

Sample Size for a Clinical Trial: Why Do Some Trials Need Only 100 Patients and Others 1000 Patients or More?

Roy W. Beck, MD, PhD - Tampa, Florida

The objective in determining the number of patients for a randomized clinical trial is to select a sample size that is large enough so that the trial will have a good chance of detecting a benefit of treatment, if one exists, but is no larger than it needs to be. Why do some trials need only 100 patients, others 500, and others 1000 or more?

To understand how the sample size is derived for a randomized trial, it is useful to start with the concept of the null hypothesis. In an efficacy trial, the null hypothesis is that the treatment and placebo have the same effect. The aim of the trial is to provide evidence that the null hypothesis is wrong. If there is no true benefit of treatment (i.e., the null hypothesis is true), then the proportion of outcome events that occurred in the trial would be expected to be the same in the treatment and control groups. However, by chance one group is likely to have a higher proportion of the outcomes than the other group. Because of this, we can never prove that the treatment and control groups are exactly equal; rather, we can estimate the chance that we would find the observed difference in the distributions (or differences that are even more extreme) if in truth the outcome risks were exactly equal. The *P* value of the statistical test comparing the groups estimates this probability. If, a priori, it is specified that we are willing to accept a 5% type 1 error (i.e., a 5% chance we will conclude that there is a treatment group difference when, in truth, treatment is not better or worse than a placebo), then we can conclude that the treatment is beneficial when the treatment group has a lower failure rate than the placebo group and the *P* value is <0.05 . An example will help to illustrate this.

Assume that we need to determine the sample size required for a trial evaluating whether a treatment can reduce the proportion of patients whose vision worsens (such as a 3-line loss of visual acuity) compared with a placebo control by 50% (i.e., half as many persons in the treated group will lose vision compared with the control group). The sample size for such a trial is determined mainly by how often we expect the control group will have the bad outcome. This is often referred to as the failure rate or event rate. A low control group failure rate means that a large number of subjects are going to be needed to be able to evaluate whether the treatment can reduce the number of failures. The lower the control group failure rate, the larger the sample size will need to be. I will try to explain why the control group outcome rate is key.

If the study randomized 100 subjects to 2 groups (50 per group) and 10 of the subjects developed the study outcome (10% overall outcome rate), how would these 10 outcomes have to be distributed between the 2 groups to reject the null hypothesis of no treatment effect (i.e., to have confidence that a group difference is not due to chance)? If the split

were 7 outcomes in the control group and 3 in the treatment group, this would be greater than a 2:1 difference between groups ($>50\%$ reduction in outcomes). However, the *P* value indicating the probability of this split, or one more extreme, occurring by chance is 0.32. So, if we had a sample size of 100 (50 per group) and found 7 outcomes in the control group and 3 in the treatment group, we would not have a high degree of confidence that this did not occur by chance, even though more than twice as many failures occurred in the control group than in the treatment group. If we repeated this 100-patient trial over and over, and the true treated group and control group failure rates were exactly equal, even an 8 to 2 or more extreme split would happen more than one time in 20 trials. We would need to have a 9:1 or 10:0 split of the 10 outcomes for the *P* value to be $<5\%$.

Therefore, it can be seen that we will need to have considerably more than 100 subjects to evaluate a projected outcome split of 2:1 between the groups. With 200 subjects (100 per group) and an outcome rate of 10%, there would be 20 outcomes. The *P* value for the probability of a 14:6 or more extreme split between the 2 groups is 0.11, and that for a 15:5 (or more extreme) split is 0.04. Thus, if we found a 2:1 difference in outcomes between groups, again we would not have sufficient confidence to reject the null hypothesis. If we had 400 subjects (200 per group) and 40 outcomes, then a split of 27:13 would have a *P* value of $<5\%$. The actual sample size needed for a study would be larger than 400 because we need to account for losses to follow-up and both type 1 and type 2 errors (the type 2 error is the probability of not finding a treatment group difference when in truth treatment is beneficial), but this example provides an indication of why sample size is so dependent on the outcome rate.

As a general rule of thumb, within a fairly wide range of control group outcome rates a trial needs to have on the order of 40 to 60 outcomes in the control group to be able to detect a benefit of treatment that is on the order of a 50% reduction in the failure rate. So, if the control group rate is 10%, each group will need to have roughly 500 patients; if the control group rate is 5%, this increases to roughly 1000 patients per group, and if the control group rate is 50%, this decreases to roughly 100 patients per group. So, to get a ballpark estimate of the sample size per group for a 50% relative treatment effect, divide 50 by the control group outcome proportion (e.g., $50/0.1 = 500$ in the above example).

As has been demonstrated so far, the larger the projected failure rate for the control group, the smaller the sample size. Although there is always a desire to have the sample size be as small as possible, it is important not to overesti-

mate the control group failure rate. Here is an example of the consequences of this happening. One of the trials conducted by the Herpetic Eye Disease Study was designed to evaluate whether treating acute epithelial keratitis with a 21-day course of oral acyclovir, compared with a placebo, was beneficial in reducing the development of stromal keratitis or iritis in the ensuing 12 months.¹ In designing the trial, the sample size was calculated to be 502 patients, based on projecting an outcome rate of 20% in the placebo group and 10% in the acyclovir group (after accounting for losses to follow-up). About halfway through the trial, an interim analysis showed that the outcome rate in the placebo group was only 10%. If 10% had been used for the original sample size estimation, then the required sample size would have been more than double, for a 50% treatment effect. The Data and Safety Monitoring Committee concluded that it was unlikely that a treatment benefit could be shown in the original sample size of 502, and doubling the sample size was not feasible. Therefore, the trial was stopped early.

The second important factor that determines sample size is the magnitude of difference between treatment groups. In the above example, I projected a 50% reduction in the failure rate comparing the treatment and placebo groups, a percentage that is commonly used. If there is a desire to evaluate a smaller treatment effect, the sample size will need to be larger, and vice versa. If a large treatment effect is projected (often in a misguided effort to achieve a small sample size), it is important to keep in mind that the study will be underpowered to detect a treatment effect any smaller than this. Therefore, the study results might be negative, even though the treatment has a meaningful benefit (but not as effective as originally projected).

The 2 other factors that need to be set to compute sample size are the type 1 error and type 2 error rates. In conducting a trial, we can never be certain that we will obtain the correct answer. So we need to decide (1) what risk we are willing to take that the study will conclude erroneously that treatment is beneficial when the treatment has no effect and the null hypothesis is true (type 1 error, also known as α) and (2) what risk we are willing to take that the trial will miss a true treatment effect (type 2 error, also known as β) and conclude erroneously that treatment is not beneficial. The type 1 error is generally set at 5% and the type 2 error at 10% or 20%, often without much consideration for the specific objectives of the trial. More often, the type 2 error is expressed in terms of the statistical power of a study to detect a benefit of treatment if one exists; computationally, statistical power is $1 - \beta$. The thought process behind setting the type 1 and type 2 error rates is a topic for another article.

In conclusion, trials need large numbers of patients when the outcome event rate is low or there is a desire to detect a small benefit of treatment. If the failure rate for the control group is overestimated in calculating the sample size, then the trial will be underpowered to detect a treatment benefit and may not provide useful results. Studies designed to detect only large benefits of treatment will require smaller sample sizes, but are likely to miss meaningful treatment benefits.

Reference

1. Herpetic Eye Disease Study Group. A controlled trial of oral acyclovir for the prevention of stromal keratitis or iritis in patients with herpes simplex virus epithelial keratitis. The Epithelial Keratitis Trial. *Arch Ophthalmol* 1997;115:703–12.