Journal of Clinical Research Best Practices

Vol. 8, No. 6, June 2012 “Happy Trials to You”

Optimizing Sample Size

By Carleton Southworth

Introduction

If the sample size in a clinical trial is larger than necessary, the excess time and money will be wasted. Even worse, if it is too small, all the time and money will be wasted. Efficient clinical trials are designed with a sample size in the “Goldilocks zone” — just the right size.

Determining a sample size that is “just right” primarily depends on the following five factors:

- The characteristics of the treatment effect
- The magnitude of the difference between treatments that needs to be detected (a clinically meaningful difference sometimes referred to as a “delta”)
- The statistical “background noise”
- The type of measurement scale being used
- The error properties of the measurement system

Understanding these factors can help you design your clinical trial to minimize sample size, while still providing a reliable answer to your research question. Put another way, the statistical power associated with a given sample size can increase or decrease depending on the five factors listed above. Without understanding these factors, it is very difficult to choose the optimal sample size. In many cases, some of these factors can be controlled to reduce the sample size requirement.

Except for exploratory trials, almost every clinical trial is designed to test a hypothesis — to answer a research question or questions. When selecting a sample size, it is necessary to identify a primary research question — a primary outcome — that is more important to answer than other, ancillary questions. This primary question drives sample size. To illustrate the principles outlined here, we will use the hypothetical study to test the primary hypothesis that a new drug, Newocol, lowers LDL cholesterol more than an existing drug, Oldocol.

Determine the Likely Size and Consistency of the Treatment Effect

If the likely treatment effect is large, a relatively small sample is required. For example, it is easier to tell if a treatment extends a life by five years than if it slightly shrinks the size of a solid tumor over time. Similarly, if the effect of a treatment is consistent and predictable, the sample can be smaller.

In our example, if Oldocol typically reduces LDL cholesterol by 100 mg/dL, it is a lot easier to measure a 5% improvement with Newocol than if Oldocol typically reduces LDL cholesterol by only 20 mg/dL.

Select a Minimum Detectable Difference

It is critical to identify a minimum clinically meaningful difference that you want to be able to detect with your primary outcome.[2] This difference (delta) is often based on expert judgment. Consequently, input from clinical investigators can be essential to ensure that the difference is identified accurately. On one hand, if you choose a delta that is too large, your sample will be too small to reliably detect a difference that would have been important to detect. On the other hand, if you choose a small difference that is not clinically important, then sample size and costs will be inflated trying to find a clinically insignificant difference that will not ultimately cause physicians to start prescribing the new drug.

If equivalent performance is adequate, e.g., because there are dose regimen, side effect, or economic advantages, the question then becomes how close is close enough. "Very close" requires a larger sample than "fairly close."

In our example, we will assume that Newocol is very similar to Oldocol, so lowering LDL cholesterol 10 mg/dL below the control drug would be clinically meaningful. If we were to pick a difference of 1 mg/dL ("new and improved!"), then the sample size required would increase significantly. We could detect a tiny difference, but it would be one so small that it would not induce prescribers of cholesterol lowering drugs to change their practice. Selecting too large a difference, say 30 mg/dL, would yield a smaller sample size, but the statistical power yielded by the small sample size would likely miss the 10 mg/dL difference. If a delta of 10mg/dL would create a successful drug, we might still want to size the sample to reliably detect a delta of, say, 8 mg/dL to reduce the risk of a false negative.

Minimize Statistical "Background" Noise

Many things can affect LDL cholesterol levels. Even with no changes in medication, the level can change from day to day due to diet, exercise or completely unsuspected factors. These changes are called "background noise." With more noise, it is harder to "hear" the signal we are trying to detect. The harder it is to hear the signal, the larger the sample must be. Therefore, by reducing background noise, we can "sharpen" the ability of a given sample size to measure the effect of the treatment and thus lower the sample size required.[3]

The placebo response is a special example of background noise. In some therapeutic areas, e.g., depression, it can exceed 50%. Unfortunately, it can also vary unpredictably from study to study and individual to individual. If you expect an average positive response in the placebo arm of a study of 15%-25% and a treatment effect of 5%, a very large sample will be required. Fortunately, in our LDL cholesterol example, while placebo effects can occur, they are minimal.

Subject Homogeneity

If all of the subjects in your trial are nearly identical, then their baseline values prior to treatment will be relatively similar. Likewise, the variables that affect LDL levels will be nearly alike. Subjects who are nearly alike are "homogeneous," whereas subjects who differ are "non-homogeneous." Reducing between-subject variation in variables that affect the primary outcome variable reduces background noise, and thus makes it possible to reduce sample size. However, there is a trade-off: If you conduct a clinical trial only on women aged 20 to 25 with BMIs between 18 and 28, then your results would be generalizable only to future patients who fit into this category. Nonetheless, the cost savings may be worth the limitation (especially if physicians choose to prescribe your drug off label). It can be a particularly good strategy if your new treatment is aimed primarily at a relatively homogeneous group of patients, or if your new treatment is likely to be especially effective in a select group of patients. It may be possible to expand your indications later. In our example, it might make sense to test Newocol on sedentary patients in their 60s and 70s with a baseline LDL of 150-200 mg/dL.

Use Subjects as Their Own Controls

In medical device trials, it may be possible to use a subject as his or her own control. For example, in a total joint study, it may be possible to find patients who require bilateral hip replacement. By randomly assigning a new treatment hip to either the right or left side, and then using the control hip on the other side, you can ensure that the two treatment groups are identical because each subject will have both an experimental and a control hip. This approach reduces statistical background noise and thus reduces the sample size required to detect a difference.

In a pharmaceutical trial like our Newocol trial, a similar result is sometimes feasible with a case-crossover trial.[4] In such a trial, two groups are randomly formed. One group starts on the experimental treatment and the other on the control treatment. Each group then switches treatment (after a period to allow the experimental and control drugs to "wash out"). In some cases, the groups switch treatments more than once. By the end of a case-crossover trial, the response of every subject to both drugs will have been measured. A case-crossover design might be feasible for the Newocol trial.

Control the Environment

Most Phase I trials are conducted in facilities where variability in exercise, diet, sleep, etc., can be minimized. Similarly, requiring ambulatory subjects to take study medications at specific times of day limits variability. If variability cannot be limited, sometimes it can at least be measured, e.g., with diaries; known variability is statistically better than unknown variability because it can be correlated with treatment effects. In the Newocol trial, practical considerations and the requirement for generalizable data might prevent attempts at controlling diet or exercise, but these factors can still be recorded in diaries.

Pick a Variable with the Optimal Scale Type

There are different kinds of variables — (1) categorical variables such as "cure" versus "no cure," (2) rank-order variables such as "poor," "fair," "good," and "excellent" — where multiple potential responses can be arranged on a continuum but where the difference between each of the levels may not have a consistent meaning, and (3) interval-level variables, such as blood pressure and temperature, where every step up the scale measures the same amount of increase. Moving from categorical to rank order to an interval-level scale increases the amount of information in a piece of data. Therefore, sample size requirements become smaller as one moves from categorical to rank order to interval level measurement scales.[1] Thus, where feasible, pick an interval-level variable to measure your primary endpoint. Notice that an interval-level scale can be converted to a rank-order scale — blood pressure could be arranged into "too low," "optimal" and "too high" categories — but doing this loses information. It may make sense to do this in some instances, but be aware of the consequences. In our example, we have an easy choice: we select mg/dL to measure LDL cholesterol. We could use a rank-order variable to capture LDL cholesterol level, such as "optimal," "near optimal," "borderline high," "high" and "very high," but at the cost of increasing the required sample size.

Improve Your Measurement Method

Most measurement methods include some error. For example, when measuring blood pressure, there can be variability associated with the person who is taking the blood pressure, the measurement device, the time of day, whether the subject is standing, sitting or prone, and whether the measurement is taken at the beginning or end of an examination.

Think of Your Measurement Method as a System and Standardize it

In our example, ensuring that all individuals taking blood pressure are trained to take it in exactly the same way can help to minimize measurement error and, consequently, help to keep sample sizes smaller. Think of your measurement method as a system that may involve individuals, measurement instruments, environmental factors, and so on. Each element of the system can create variability, but some elements are more important than others and therefore deserve more attention.

Standardizing the measurement system can reduce error and thus reduce sample-size requirements.[5] In our LDL trial, it makes sense to select a single laboratory to measure all samples. We should ensure that the lab is fully qualified under GLPs (Good Laboratory Practices). We should find a laboratory that has already standardized its measurement system. We might want to send a standard blood sample to the lab each month to detect any variations.

In other trials, especially those measuring psychological variables (as might be the case with psychiatric drugs), it might be necessary to train the evaluators and assess their performance to ensure that measurements are standard across sites and across evaluators. It might be necessary to have two evaluators read X-ray images or biopsies, with an adjudication process, to ensure consistency.

Measure Multiple Times

In some lab tests, a good method for reducing error is to measure the same sample multiple times and then average the result.[6] The measurement error associated with an average of two or three measurements (or more) will be less than the error associated with a single measurement. Reducing measurement error using this method reduces the sample size that is required. This method may be especially effective when measurement errors are known to be large. The author knows of one example where a statistician measured a bowling ball to within a fraction of an ounce by having a large room full of people take turns holding the bowling ball and guessing the weight. The average of all of the guesses was highly accurate. In our example, we might be able to lower the sample size by measuring each aliquot of blood several times. However, if we are looking for a delta of 10 mg/dL and the lab generates readings consistent to within 0.1 mg/dL, multiple measurements will not be useful.

Conclusion

The primary factors that affect optimal sample size are the characteristics of the likely treatment effect, the minimum clinically meaningful difference to be detected, the level of background noise, the type of measurement scale, and the error properties of the measurement system. Some of these factors can be controlled to reduce sample size and thus the time and cost of a clinical trial. Some factors have more impact than others, so attention should be given to them accordingly. Involving a biostatistician during study design can help ensure that these factors are addressed properly and the clinical trial will be able to answer the primary research question efficiently.

References

1. Sullivan L. Essentials of biostatistics in public health. Jones and Bartlett: Sudbury, MA; 2008.
2. Wells G, Li T, Maxwell L, MacLean R, Tugwell P. Determining the minimal clinically important differences in activity, fatigue, and sleep quality in patients with rheumatoid arthritis. J Rheumatol. February 2007: 34(2):280-289.
3. Sawyer S. Analysis of variance: the fundamental concepts. Journal of Manual & Manipulative Therapy. 2009: 17, E27-E38.
4. Maclure MA, Maclure M, Robins MJ. The case-crossover design: a method for studying transient effects on the risk of acute events. American Journal of Epidemiology 1991; 133(2):144-153.
5. Warnick G, Kimberly M, Waymack P, Leary E, Myers G. Standardization of measurements for cholesterol, triglycerides, and major lipoproteins. Labmedicine, August 2008 j Volume 39 Number 8.
6. Fridman A. The quality of measurements: a metrological reference. Springer, New York: 2012.

Author

Carleton Southworth, AB MS RAC, is Director, Biostatistics at American Research Partners. Contact him at 1.574.367.8076 or csouthworth@americanresearchpartners.com.