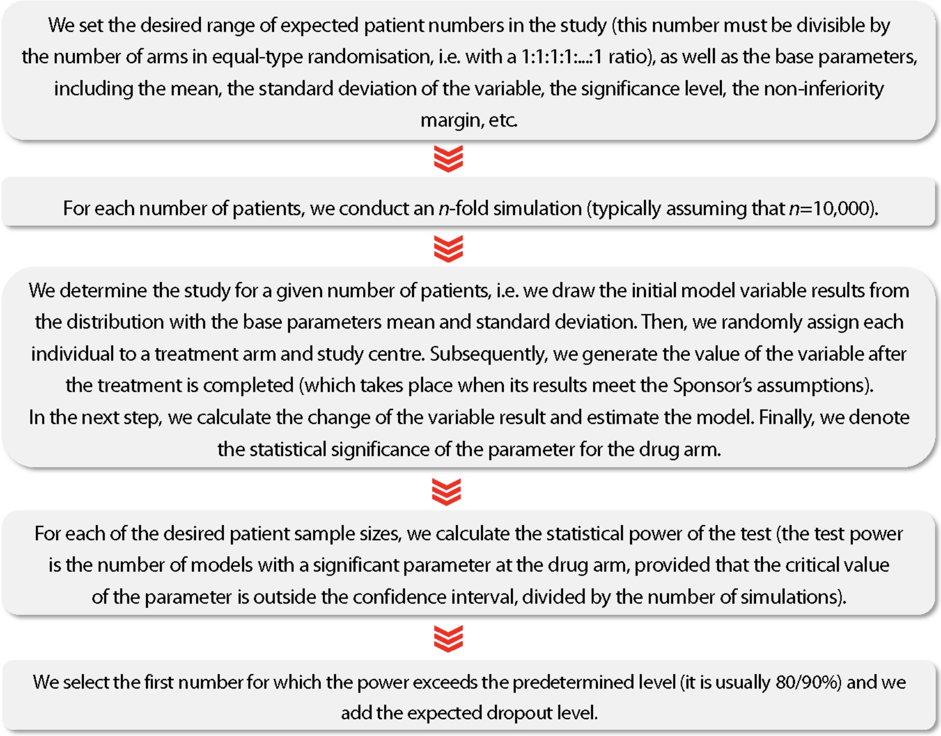
3.1. Monte Carlo methods – a brief review

Monte Carlo methods constitute a large class of computational algorithms that are based on repeated random sampling aiming to obtain some numerical results. In general, the application of a Monte Carlo method involves the limitation of the power of the test in order to obtain the sample size, i.e. it allows the evaluation of the parameter bias and the power of the test based on computer-generated population data. The most popular statistical methods include using parameter values determined in a study from the past, meta-analysis techniques or self-estimated statistical parameters. However, if the research concerns newly discovered diseases, there is basically only a third way: using self-estimated statistical parameters. When the population data is obtained, samples of a given size from a certain range are generated. This enables the calculation of model parameters and the power of the test is able to satisfy the expected demands. It must be noted that the parameters' bias is strongly associated with the deviation of the experimentally estimated value from the set of the replicated estimates. The details of the approach estimating the sample size with many successful experimental applications is presented in the next section.

3.2. Monte Carlo-based approach

This part of the paper focuses on an approach to sample size estimation that combines the concepts of p-values with the test power, the effect size and the use of confidence intervals. As a result, the estimate is not based on the debatable p-value alone, but also on other relevant significance criteria stated in the study. Monte Carlo methods have been present in statistics for quite some time (see, for example, papers like Jiang et al., 2012). However, this paper attempts to systematise this approach in the form of a fairly simple scheme with reference to several important aspects. It must be emphasised here that the sample size is calculated for a specific research hypothesis and for a specific statistical test which was selected on the basis of this hypothesis. The adopted approach is based on the scheme illustrated in the following Figure.

Figure. A structured approach to sample size estimation in a typical clinical trial



We set the desired range of expected patient numbers in the study (this number must be divisible by the number of arms in equal-type randomisation, i.e. with a 1:1:1:1:...:1 ratio), as well as the base parameters, including the mean, the standard deviation of the variable, the significance level, the non-inferiority margin, etc.

For each number of patients, we conduct an *n*-fold simulation (typically assuming that *n*=10,000).

We determine the study for a given number of patients, i.e. we draw the initial model variable results from the distribution with the base parameters mean and standard deviation. Then, we randomly assign each individual to a treatment arm and study centre. Subsequently, we generate the value of the variable after the treatment is completed (which takes place when its results meet the Sponsor's assumptions). In the next step, we calculate the change of the variable result and estimate the model. Finally, we denote the statistical significance of the parameter for the drug arm.

For each of the desired patient sample sizes, we calculate the statistical power of the test (the test power is the number of models with a significant parameter at the drug arm, provided that the critical value of the parameter is outside the confidence interval, divided by the number of simulations).

We select the first number for which the power exceeds the predetermined level (it is usually 80/90%) and we add the expected dropout level.

Source: authors' work.

The presented approach strengthens the role of not only the effect size, but also the confidence intervals, because the result is not considered significant if the confidence interval includes a critical value suggesting that there is no difference for the estimated parameters.

This part of the paper provides an example illustrating the discussed approach. The Visual Analogue Scale (VAS) is a pain rating scale used for the first time in 1921 by Hayes and Patterson (1921). We consider a study with a primary endpoint[7] measured by a decrease in VAS between the pre-treatment and the 30-day treatment. The study specifies that randomisation[8] assigns the patient to one of the four arms:[9] placebo, low dose, medium dose and high dose. Three medical centres[10] participated in the study. The relevant literature indicates that patients with the examined disease evaluate their pain on the VAS scale, e.g. at an average of 7.5 with a standard

[7] Main hypothesis in a study.

[8] Patients are randomly assigned to the control group and to the treatment group.

[9] Arm in a clinical trial refers to each group or subgroup of participants that receives specific interventions (or placebo) according to the protocol.

[10] A place where an experiment is conducted.

deviation of 1.0. The Sponsor expects the patients receiving placebo to have an average VAS score of 5.5 with a standard deviation of 1.5 after 30 days, while the patients receiving the study drug an average VAS of 4 with a standard deviation of 1.0. We carry out 10,000 study simulations in order to estimate certain parameters in the study, i.e. from the distribution with base parameters m (mean) and s (standard deviation) and for the appropriate number of patients, we draw their initial VAS result, then we randomly assign them to an arm and a centre; in the next step we generate a VAS result after 30 days of treatment based on the Sponsor's assumptions; we then calculate the change on the VAS scale and estimate the ANCOVA model, which takes the following form:

$$\Delta VAS_{ijk} = \beta_0 + \tau_i + \beta_1 x_{ijk} + \varphi_k + \epsilon_{ijk}, \tag{2}$$

where:

ΔVAS_{ijk} is the difference between the outcome and the baseline of the VAS grade for the j-th patient under the i-th treatment in the k-th centre

x_{ijk} is the baseline VAS grade for the j-th patient under the i-th treatment in the k-th centre,

τ_i are the fixed treatment effects,

φ_k are the fixed centre effects,

β_0 and β_1 are regression coefficients,

ϵ_{ijk} are independent random variables with the $N(0, \sigma)$ distribution.

In the final step, we record the significance of the parameter at the treatment arm. The percentage of statistically significant results (provided that the confidence interval for this parameter does not contain zero) constitutes the estimate of the power of the test. We select the smallest number n for which the power exceeds 80%. The full implementation of our single Monte Carlo experiment in the R programming language for a sample dataset is provided in the supplementary material (available upon request submitted to the authors via e-mail).

4. Limitations of the study

This section outlines certain limitations of the approach presented above. Besides the previously mentioned lack of literature on the new issues, the problem is the lack of extensive information in the literature on the proper distribution of data. In many cases, only the basic central tendency measures are calculated. In order to perform a simulation, however, we need to specify the distribution precisely. In the proposed solution, we use a normal distribution, but the question is how precisely this distribution approximates the values of the studied parameter. Therefore, some assumptions have to be made at this stage as well. Sometimes we can use a chart from an article, although it still tends to provide quite scarce information. Whether a given value has any constraints (i.e. discrete variables, possible minimums and maximums) should also be taken into account. These problems may escalate when more advanced statistical methods (e.g. mixed models, survival analysis) or a more complicated study design (e.g. crossover arms as described in Yeh et al., 2020) are used in the clinical trial. It also raises questions as to the validation method and the strictly numerical accuracy of the model. In these cases the number of uncertainties resulting from the lack of information can increase in the future.

5. Conclusion

This paper discussed the issue of sample size estimation in a clinical trial. Various approaches used to calculate this value were described. We proposed an approach in the form of a diagram based on the Monte Carlo methods. Our aim was to draw researchers' attention not only to the problem itself, but also to the role of effect sizes and confidence intervals in such estimates. This is a serious challenge for the further development of statistics, involving almost all the key concepts of modern statistics. The inclusion of non-inferiority tests or meta-analyses as new directions of change give hope that the raised aspect will be further discussed and developed.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have occurred and influenced the work reported in this paper.

References

Amrhein, V., Greenland, S., & McShane, B. (2019, March 20). *Scientists rise up against statistical significance*. https://www.nature.com/articles/d41586-019-00857-9.

Chin, R. (2012). *Adaptive and Flexible Clinical Trials*. CRC Press.

Chow, S.-C., Wang, H., & Shao, J. (2007). *Sample Size Calculations in Clinical Research* (2nd edition). CRC Press. https://doi.org/10.1201/9781584889830.

Delgado, D. A., Lambert, B. S., Boutris, N., McCulloch, P. C., Robbins, A. B., Moreno, M. R., & Harris, J. D. (2018). Validation of Digital Visual Analog Scale Pain Scoring With a Traditional Paper-based Visual Analog Scale in Adults. *Journal of the American Academy of Orthopaedic Surgeons. Global Research & Reviews*, *2*(3), 1–6. https://doi.org/10.5435/JAAOSGlobal-D-17-00088.

European Medicines Agency. (n.d.). *ICH E9 statistical principles for clinical trials – Scientific guideline*. Retrieved October 14, 2019, from https://www.ema.europa.eu/en/ich-e9-statistical-principles-clinical-trials-scientific-guideline.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149.

Hayes, M. H. S., & Patterson, D. G. (1921). Experimental development of the graphic rating method. *Psychological Bulletin*, *18*(2), 98–99.

Jiang, Z., Wang, L., Li, C., Xia, J., & Jia, H. (2012). A Practical Simulation Method to Calculate Sample Size of Group Sequential Trials for Time-to-Event Data under Exponential and Weibull Distribution. *PLOS ONE*, *7*(9), 1–10. https://doi.org/10.1371/journal.pone.0044013.

National Institutes of Health. (n.d.). *Clinical Research Trials and You: The Basics*. Retrieved October 3, 2022, from https://www.nih.gov/health-information/nih-clinical-research-trials-you/basics.

Platek, R., & Särndal, C. E. (2001). Czy statystyk może dostarczyć dane wysokiej jakości?. *Wiadomości Statystyczne*, *46*(4), 1–21.

Raosoft. (2004). *Raosoft Sample Size Calculator*. http://www.raosoft.com/samplesize.html.

Szreder, M. (2022). Szanse i iluzje dotyczące korzystania z dużych prób we wnioskowaniu statystycznym. *Wiadomości Statystyczne. The Polish Statistician*, *67*(8), 1–16. https://doi.org/10.5604/01.3001.0015.9704.

U.S. Food and Drug Administration. (n.d.). *Guidance for Industry. E9 Statistical Principles for Clinical Trials*. Retrieved September 11, 2019, from https://www.fda.gov/regulatory-information/search-fda-guidance-documents/e9-statistical-principles-clinical-trials.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's Statement on *p*-Values: Context, Process, and Purpose. *The American Statistician*, *70*(2), 129–133. http://dx.doi.org/10.1080/00031305.2016.1154108.

Yeh, J., Gupta, S., Patel, S. J., Kota, V., & Guddati, A. K. (2020). Trends in the crossover of patients in phase III oncology clinical trials in the USA. *Ecancermedicalscience*, *14*, 1–8. https://doi.org/10.3332/ecancer.2020.1142.