Chapter 16
Sample Size Calculation

Eileen M. Stock and Kousick Biswas

Introduction

Every clinical trial should be planned in advanced. This plan should include the study's objectives, primary and secondary endpoints, data collection, inclusion and exclusion criteria, required sample size with scientific justification, statistical methodology, and an approach to handle missing data [1]. A sample size calculation is used to determine the minimum number of participants needed in a clinical trial in order to be able to answer the research question under investigation. During the planning phase of a clinical trial, sample size estimation should be one of the very first and key components to consider in the design of a study. Knowing the anticipated sample size allows investigators to determine whether a study is feasible and to develop an appropriate budget and identify needed resources to carry out the study. The calculation of sample size with a sufficient level of significance and power is essential to the success of a trial.

Requirements for Sample Size Calculation

The estimation of sample size involves the consideration of multiple components, including the study's objective and primary hypothesis, type of endpoint to be analyzed, expected treatment effect and variability, treatment allocation ratio if it is

E.M. Stock (✉) · K. Biswas
Cooperative Studies Program Coordinating Center,
Office of Research and Development, U.S. Department of Veterans Affairs,
VA Medical Center, 5th Boiler Street, Perry Point, MD 21902, USA
e-mail: Eileen.Stock@va.gov

K.M.F. Itani and D.J. Reda (eds.), *Clinical Trials Design in Operative and Non Operative Invasive Procedures*, DOI 10.1007/978-3-319-53877-8_16

desirable to have more randomized to one group than another, anticipated recruitment rate, and the estimated number of dropouts. Other parameters influencing sample size calculation include types of error (I and II) and power [1, 2].

Types of Error and Power

Consider the multisite randomized clinical trial comparing operative and nonoperative treatment using accelerated functional rehabilitation for acute Achilles tendon ruptures [3]. For the primary outcome of rerupture, the *null hypothesis*, denoted H_0, would be that there exists no difference between the two population proportions of rerupture. That is, there is no difference in the rate of rerupture between those with acute Achilles tendon rupture undergoing surgical repair and those treated nonoperatively. The *alternative hypothesis* (for a two-sided test; typically denoted H_a) is that there is a difference in the rate of rerupture. A *Type I error*, commonly referred to as *significance level* and denoted as α, is defined as the probability of erroneously rejecting the null hypothesis when it is in fact true. In this example, a Type I error would be concluding a difference in the rate of rerupture between treatment procedures that is unlikely to actually exist, i.e., a false positive. A *Type II error*, denoted as β, is the probability of failing to reject a false null hypothesis. That is, erroneously missing an actual difference in rerupture rates between treatment procedures, a false negative. *Power* (equal to $1 - \beta$) is the probability of rejecting the null hypothesis when it is false and should be rejected (Table 16.1) [1, 2].

Study's Primary Hypothesis

The primary purpose of a clinical trial, written as a scientific hypothesis, guides the design of the trial. Traditionally, a two-arm parallel-group design is employed to look for a difference between treatments (two-sided). Two-sided *p-values* provide the probability that the results are compatible with the null hypothesis (H_0 true).

Table 16.1 Summary of type I and II errors

	True state	
Statistical decision	H_0 true (No treatment benefit) Should fail to reject H_0	H_0 false (Treatment benefit) Should reject H_0
Fail to reject H_0 (No treatment benefit)	Correct decision	Type II error (β)
Reject H_0 (Treatment benefit)	Type I error (α)	Correct decision, power ($1 - \beta$)

When the p-value is small (say, *p-value* <0.05), the null hypothesis is rejected (reject H_0) and there is evidence to support a difference in treatment effects. The direction of the test statistic establishes whether the new treatment is superior or inferior to the control treatment. In some instances, there is no interest in rejecting a null hypothesis in both directions (i.e., there is no interest in an inferiority results) and a superiority trial may be preferred to examine whether a new treatment is superior (better) than the established alternative (one-sided) [4].

While the traditional approach is intended to determine whether there is a difference between the experimental treatment and control, this may not be the relevant approach when the control is known to be effective and it is hoped that the experimental treatment can be shown to be as effective. In this instance, it is usually the case, that the experimental treatment may offer other advantages to the control treatment, such as convenience or tolerability, if it can be shown to be as effective to the control. Equivalence trials are designed to establish that the new procedure cannot be worse nor better than the conventional procedure if the null hypothesis is rejected. It requires that the two treatment approaches be identical within some acceptable range, δ (normally $\pm 20\%$) [5]. Lastly, for a non-inferiority trial, the aim is to show that the new treatment is as good as or better (no worse) than the established treatment [4]. Each of the mentioned designs will be selected according to the study's primary hypothesis and rely on prior information about the effects of the new procedure on a specific endpoint [1].

Study Design Considerations

Various study designs, such as a parallel-group, crossover, factorial, or cluster, may be employed to address a study's objectives and ensure the required sample size is achieved. Each design will vary in their approach for sample size calculation. In the case of rare events, the need for a multisite trial is higher.

Study Endpoint Expected Response

A study's endpoint, whether continuous, dichotomous, or time-to-event, will govern the type of model and sample size calculation. In the case of multiple comparisons, an adjustment to the significance level may be necessary. For a continuous endpoint, information on the expected central tendency (mean score) and variability (standard deviation) of the new procedure and its comparator are needed to more precisely estimate the sample size. The greater the variation within groups or the smaller the expected difference between groups, the larger the sample

size will need to be in order to produce the same result. For a dichotomous variable, the proportion of participants achieving success in each group is needed. Most importantly, the expected treatment effect, as compared to its comparator, should be clinically meaningful [1].

Participant Retention Rates and Treatment Allocation

While sample size calculations determine the required number of participants for specific analyses, other aspects of recruitment should also be considered such as screen-failures, dropouts, and patients who are lost to follow-up. A trial should enroll more subjects to account for potential dropouts and those lost to follow-up. Attrition rates can vary tremendously, where $\leq 5\%$ is of little concern but $\geq 20\%$ poses serious threats to the validity of the trial [6]. Most RCTs (60–89%) published in leading journals have missing endpoint data, with complete case analysis the most frequently used strategy for handling this missing data [7, 8]. For many of these trials (18%), dropout rates exceeded 20% [8, 9]. For this reason, the number of enrollments, in trials where the primary outcome measure is continuous or binary, can be determined using an adjustment to the sample size and estimated dropout rate in the formula, $Enrollment = Sample\ Size/(1 - dropout\ rate)$ [1]. For time-to-event, or survival data, the adjustment for dropout rate is more involved. In some instances, interim analyses may be requested to monitor treatment effects and ensure enrollment follows a specific trajectory [10, 11].

If one treatment arm is anticipated to have a greater dropout rate than its comparator, an unequal treatment allocation may be employed to ensure a balanced distribution at the end of the trial. Additionally, varied allocation and enrollment can occur in cases where it is unethical to assign an equal number of patients to each arm (e.g., placebo or sham treatment) [1]. Thus, sample size is adjusted in these scenarios. Note that departures form 1:1 randomization will increase the sample size requirement.

Conventional Guidelines

In sample size calculations, the level of significance (α) for a study is typically assumed to be 0.05 (or 5%) [12]. However, 1% or less may be used for larger samples and 10% for smaller samples. Also, the minimum power for which sample size is calculated is 80%. Larger power may be used to estimate sample size in order to provide a more conservative estimate in case treatment effects or recruitment are less than anticipated.

Calculation of Sample Size

There are many approaches to sample size estimation, with some of the more common calculations involving the comparison of two means, proportions, or a time-to-event measure and testing for a difference between groups. The next few sections describe these in more detail.

Comparing Two Means

The formula for calculating sample size comparing the mean of two treatment arms is

$$n_1 = \kappa n_2; \quad n_2 = \left(1 + \frac{1}{\kappa}\right)\left[\frac{\left(z_{\alpha/2} + z_\beta\right)^2}{d^2}\right] = \left(1 + \frac{1}{\kappa}\right)\left[\frac{\left(z_{\alpha/2} + z_\beta\right)^2\left(\sigma_1^2 + \sigma_2^2\right)}{2\left(\bar{\mu}_1 - \bar{\mu}_2\right)^2}\right],$$

where $z_{\alpha/2}$ is the critical value of the standard normal distribution at $\alpha/2$ (e.g., 1.96 for a 95% confidence interval with Type I error $\alpha = 0.05$), z_β is the critical value of the standard normal distribution at β (e.g., 0.84 for 80% power and Type II error $\beta = 20\%$), κ is the matching ratio, μ_i is the population mean of the endpoint in group i, σ_i^2 is the population variance of the endpoint in group i, and d is Cohen's effect size [13]. For studies with 1:1 randomization, $\kappa = 1$.

Comparing Two Proportions

The formula for calculating sample size comparing two proportions is

$$n_1 = \kappa n_2; \quad n_2 = \left[\frac{p_1(1 - p_1)}{\kappa} + p_2(1 - p_2)\right]\left(\frac{z_{\alpha/2} + z_\beta}{p_1 - p_2}\right)^2,$$

where p_i is the population proportion of group i, and $p_1 - p_2$ is the effect size or difference desired to be detected [13].

Comparing Time-to-Event

The formula for calculating sample size for a time-to-event analysis (Cox proportional hazards model) is

$$n = \frac{1}{p_1 p_2 p_A}\left(\frac{z_{\alpha/2} + z_\beta}{\ln(\theta) - \ln(\theta_0)}\right)^2,$$

where p_i is the proportion with the event in group i, p_A is the overall event rate, θ is the hazard rate, θ_0 is the hypothesized hazard rate under the null hypothesis, and $\ln(\theta) - \ln(\theta_0)$ the regression coefficient (treatment indication) [13, 14]. Note that sample size formulae accounting for the length of the recruitment and follow-up periods, and drop-outs, are more sophisticated.

Available Software

Statistical software packages with tools for sample size and power analysis calculations include SAS (SAS Institute, Inc.; Cary, NC), G*Power (Faul, Erdfelder, Lang, & Buchner, 2007), PASS (NCSS, LLC.; Kaysville, Utah), R (The R Foundation for Statistical Computing; Auckland, New Zealand), M*plus* (Muthén & Muthén; Los Angeles, CA), and PS available online at Vanderbilt University (Dupont & Plummer, 1990) [15]. Several of these packages are available at no cost.

Common Pitfalls Related to and Affecting Sample Size

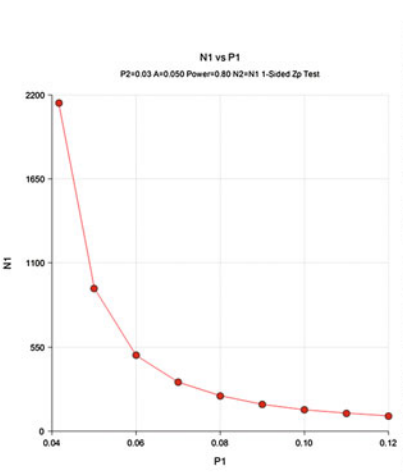
Sample size calculations pose several challenges, including obtaining an accurate estimate of treatment effects, selecting an appropriate power and significance level, and even selecting the correct formula to be used [16]. As a result, sample size underestimation or overestimation may occur.

Sample Size Underestimation

Sample size underestimation refers to a sample size for a trial that was calculated to be less than that required [16]. This results in lower power than is needed and may lead to misleading results such as the determination of no treatment effect (*p-value* > α) when one really existed. The treatment effect was not statistically significant even though it was clinically significant. That is, recruiting too few participants can lead to inconclusive results because of the low likelihood of finding a clinically relevant difference statistically significant.

Revisiting the Achilles Tendon Rupture trial, small sample size was a limitation of the study (72 participants per each arm), and therefore was underpowered to

definitively make a conclusion about rerupture rates [3]. A meta-analysis had shown rerupture rates to be approximately 2.8% following operative repair and 11.7% for nonoperative treatment [17]. ConsequentlyH, the Rupture trial underestimated the sample size required. Instead, rates of 2.8% and 4.2% for operative and nonoperative treatment, respectively, were observed. The former would require a sample size of 104 participants in each group using a one-sided 2-sample independent proportions test, assuming a significance level of $\alpha = 0.05$. The latter would require 2148 participants per arm (Fig. 16.1). Although the actual power for comparing rerupture rates was 12% (Fig. 16.2), with a Type II error of 88%, this study was the largest to date of its kind and findings would provide clinical insight and pilot data should a larger trial be pursued.



Tests for Two Proportions

Numeric Results for Testing Two Proportions using the Z-Test with Pooled Variance
H0: P1 - P2 ≤ 0 vs. H1: P1 - P2 = D1 > 0.

Target Power	Actual Power*	N1	N2	N	P1	P2	Diff D1	Alpha
0.80	0.80013	2148	2148	4296	0.0417	0.0278	0.0139	0.0500
0.80	0.80035	936	936	1872	0.0500	0.0278	0.0222	0.0500
0.80	0.80015	499	499	998	0.0600	0.0278	0.0322	0.0500
0.80	0.80053	322	322	644	0.0700	0.0278	0.0422	0.0500
0.80	0.80130	231	231	462	0.0800	0.0278	0.0522	0.0500
0.80	0.80006	176	176	352	0.0900	0.0278	0.0622	0.0500
0.80	0.80065	141	141	282	0.1000	0.0278	0.0722	0.0500
0.80	0.80217	117	117	234	0.1100	0.0278	0.0822	0.0500
0.80	0.80198	99	99	198	0.1200	0.0278	0.0922	0.0500

* Power was computed using the normal approximation method.

References
Chow, S.C., Shao, J., and Wang, H. 2008. Sample Size Calculations in Clinical Research, Second Edition. Chapman & Hall/CRC. Boca Raton, Florida.
D'Agostino, R.B., Chase, W., and Belanger, A. 1988.'The Appropriateness of Some Common Procedures for Testing the Equality of Two Independent Binomial Populations', The American Statistician, August 1988, Volume 42 Number 3, pages 198-202.
Fleiss, J. L., Levin, B., and Paik, M.C. 2003. Statistical Methods for Rates and Proportions. Third Edition. John Wiley & Sons. New York.
Lachin, John M. 2000. Biostatistical Methods. John Wiley & Sons. New York.
Machin, D., Campbell, M., Fayers, P., and Pinol, A. 1997. Sample Size Tables for Clinical Studies, 2nd Edition. Blackwell Science. Malden, Mass.
Ryan, Thomas P. 2013. Sample Size Determination and Power. John Wiley & Sons. Hoboken, New Jersey.

Report Definitions
Target Power is the desired power value (or values) entered in the procedure. Power is the probability of rejecting a false null hypothesis.
Actual Power is the power obtained in this scenario. Because N1 and N2 are discrete, this value is often (slightly) larger than the target power.
N1 and N2 are the number of items sampled from each population.
N is the total sample size, N1 + N2.
P1 is the proportion for Group 1 at which power and sample size calculations are made. This is the treatment or experimental group.
P2 is the proportion for Group 2. This is the standard, reference, or control group.
D1 is the difference P1 - P2 assumed for power and sample size calculations.
Alpha is the probability of rejecting a true null hypothesis.

Summary Statements
Group sample sizes of 2148 in group 1 and 2148 in group 2 achieve 80.013% power to detect a difference between the group proportions of 0.0139. The proportion in group 1 (the treatment group) is assumed to be 0.0278 under the null hypothesis and 0.0417 under the alternative hypothesis. The proportion in group 2 (the control group) is 0.0278. The test statistic used is the one-sided Z-Test with pooled variance. The significance level of the test is 0.0500.

Fig. 16.1 Sample size estimation for comparing rerupture rates, varying rates in the nonoperative group [created through the use of: PASS 14 Power Analysis and Sample Size Software (2015). NCSS, LLC. Kaysville, Utah, USA, ncss.com/software/pass]

```
> power.prop.test(n=72, p1=(2/72), p2=(3/72), sig.level=0.05, power= , alternative="one.sided")

     Two-sample comparison of proportions power calculation

              n = 72
             p1 = 0.02777778
             p2 = 0.04166667
      sig.level = 0.05
          power = 0.1169202
    alternative = one.sided

NOTE: n is number in *each* group
```

Fig. 16.2 Power analysis for observed difference in rerupture rates [created through the use of: R (The R Foundation for Statistical Computing; Auckland, New Zealand)]

Sample Size Overestimation

On the contrary, a sample size selected to be much larger than was required describes *sample size overestimation* [16]. Studies that are too large are also problematic for at least two reasons. This scenario may be evident from an exceptionally strong statistical significance (very small *p-value*), which raises ethical concerns if more subjects were exposed to an inferior treatment than were required or resources wasted. Additionally, for larger sample sizes, smaller differences can be detected and be statistically significant even when the difference is not clinically meaningful. In trial design, each assumption may be made too conservatively, to avoid the risk of failure, and the analysis of the study's primary objective becomes overpowered as a result.

Selecting a Clinically Meaningful Difference

Determining the clinically meaningful difference for which a study is powered to detect is generally the most difficult task of the sample size calculation process. A very thorough literature search should be conducted to obtain any available data on the potential effect of the proposed new treatment. This may include published abstracts, results of phase II trials or pilot studies, and subgroup analyses from a previously conducted trial. If enough publications are available, meta-analysis techniques can be used to obtain an estimate of the potential treatment effect.

Often data are limited to help inform the potential treatment effect estimate. In those cases, an investigator might look to other published studies in this area to determine the magnitude of effect that was used when that study was designed. Often, FDA has determined the degree of treatment effect needed to establish efficacy and their guidelines may be useful as a resource. Additionally, a panel of experts in the area of investigation can be convened to develop a consensus estimate of treatment effect.

Available Databases

There are multiple databases available for use in obtaining estimates for sample size calculations. In 1994, the VA established a VA National Surgical Quality Improvement Program (NSQIP) in which all medical centers performing major surgery participated [18]. The database contains 135 variables collected preoperatively and up to 30 days postoperatively. Data is categorized as demographic, surgical profile, preoperative, intraoperative, or postoperative. Each hospital submits an average of 1,600 major operations per year into the database [19]. While the aim of NSQIP was initially quality improvement in surgical care through periodic

reports and assessments of performance, VA investigators can also query the database for scientific research purposes and to obtain estimates of event rates for a power analysis such as mortality, cardiac and noncardiac complications, postoperative pneumonia, intubations, pulmonary embolism and venous thrombosis, renal dysfunction, and infections. Similarly, the American College of Surgeons National Surgical Quality Improvement Program (ACS NSQIP) can be used for sample size estimation as in the comparison of postoperative complication rates for regional versus general anesthesia among surgical patients with chronic obstructive pulmonary disease [20, 21]. Other useful available databases include the Society of Thoracic Surgeons (STS) National Database including separate databases for adult cardiac, general thoracic, and congenital heart surgery [22], and the Centers for Disease Control (CDC) Cancer Registry [23].

References

1. Sakpal TV. Sample size estimation in clinical trial. Perspect Clin Res. 2010;1:67–9.
2. Jones SR, Carley S, Harrison M. An introduction to power and sample size estimation. Emerg Med J. 2003;20:453–8.
3. Willits K, Amendola A, Bryant D, et al. Operative versus nonoperative treatment of acute Achilles tendon ruptures: a multicenter randomized trial using accelerated functional rehabilitation. J Bone Joint Surg Am. 2010;92:2767–75.
4. Christensen E. Methodology of superiority vs. equivalence trials and non-inferiority trials. J Hepatol. 2007;46:947–54.
5. Steinijans V, Hauschke D. International harmonization of regulatory bioequivalence requirements. Clin Res Regul Aff. 1993;10.
6. Fewtrell MS, Kennedy K, Singhal A, et al. How much loss to follow-up is acceptable in long-term randomised trials and prospective studies? Arch Dis Child. 2008;93:458–61.
7. Akl EA, Briel M, You JJ, et al. LOST to follow-up information in trials (LOST-IT): a protocol on the potential impact. Trials. 2009;10:40.
8. Wood AM, White IR, Thompson SG. Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals. Clin Trials. 2004;1:368–76.
9. Bell ML, Kenward MG, Fairclough DL, Horton NJ. Differential dropout and bias in randomised controlled trials: when it matters and when it may not. BMJ. 2013;346:e8668.
10. Floriani I, Rotmensz N, Albertazzi E, et al. Approaches to interim analysis of cancer randomised clinical trials with time to event endpoints: a survey from the Italian National Monitoring Centre for Clinical Trials. Trials. 2008;9:46.
11. Broglio KR, Stivers DN, Berry DA. Predicting clinical trial results based on announcements of interim analyses. Trials. 2014;15:73.
12. Kadam P, Bhalerao S. Sample size calculation. Int J Ayurveda Res. 2010;1:55–7.
13. Chow S, Shao J, Wang H, editors. Sample size calculations in clinical research. 2nd ed. Boca Raton: Chapman & Hall/CRC; 2008.
14. Wang H, Chow S. Sample size calculation for comparing time-to-event data. In: D'Agostino R, Sullivan L, Massaro J, editors. Wiley encyclopedia of clinical trials. New York: Wiley; 2007.
15. Dupont WD, Plummer WD. Power and sample size calculations: a review and computer program. Control Clin Trials. 1990;11:116–28.

16. Noordzij M, Tripepi G, Dekker FW, Zoccali C, Tanck MW, Jager KJ. Sample size calculations: basic principles and common pitfalls. Nephrol Dial Transplant. 2010;25:1388–93.
17. Lo IK, Kirkley A, Nonweiler B, Kumbhare DA. Operative versus nonoperative treatment of acute Achilles tendon ruptures: a quantitative review. Clin J Sport Med. 1997;7:207–11.
18. Khuri SF, Daley J, Henderson WG. The comparative assessment and improvement of quality of surgical care in the Department of Veterans Affairs. Arch Surg. 2002;137:20–7.
19. Fuchshuber PR, Greif W, Tidwell CR, et al. The power of the National Surgical Quality Improvement Program–achieving a zero pneumonia rate in general surgery patients. Perm J. 2012;16:39–45.
20. ACS National Surgical Quality Improvement Program (ACS NSQIP). Participant Use Data Files. Available from https://www.facs.org/quality-programs/acs-nsqip. American College of Surgeons.
21. Hausman MS Jr, Jewell ES, Engoren M. Regional versus general anesthesia in surgical patients with chronic obstructive pulmonary disease: does avoiding general anesthesia reduce the risk of postoperative complications? Anesth Analg. 2015;120:1405–12.
22. The Society of Thoracic Surgeons. STS National Database. Available from http://www.sts.org/national-database.
23. Centers for Disease Control and Preventions. National Program of Cancer Registries (NPCR). Available from http://www.cdc.gov/cancer/npcr/.