

Sample-Size Estimation

Reprinted from New Drug Development: Design, Methodology, and Analysis, by J. Rick Turner, with permission of John Wiley & Sons, Inc.

9.1. Introduction

Sample-size estimation is the process by which a researcher decides how many subjects to include in a given clinical trial. Sample-size estimation is a critical part of the design of clinical trials, and, like all design issues, this must be addressed in the study protocol before the trial commences.

Many sources use the terms "sample-size determination" or "sample-size calculation" when discussing this issue. This book uses the term sample-size estimation to emphasize that deciding on the sample size that will be employed in a clinical trial is a process of estimation that involves both statistical and clinical informed judgment and not a process of simply calculating the "right" answer. It is true that mathematical calculations are made in this process, and, for a given set of values that are placed into the appropriate formula in any given circumstance, a precise answer will be given. However, the values that are placed into the formula are chosen by the sponsor.

Some of the values that need to be entered into the formula are typically chosen from a standard set of possibilities, with the researcher deciding which of several generally acceptable values is best suited for the intentions of a given trial. Other values are estimates based on data that may be available in existing literature or may have been collected in an earlier trial in the clinical development program. These include the estimated treatment effect and the variability associated with the estimated treatment effect. Deciding upon the sample size for a given clinical trial is a balancing act in which several factors need to be considered to achieve the balance desired by the sponsor.

The likelihood of a successful outcome (at least from the point of view that "success" means obtaining a statistically significant result) can be increased by increasing the sample size. When designing a study, the researcher wants to ensure that a large enough sample size is chosen to be able to detect an important difference that does in fact exist. It is certainly possible that a trial can fail to demonstrate such a difference simply because the sample size chosen was too small. Therefore, it might appear reasonable to think that a very big sample size is a good idea. However, there are ethical issues that must be considered when choosing the sample size (see the following section). Additionally, increasing the sample size increases the expenses, difficulties and overall length of a trial. Somewhere, for each sponsor and each study, an acceptable sample size needs to be chosen that balances the likelihood of a statistically significant result with the cost and time involved in conducting the clinical trial. Indeed, some sponsors have proposed decision-making models to estimate the ideal sample size incorporating various factors such as the length of the study, the financial costs, and usual statistical considerations (see Pallay, 2000).

9.2. Ethical Issues in Sample-Size Estimation

One of the key elements in conducting an ethical clinical trial is the principle of beneficence. This principle requires that the study design is scientifically sound and that any risks of the research are acceptable in relation to the likely benefits from the study. In the context of our present discussions, the phrase "requires that the study design is scientifically sound" is particularly pertinent. Research subjects voluntarily take part in clinical trials not for their

personal gain but for the greater good. They are told that their participation will provide information that is useful and generalizable to a much larger group of people. This is one of the benefits that are weighed against the risks of being exposed to a drug under development. If the design of the trial is such that the data collected do not permit the best possible information to be obtained, the subjects' expectations have been violated.

Sample-size estimation therefore has an important ethical component. There are ethical issues involved in recruiting both too few and too many subjects (Matthews, 2006). Recruiting too few subjects means that the study may be underpowered and unable to detect a treatment effect of interest that actually exists. Such a design is scientifically inadequate to answer the research question of interest (i.e., to address the primary objective of the trial). It is also unethical. Subjects may have taken part in a study that did not have a chance of detecting a treatment effect that may have existed, and thus their expectation that participation may add to the knowledge base about the investigational drug was violated.

It is also unethical to recruit many more subjects than were actually necessary to provide an answer to the research question. Imagine a trial testing a new drug against a placebo in which an answer to the research question could have been obtained with 1,000 subjects (500 in each treatment group). That is, after 1,000 subjects had participated, there could have been compelling evidence that the new drug was safe and more effective than the placebo. If 2,000 subjects actually took part in the trial, one-half of the second thousand subjects would have received the placebo. That is, 500 subjects would have been given a treatment that was inferior. Asking subjects to participate in clinical trials in which a new drug is evaluated against a control drug is only ethical if there is no existing evidence at that time that the new drug is more effective (i.e., clinical equipoise).

Sample-size estimation therefore takes on a special significance in clinical trials. As noted in Section 9.1, this process of estimation does not produce the "right" answer, so it is not possible to specify precisely what constitutes "too few" or "too many" subjects. However, it is imperative to estimate a reasonable sample size based on the best evidence that is available at the time and with full knowledge of the implications of this estimate.

9.3. Variables involved in Sample-Size Estimation

Several variables need to be considered in the process of sample-size estimation. The values of these variables in any given case can be chosen by the sponsor based on several considerations. Some terms that will be useful for present discussions are:

- Type I errors and Type II errors. A Type I error occurs when a significant result is "found" when it does not really exist, and a Type II error occurs when one fails to find a significant difference that actually exists.
- The probability of making a Type I error, α . This is also the level of statistical significance chosen, typically 0.05, but it is possible to choose 0.01 or even more conservative values.
- The probability of making a Type II error, β . A probability value must be between 0 and 1: therefore, β will be between 0 and 1.
- Power, calculated as 1 minus β . Since the probability represented by β will be between 0 and 1, power will also be between 0 and 1 since it is defined as 1 minus β . In the context of our ongoing example of developing a new antihypertensive, the power of a trial describes its ability to find a difference between treatment groups when such a difference actually exists. The power of a statistical test is the probability that the null hypothesis is rejected when it is indeed false. Since rejecting the null hypothesis when it is false is extremely

desirable, it is generally regarded that the power of a study should be as great as practically feasible.

- An estimation of the treatment effect. This is usually the difference between means or proportions. In this book's ongoing example involving a new investigative antihypertensive drug, the treatment effect is the difference between the mean drug treatment group SBP (systolic blood pressure) response and the mean placebo treatment group SBP response.
- An estimation of the variance in the treatment effect (the standard deviation is typically used here).
- The standardized treatment effect, calculated by dividing the estimated treatment effect by its estimated standard deviation.
- The sample size, N, that is provided by the calculation performed using the values chosen by the researcher.

9.4. Type I and Type II Errors

In the context of this book's ongoing example involving a new investigative antihypertensive drug, a Type I error occurs when a significant difference between treatment groups is found when it does not really exist. This occurrence is also known as a false-positive finding. A Type II error occurs when the sponsor fails to find a significant difference that does actually exist, an occurrence also known as a false negative. (It should be noted here that different types of study designs, such as equivalence and noninferiority designs, require different types of null hypotheses and different expressions of Type I and Type II errors. They also require different formulas for sample-size estimation. This chapter's discussions address the book's ongoing example involving the development of a new antihypertensive drug.)

Setting β at 0.10 means that the sponsor is willing to accept a 10% chance of missing an association of a given treatment effect size (the treatment effect size chosen by the sponsor). That is, the sponsor is willing to accept a 10% chance of a Type II error occurring. Put the other way, this means that there is a 90% chance of finding a treatment effect of the magnitude chosen (or greater) for the sample-size estimation. Thus, in 9 out of 10 studies (90%), the investigator would likely be able to correctly reject the null hypothesis (given that the assumed standard deviation is correct).

Table 9.1 provides a concise picture of the implications of false-positive findings and false-negative findings. One of two actions – rejecting the null hypothesis or failing to reject the null hypothesis – must occur at the end of all hypothesis testing, and the action taken is determined by the significance of the test statistic obtained in the statistical analysis conducted. If the test statistic attains statistical significance, the sponsor rejects the null hypothesis; if the test statistic does not attain statistical significance, the sponsor fails to reject the null hypothesis.

Table 9.1. Type 1 Errors and Type 11 Errors

	Reality	
Action Based on Study Results	Research Hypothesis is True	Null Hypothesis is True
Reject null hypothesis	Correct action	Type I error (false positive)
Fail to reject null hypothesis	Type II error (false negative)	Correct action

9.4.1. The Implications of Type I and Type II Errors

Table 9.1 shows that the results from a clinical trial can lead the sponsor to an inappropriate conclusion in some cases. In these cases, one of two types of error occurs:

- Type I error (false-positive): In this scenario, the sponsor rejects the null hypothesis, e.g., “finds” a statistically significant difference between the drug treatment group mean response and the placebo treatment group mean response in the type of study used in our ongoing example of an antihypertensive drug. The inference from this finding, based on the sample of subjects employed in this trial, is that the drug would be effective in the population from which the sample was chosen.
- Type II error (false-negative): In this case, the sponsor fails to reject the null hypothesis, i.e., fails to find a statistically significant difference between the drug treatment group mean response and the placebo treatment group mean response. The inference from this finding, based on the sample of subjects employed in this trial, is that the drug would not be effective in the population from which the sample was chosen.

Ideally, the likelihood of either type of error would be zero, or at least as low as possible. In reality, the possibility of making these errors cannot totally be eliminated, but their likely occurrence can be balanced one against the other. This is done by choosing various combinations of α and β , and therefore various combinations of α and power, since power is defined as 1 minus β .

9.5. Choosing the Variables Needed for Sample-Size Estimation

As noted in Section 9.1, several variables are needed for sample-size estimation, and the researcher can choose the values to be used in the formula that will yield the sample size, N . These are α , β , the estimated treatment effect, and its variance.

9.5.1. Alpha and Beta

The estimation of a sample size requires several variables to be chosen. In each case, the following need to be selected:

- The significance level (α), usually 0.05. Choosing $\alpha = 0.05$ means that on 95% of occasions the null hypothesis will be rejected correctly. That is, the researcher is willing to accept a 5% chance that a positive finding will result by chance alone. Generally, regulatory agencies are concerned about Type I (false-positive) errors: They do not want to grant marketing approval on the basis of erroneously favorable data. This occurrence is made acceptably low by typically choosing $\alpha = 0.05$ and sometimes choosing $\alpha = 0.01$ to be really conservative.
- Adequate power. In the context of our ongoing example, power is the ability of a study to find a difference between treatment groups when such a difference actually exists. Power is calculated as 1 minus β . In most clinical trials, adequate power is regarded as at least 80%, and it is typically 90%. In contrast to regulatory agencies' concern with Type I errors, sponsors are generally concerned about Type II (false-negative) errors: They do not want to erroneously conclude that their drug does not work by obtaining a nonsignificant result when the drug does actually work. Doing this will likely result in the drug not being brought to market when it might have been. So, sponsors want enough power to detect a real difference when it exists, i.e., to reject the null hypothesis when it should be rejected. So, ideally, they probably want (at least) 90% power. Selecting 90% power sets β at 0.10, since power equals 1 minus β .

9.5.2. The Treatment Effect, Its Variance, and the Standardized Treatment Effect

For the ongoing example in this book, i.e., testing a new antihypertensive drug against a placebo control, the following values must also be chosen:

- The clinically relevant difference (CRD) that the test is required to detect. This is the treatment effect size, i.e., the difference between the mean drug treatment group response and the mean placebo treatment group response that the sponsor deems clinically relevant.
- The standard deviation (SD) of the treatment effect.
- The standardized effect size, calculated as the ratio CRD/SD.

Determining the clinically relevant difference to look for in the study is relatively straightforward. Another way to conceptualize the clinically relevant difference is as the smallest effect size that is clinically meaningful. This can be based on clinical input. For example, a decrease in SBP of 10 mmHg may be thought by the sponsor to be clinically relevant in this context.

The standard deviation for change in SBP can be harder to determine. One way is to examine previously published data on similar outcomes (maybe other drugs in the same class). If few data are available from this source, consulting with experts in this research domain may be helpful. Another possibility is to conduct a small pilot study. In the later phases of a clinical development program, data from earlier studies may be informative. This often means that for confirmatory clinical drug trials there will be results from earlier trials, so information is readily available. From these two items, the standardized effect size can be calculated as the ratio CRD/SD.

9.6. Using the Appropriate Formula to Yield the Sample Size

Sample-size estimation can be performed for any study design. In each case, the respective formula will be used to estimate the sample size required (see Chow et al., 2003). For the formula used in the type of study design that we are using as our ongoing example, each of the variables we have discussed will have certain influences on the sample size, N , that will be given by the formula. These influences, i.e., their relationships with N given that all of the others remain the same, can be summarized as follows:

- The smaller the chosen value of α , the larger the value of N that will be given.
- The smaller the chosen value of β , the larger the value of N that will be given. This is because power is defined as 1 minus β . As β decreases, power increases; as power increases, the larger the value of N that will be given.
- The larger the standardized effect size, the smaller the value of N that will be given.

The third of these relationships, the relationship between standardized effect size and N , is actually influenced by the two separate factors that determine the standardized effect size once each of these factors has been chosen. As noted in Section 9.5.2, the standardized effect size is calculated as the ratio CRD/SD. Since the CRD is the numerator in this ratio, the larger the CRD, the larger the standardized effect size will be for a given SD. And, conversely, since SD is the denominator in the ratio, the larger the SD, the smaller the standardized effect size will be for a given CRD. Therefore, the larger the SD, the larger the N that will be given by the sample-size estimation.

9.7. Influences on the Sponsor's Choice of These Values

As we have discussed, when conducting a sample-size estimation, the researcher has to choose values for α and β and has to come up with a standardized treatment effect, which in turn is the result of finding the best possible estimates of a clinically significant difference and its variation. What are the influences that lead the sponsor to choose certain values for α , β , and the standardized treatment effect?

For financial, time demand, and logistical reasons, a smaller sample size is preferable to a sponsor than a larger one. There are also ethical factors that need to be borne in mind. It is unethical to choose a sample that can reasonably be considered either too small or too large (see Section 9.2). The optimum sample size can be considered to be the smallest sample size that can reasonably be expected to answer the primary research question, i.e., evaluating the primary objective as stated in the study protocol.

What might influence the sponsor's choice of α and β ? It was noted in Section 9.5.1 that a typical value for α is 0.05 and a typical value for β is 0.10 (i.e., a typical power is 90%). Circumstances in which it may make sense to choose different values include the following:

- For a drug with nasty side effects, it will be necessary to have particularly compelling evidence that it is effective. That is, it will be necessary to demonstrate highly statistically significant efficacy, and there is a strong need to avoid false-positive results, i.e., to avoid a Type I error. Therefore, the sponsor will likely set α lower than usual, perhaps at 0.01 or even 0.001. This choice of α being set at lower than the typical 0.05 will result in a greater N being given by the sample-size estimation.
- If a study is particularly difficult to repeat, and therefore the trial being planned is the sponsor's "one shot" at getting relevant data, it is a good idea to give the study as much power as is practically possible. So, the sponsor may increase the study's power to a higher value than usual. Since power is calculated as 1 minus β , the sponsor needs to reduce β in order to increase the study's power, which means that the chance of a Type II error, i.e., not finding a treatment effect that actually exists, is reduced. This choice of β as lower than the typical 0.10 will result in a greater N being given by the sample-size estimation.
- If the sponsor is conducting a study early in product development and wishes to optimize power on that occasion while "compromising" α , β might be chosen as 0.10 and α as 0.10 or even 0.15 or 0.20 (Donahue and Ruberg, 1997).

While either action, i.e., reducing α or β , will increase the value of N given by the sample-size estimation and therefore result in additional cost to the sponsor, the sponsor may well decide that, in the overall balancing act of estimating sample size, there are good reasons to do this in cases such as these examples.

At each stage of the drug development program and for each trial within that stage, sponsors need to be aware of the implications of their choice of α and their choice of β and the acceptability of these implications. The implications of each choice and the acceptability of these implications may change throughout the course of a clinical development program.

9.8. Choosing the Objective(s) on Which to Base the Sample-Size Estimation

A sample-size estimation must be based on a specific objective in a clinical trial's study protocol. By the time sample-size estimation becomes particularly meaningful, i.e., in later-stage clinical trials designed to demonstrate efficacy, it is a very good idea to have a single objective (the primary objective) and a single identifiable endpoint or outcome of interest. In this case, the sample-size estimation is based on this objective. However, this situation is

not always the case, and more than one outcome measure is regarded as equally important by the researcher. In these situations, a common approach is to conduct the sample-size estimates for each outcome measure and then select the largest of these as the sample size required to answer all the questions of interest (Machin and Campbell, 2005).

This approach, however, raises issues of multiplicity. Accordingly, lower p-values may be required to be able to declare a result as statistically significant. This means that an adjustment to the sample-size estimation formula is appropriate, with the precise nature of the adjustment being related to the number of outcomes to be tested. This adjustment raises the magnitude of the estimated sample size (Machin and Campbell, 2005).

9.9. Other Issues to Keep in Mind

It is useful to keep several other issues in mind when conducting sample-size estimations, including the following:

- Possible attrition. It is likely that all of the subjects that start a clinical trial will not complete it. This attrition may increase with the demands of a trial (e.g., number of clinic visits required, or degree of discomfort caused by any procedures or measurements). It is important to consider the possible (likely) attrition rate when estimating the number of subjects needed for the "successful" analysis of the data, and to increase the number chosen appropriately.
- Overly optimistic choice of treatment effect. As Machin and Campbell (2005) noted in the context of comparative clinical trials, researchers "are often optimistic about the magnitude of the improvement of the new treatments over the standard." Since a larger estimated treatment effect leads to a smaller sample size being chosen, overestimating the estimated treatment effect may lead to a smaller but still clinically important effect not being detected, since the sample size adopted was too small to detect it.
- In some instances (e.g., late-stage development) the sample size required is driven not by statistical considerations for demonstrating effect but by minimal exposure requirements for safety considerations (see ICH Guideline E1).

References

Chow, S-C., Shao, J., and Wang, H., *Sample size calculation in clinical research*, CRC/Taylor Francis.

Donahue, R.M.J. and Ruberg, S.J., 1997, Standardizing clinical study designs for accelerating drug development, *Drug Information Journal*, 31:655-663.

Machin, D. and Campbell, M.J., 2005, *Design of studies for medical research*, John Wiley & Sons.

Matthews, J.N.S., 2006, *Introduction to randomized controlled clinical trials*, 2nd Edition, Chapman & Hall/CRC.

Pallay, A., 2000, A decision analytic approach to determining sample sizes in a Phase III program, *Drug Information Journal*, 34:365-377.