Chapter 9. Study Size Planning

Eric S. Johnson, Ph.D., M.P.H.
Kaiser Permanente Northwest, Portland, OR

M. Alan Brookhart, Ph.D.
University of North Carolina at Chapel Hill Gillings School of Global Public Health, Chapel Hill, NC

Jessica A. Myers, Ph.D.
Harvard Medical School and Brigham and Women's Hospital, Boston, MA

Abstract

The feasibility of a study often rests on whether the projected number of accrued patients is adequate to address the scientific aims of the study. Accordingly, a rationale for the planned study size should be provided in observational comparative effectiveness research (CER) study protocols. This chapter provides an overview of study size and power calculations in randomized controlled trials (RCTs), specifies considerations for observational comparative effectiveness research (CER) study size planning, and highlights study size considerations that differ between RCTs and observational studies of comparative effectiveness. The chapter concludes with a checklist of key considerations for study size planning for a CER protocol.

Introduction

An important aspect of the assessment of study feasibility is whether the projected number of accrued patients is adequate to reasonably address the scientific aims of the study. Many journals have endorsed reporting standards that ask investigators to report the rationale for the study size. For example, the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) checklist asks investigators to report their rationale, which may include a statistical power calculation. However, such a rationale is often missing from study protocols. This is problematic when investigators interpret study findings in terms of the statistical significance in relation to the null hypothesis, which implies both a prespecified hypothesis and adequate statistical power (e.g., ≥80% for detecting a clinically important increase in harm). Without the context of a numeric rationale for the study size, readers may misinterpret the lack of a statistically significant difference in effect as false reassurance of lack of harm, or falsely conclude that there is no benefit when comparing two interventions.

Study Size and Power Calculations in RCTs

The study planning needed to achieve various study sizes and an understanding of statistical power that a given study size can yield are important aspects in the design of randomized controlled trials (RCTs). Reporting on the rationale underlying the size of treatment arms is clearly specified in the Consolidated Standards of Reporting Trials (CONSORT) and STrengthening the Reporting of OBservational studies in Epidemiology (STROBE) reporting guidelines, and institutional review boards (IRBs) often require such statements in a study protocol before data collection can begin.[1] The rationale for study size in an RCT usually depends on calculations of the study size needed to achieve a specified level of statistical power for the primary hypothesis under study, defined as the probability of rejecting the null hypothesis when a specific alternative hypothesis (the primary hypothesis under study) is true. In the case of a trial comparing treatments, this is the probability of finding a statistically significant difference between treatments in the primary outcome if the treatments do indeed

differ by the amount specified. Several software packages and online tools exist for performing these calculations, such as Stata and Power Analysis and Sample Size (PASS).[2-3] Textbooks give more detail on the calculations for a wide variety of data structures and statistical models.[4]

Calculating statistical power requires specification of several investigator choices and assumptions, each of which has important implications and must be specified with sufficient scientific rationale. Most importantly, investigators must specify a primary study outcome and a minimum treatment effect of interest for that outcome. This quantity, often referred to as the clinically meaningful or minimum detectable difference, identifies the size of the smallest potential treatment effect that would be of clinical relevance. Study size is calculated assuming that this value represents the true treatment effect. If the true treatment effect is larger than this quantity, then the power for a given study size will be even higher than originally calculated.

In addition to the minimum treatment effect of interest, calculating the needed study size requires specifying a measure of data variability. In trials with a continuous outcome (e.g., LDL cholesterol), investigators must make assumptions about the standard deviation of the outcome in each trial arm; when the outcome is the occurrence of an event (e.g., death), then an assumed event rate in the control group is necessary. If the assumed event rate in the control group is combined with the specified treatment effect of interest, then one can calculate the expected event rate in each group if the minimum clinically important treatment effect is achieved. The CONSORT statement recommends reporting these quantities (the expected results in each group under the minimum detectable difference) rather than the minimum detectable difference. It is recommended that estimates of standard deviations and event rates used in study size calculations be taken from existing literature or pilot studies when available.

Finally, needed study size depends on the chosen Type I error rate (α) and the required statistical power. For the majority of studies, the conventional cutoff for statistical significance, $\alpha = 0.05$, is used, but this quantity should be clearly specified nonetheless. Many studies also use a standard required power of 80 percent, although other values are often considered. In RCTs that have study size constraints, due to budget or the pool of available patients, the power obtained from the achievable study size should be described. Potential reductions in the number of recruited patients available for analysis (e.g., due to loss to followup) should also be discussed.

Table 9.1 shows an example of an adequate consideration of study size under several potential scenarios that clearly specify assumptions about the baseline risk of the primary outcome under study, the minimum clinically relevant treatment effect, and the required power. In this table, all of the necessary quantities are reported for determining the adequacy of the chosen study size; and investigators, funding agencies, and ethics review boards can make informed decisions about the potential utility of the planned study.

Table 9.1. Example study size table for an RCT comparing the risk of death for two alternative therapies*

Scenario	Effect of Interest	Therapy 1 Risk	Therapy 2 Risk	Desired Power	Needed Study Size	Needed Recruitment
1	0.75	0.020	0.015	80%	10,795	13,494
2	0.75	0.100	0.075	80%	2,005	2,507
3	0.50	0.100	0.050	80%	435	544
4	0.50	0.100	0.050	90%	592	728*

All calculations assume a Type I error rate of 0.05. The effect of interest is specified as a risk ratio. Study size is reported per treatment arm, and a 20% dropout rate is assumed for calculating the needed recruitment.

These considerations in sample size and power in the context of RCTs are also relevant for nonrandomized studies, but their application in nonrandomized studies may differ. The following section is for additional consideration, particularly for nonrandomized studies.

Considerations for Observational CER Study Size Planning

Bland has commented that funding agencies and journals put investigators in an inconsistent position: Funding agencies ask for statistical power calculations to test one hypothesis for the primary outcome, yet journals ask for confidence intervals.[5] In his commentary, Bland proposed that we resolve that inconsistency by asking investigators to base their study size on the expected precision of all relevant comparisons. Goodman and Berlin recommended a similar idea in 1994 (page 204 of their article):[6]

> In our experience, expressing the implications of sample size calculations in the same language as is used in a published paper, instead of the language of power and detectable differences, helps researchers to understand the implications more clearly and take them more seriously. This in turn can produce meaningful discussions about the aims of the study, which power considerations rarely seem to inspire.

Basing the study size on the expected width of confidence intervals offers another advantage: Investigators no longer need to commit to a primary outcome and a primary comparison (e.g., among alternative interventions).

Many funding agencies, however, rely on the conventional power calculations advocated by most trialists. Therefore, this section primarily focuses on power calculations and adapts trialists' conventional advice to nonrandomized or observational studies because they introduce complexities that randomized trials do not need to consider. For example, investigators may not be able to estimate the power or precision of their proposed comparisons until they have generated the propensity score and constructed matched cohorts, which may exclude patients and interventions that appeared eligible when the cohort was assembled.

Case Studies

Schneeweiss and colleagues published one of the first Developing Evidence to Inform Decisions about Effectiveness (DEcIDE) Program studies on comparative effectiveness; they compared the short-term risk of mortality in elderly patients who started a conventional versus an atypical antipsychotic medication regimen,[7] reproducing an earlier study by Wang and colleagues.[8] Consistent with most nonexperimental studies, especially in the pre-STROBE era, their methods section does not offer a rationale for the cohort study's size. Based on their patient counts for each class of antipsychotic medication and the number of deaths observed during the first 180 days after starting medication, we calculated the statistical power for their study question: Do conventional antipsychotic medications pose a higher risk than atypical antipsychotic medications as measured by all-cause mortality?

We considered an inferiority hypothesis by using the crude mortality risk observed in the control cohort of atypical medication patients (9.58 percent), and then assigning the conventional medication cohort a 10-percent higher risk (10.54 percent), a clinically important excess risk. Based on the numbers of patients and deaths noted above, Stata's sample size command, *sampsi*, reported statistical power of 0.83. Their subgroup analyses would have had lower power, but the main study was appropriately powered for its primary outcome and comparison.

Considerations That Differ for Nonrandomized Studies

Power calculations may require additional considerations for application to nonrandomized studies. For a well planned and conducted RCT, the Type I and Type II errors (i.e., false positive or false negative) rank higher as possible explanations for a finding of "no statistically significant difference" because randomization has overcome the potential confounding, the protocol has reduced measurement error, et cetera. But for nonrandomized studies, Type I and Type II errors rank lower on the list of possible explanations for such a negative result. Confounding bias, measurement error, and other biases should concern investigators more than the expected precision when they consider the feasibility of a

comparative effectiveness study. For example, the new user design trades precision for a reduction in confounding bias by restricting the study to incident users of the interventions under study. (See chapter 2 for a discussion of new user design.)[9] As retrospective database studies become larger through distributed networks, insufficient statistical power of comparative effectiveness estimates will diminish in importance as a competing explanation for negative results—at least for the primary comparison of common interventions—and readers will need to consider whether small observed clinical differences matter for decisionmaking. For example, database studies may identify small excess risks of about 5 percent that would fall below the minimum clinically important difference specified in a prospective study.

In some cases, controlling for confounding can also reduce the precision of estimated effects. The reduction in precision is perhaps most clearly seen in studies that use propensity score matching. With propensity score matching and strong preferential prescribing in relation to patient characteristics (i.e., less overlap in propensity score distributions across cohorts), many patients will drop out of the analysis.[10] For example, Solomon and colleagues identified a cohort of 23,647 patients who were eligible for a comparative effectiveness study, but only 12,840 (54 percent) contributed to the final analysis after matching on the propensity score.[11] Inconveniently, the development of the propensity score occurs after the study protocol has been written, and the investigators have invested considerable time and effort toward completion of the comparative effectiveness study. Consequently, investigators should consider incorporating sensitivity analyses when calculating the expected precision of effects and study size estimates. For example, they might ask, "If 25 percent of the cohort were to drop out of the analysis after incorporating the propensity score, how would that reduced study size impact the expected precision?"

Because retrospective studies lack a protocol for data collection, they often suffer a higher frequency of missing data, especially for clinical examination values (e.g., blood pressure, body mass index, and laboratory results). Investigators who undertake a completed-cases analysis, which excludes patients with any missing data for key variables, may suffer from a smaller study size than they anticipated when they wrote the study protocol.[12] Depending on the nature of the missingness, it may be possible for investigators to impute certain values and retain patients in the final analysis. But as with the development of propensity scores, multiple imputation is labor intensive, and its success in retaining patients will only be known after the protocol has been written.

Conclusion

In order to ensure adequate study size, investigators should provide a rationale for study size during the planning stages of an observational CER study. All definitions and assumptions should be specified, including the primary study outcome, clinically important minimum effect size, variability measure, and Type I and Type II error rates. Investigators should also consider other factors that may reduce the effective sample size, such as loss to followup, reductions due to statistical methods to control confounding, and missing data, when making their initial assessment as to whether the sample size necessary to detect a clinically meaningful difference can be achieved.

Checklist: Guidance and key considerations for study size planning in observational CER protocols		
Guidance	**Key Considerations**	**Check**
Describe all relevant assumptions and decisions.	Describe: - The primary outcome on which the study size or power estimate is based. - The clinically important minimum effect size (e.g., hazard ratio ≥1.20). - The Type I error level. - The statistical power or Type II error level (for study size calculations) or the assumed sample size (for power calculations). - The details of the sample size formulas and calculations, including correction for loss to followup, treatment discontinuation, and other forms of censoring, and the expected absolute risk or rate for the reference or control cohort, including the expected number of events.	☐
Specify the type of hypothesis, the minimum clinically important excess/difference, and the level of confidence for the interval (e.g., 95%).	- Types of hypotheses include equivalence, noninferiority, inferiority.	☐
Specify the statistical software and command, or the formula to calculate the expected confidence interval.	- Examples include Stata, Confidence Interval Analysis, Power Analysis and Sample Size (PASS).	☐
Specify the expected precision (or statistical power) for any planned subgroup analyses.		☐
Specify the expected precision (or statistical power) in alternative special situations, as in sensitivity analyses.	Special situations include: - The investigators anticipate that strong confounding that will eliminate many patients from the analysis (e.g., when matching or trimming on propensity scores). - The investigators anticipate a high frequency of missing data that cannot (or will not) be imputed, which would eliminate many patients from the analysis.	☐

References

1. Schulz KF, Altman DG, Moher D; CONSORT Group. CONSORT 2010 statement: updated guidelines for reporting parallel group randomized trials. Ann Intern Med. 2010 Jun 1;152(11): 726-32.
2. StataCorp. Stata Statistical Software: Release 11. College Station, TX: StataCorp; 2009.
3. Hintze, J. PASS 11. NCSS, LLC. Kaysville, Utah; 2011. www.ncss.com. Accessed September 21, 2012.
4. Friedman LM, Furberg CD, DeMets DL. Sample size (chapter 8). In: Fundamentals of Clinical Trials. 4th edition. New York: Springer; 2010: 133-167.
5. Bland JM. The tyranny of power: Is there a better way to calculate sample size? BMJ. 2009;339:b3985.
6. Goodman SN, Berlin JA. The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. Ann Intern Med. 1994;121:200-6.
7. Schneeweiss S, Setoguchi S, Brookhart A, et al. Risk of death associated with use of conventional versus atypical antipsychotic drugs among elderly patients. CMAJ. 2007;176:627-32.
8. Wang PS, Schneeweiss S, Avorn J, et al. Risk of death in elderly users of conventional vs. atypical antipsychotic medications. N Engl J Med. 2005 Dec 1;353(22):2335-41.
9. Ray WA. Evaluating medication effects outside of clinical trials: new user designs. Am J Epidemiol. 2003;158:915-20.
10. Schneeweiss S. A basic study design for expedited signal evaluation based on electronic healthcare data. Pharmacoepidemiol Drug Saf. 2010;19: 858-68.
11. Solomon DH, Rassen JA, Glynn RJ, et al. The comparative safety of analgesics in older adults with arthritis. Arch Intern Med. 2010;170: 1968-78.
12. Sterne JA, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiologic and clinical research: potential and pitfalls. BMJ. 2009;338:b2393.