What Information Will a Statistician Need to Help Me With a Sample Size Calculation?

Caitlyn Ellerbe, PhD

As an investigator, one of the most pressing questions during the planning phase of a clinical trial is the sample size. It is desirable and necessary to include your statistician in the discussion of trial design well before the topic of sample size is broached, particularly in the case of innovative or adaptive designs. Although there are a multitude of available designs, we will focus on the 2-arm, superiority trial design wherein 2 groups of subjects are compared with respect to a given outcome of interest.

The foundation of a trial design is the hypothesis of interest. At the end of the trial, one of two conclusions is possible: (1) there is sufficient evidence that the 2 groups are different (ie, reject the null hypothesis) or (2) there is NOT sufficient evidence that the 2 groups are different (ie, fail to reject the null hypothesis). Given this hypothesis, the first step is to define a primary end point that will be used to test the hypothesis. For example, if the null hypothesis states that there is no difference in the proportion of deaths between the treatment and control group, then the primary end point must be the proportion of deaths. Once the primary end point has been identified, it is important to discuss the trial design that will be used to collect information on the primary outcome (eg, parallel, factorial, or crossover designs). For the purpose of this article, it is assumed that investigators have already discussed these design issues with the statistician and are now ready to calculate the required sample size given the design parameters. We begin with the key components of a sample size calculation—statistical significance level and clinical effect size; we then expand the discussion with inflation of the sample size for protocol nonadherence and additional concerns for studies using adaptive designs.

Sample Size Calculations

All power calculations can be generalized to consisting of 2 components, the degree of certainty with which to test the stated hypothesis (statistical significance and power) and the expected difference between the 2 treatment arms measured in the trial (clinical effect size).

Statistical Significance and Power

When the clinical effect size is small or the variability in the effect is large, the sample size of the trial must increase to accurately reflect the population. By sampling, we assume that the sample is representative of the population. If our assumption is correct, then the result observed in the trial is valid. Conversely, if the assumption is incorrect, there are 2 ways we can make an error (Table).

In the first case, no difference exists but the conclusion at the end of the trial rejects the null hypothesis; this is referred to as the type I error probability. By specifying the type I error probability, we declare the degree of confidence in an estimate. For most trials, this is set at 5%, which can be interpreted as a 5% chance that the null hypothesis is rejected when no difference exists (ie, false-positive error) and is equivalent to a 95% confidence interval. When multiple hypothesis tests are conducted using the same data, investigators must ensure that all tests are controlled at this same type I error probability (global type I error probability) using multiplicity adjustments. The choice of type I error is somewhat arbitrary and could be made more or less stringent depending on the risks and benefits of the trial.

In the second case, a true difference does exist, but the trial fails to reject the null hypothesis; this is referred to as a type II error probability. More commonly, investigators are interested in the power of a trial; defined as the probability of rejecting the null hypothesis if a true difference exists. Thus, by specifying power for the trial at 80%, we are accepting a 20% chance that we will fail to reject the null hypothesis when a true difference exists (ie, commit a type II error). Moreover, for the same trial, if power was increased to 90%, the odds of a failed trial when a true treatment difference exists would be reduced from 4:1 to 9:1. For this reason, it is often helpful to examine a plot of power versus sample size when considering the number of extra subjects required to increase the power of a design.

At this point, an investigator may decide that both the type I and type II errors should be set as close to zero as possible to minimize the chance of errors. As the sample size increases, both the type I and type II error probabilities will approach 0%. However, for a fixed sample size, there is an inverse relationship between the type I and type II errors. By decreasing the type I error probability, it becomes more difficult to reject the null hypothesis. If no difference exists, this has the
Table. Possible Outcomes for a Hypothesis Test

<table>
<thead>
<tr>
<th></th>
<th>Fail to Reject Null Hypothesis</th>
<th>Reject Null Hypothesis</th>
</tr>
</thead>
<tbody>
<tr>
<td>True result: no difference exists</td>
<td>Correct decision</td>
<td>Type I error</td>
</tr>
<tr>
<td>True result: difference exists</td>
<td>Type II error</td>
<td>Correct decision (power)</td>
</tr>
</tbody>
</table>

desirable effect of increasing the probability of a correct decision. However, if a difference does exist, it will still be more difficult to reject the null hypothesis; that is, the trial will have lower power and an increased type II error probability. For most trials, it is considered more dangerous to endorse an ineffective treatment (type I error) than to discount an effective treatment (type II error); thus, it is common practice to select a stringent type I error (5%) with less power (no less than 80%). However, investigators should discuss these thresholds when determining the specific sample size for a trial.

Clinical Effect Size

The second component of calculating sample size is an estimate of the effect size one wishes to observe to declare the groups different with respect to the specified end point. Effect size is defined as the minimum clinical difference that would be required to pursue further study (early phase trials) or change current practice (confirmatory trials) and is derived from clinical consensus, as well as estimates from previously reported results in similar populations or pilot studies. Furthermore, this estimate should take into account the reported effect size of the primary outcome under current standard of care and the potential downsides of a treatment (eg, cost, risk of harm).

As the difference in effect size between the control and experimental group decreases, the number of subjects required to detect this difference will increase. There is often a temptation to overestimate this difference in an effort to justify a reduced sample size; however, doing so may result in an underpowered trial. Investigators may instead decide to under-estimate the difference in an effort to protect against incorrect assumptions. Although this is less risky, it also should be undertaken cautiously as overpowered trials mean that resources are wasted and an excess number of subjects may be exposed to a suboptimal treatment.

Allocation Ratio

At the beginning of a trial, investigators have a belief, known as clinical equipoise, that there is genuine uncertainty about whether the experimental arm will perform better than the control arm. It is this uncertainty that provides the ethical justification to randomize subjects to each arm. Moreover, it is this uncertainty that also commonly leads us to allocate subjects evenly (1:1) between the arms. However, there are certain scenarios in which we may want to adjust this allocation ratio, favoring more subjects in either the control or experimental arm. This is sometimes done if subject recruitment is a problem, if strong anecdotal evidence in favor of treatment exists, if treatment safety is of concern, or if the cost makes one treatment more prohibitive. Although the decision to adjust the allocation ratio is clinical, it should be noted that the smallest total sample size (largest power) is achieved with equal allocation, and the increased number of subjects will be a function of the magnitude of imbalance between treatment arms (ie, a 1:4 allocation will require more subjects than a 1:2 allocation).

Inflation of Sample Size Because of Protocol Noncompliance

When the sample size is calculated, the number reported is the number of subjects needed at the end of the trial to achieve the specified power. However, it is rarely the case that the number of subjects enrolled in a trial is equal to the number of subjects with primary outcome at the end of a trial. As a result, it is necessary to inflate the sample size to account for these losses to maintain the statistical power. The 2 major types of losses are missing data and unplanned treatment crossover cases. In the first instance, the primary end point is not available at the end of the trial because the subject was either lost to follow-up, withdrew consent before completing the protocol follow-up, or died. The second type of loss occurs when a subject completes the trial but does not adhere to the protocol (eg, fails to take study medication).

For an intent-to-treat analysis, data from all subjects must be included. Thus, if the primary outcome is missing, it should be imputed, and if a subject does not comply, they must still be included in the group assigned at the beginning of the trial. However, the resulting difference in effect size will be diminished because subjects assigned to the treatment arm will, for instance, have outcomes expected of the control arm. For a per-protocol analysis, only data from subjects adhering to the protocol are included in the analysis of the primary outcome, thus the number of subjects in the analysis sample is reduced. During planning of the trial, investigators should estimate the rate at which these events are expected to occur so that the sample size can be adjusted accordingly. In addition and regardless of any sample size inflation, investigators should plan to reduce these events through methods, such as follow-up calls, streamlined data collection, and monitoring protocol adherence.

Adaptive Designs

Adaptive designs are characterized by a prospectively planned algorithm to modify the trial design based on accumulated data in the trial. Common adaptations include early stopping for overwhelming efficacy or futility, dropping one of the multiple arms, selecting an optimal subject population, and increasing the sample size because of higher than expected variability of the primary end point. The value of these designs is that they can account for uncertainty and insufficient information at the beginning of the trial. The cost of this flexibility is 2-fold. First, the sample size will tend to be larger than a comparable nonadaptive design. This arises because the trial seeks to answer additional questions using the collected data. Second, although investigators are permitted additional uncertainty with respect to the final design, they must determine how each potential decision will affect the assumptions of effect size.
As an example, consider the following adaptations. In a group sequential design, the primary hypothesis is tested at multiple interim analyses. If the test statistic at one of these looks crosses a prespecified boundary, the trial is stopped early for success or futility. In this way, although the trial is powered for a hypothesized effect size, it is possible to reduce the number of enrolled subjects if the effect size is significantly larger (stopping for efficacy) or significantly smaller (stopping for futility). However, because the hypothesis test is performed multiple times, each test is performed at a more stringent level. Thus a greater maximum sample size is required when compared with a trial without the group-sequential adaptation.

In the case of an adaptive dose selection design, investigators can initiate the trial with more than one candidate dose and select the best dose midway through the trial using accumulated data instead of studying a single dose that may or may not yield the maximum effect when compared with a control. In this setting, investigators need to prespecify the effect size comparing each dose to the control and must also make assumptions about the effect size comparing the doses to each other. In a nonadaptive design comparing a single dose to a control, investigators only need to specify the effect size of the dose. Thus, although adaptive designs are attractive because investigators are freed from committing to design choices without strong evidence, the sample size equation must now take into account the outcome under each of the possible adaptive design changes.

Discussion
We can now return to the question, what information will a statistician need to help me with a sample size calculation? The following are issues an investigator should consider in preparation for a discussion on calculating the sample size. An example of this relevant information and the resulting sample size calculation is given in Figure for a trial assessing the effects of blood pressure management for stroke.4

- Desired type I error probability and power
- Effect sizes reported in current body of literature for disease or population of interest
- Clinical consensus for maximum/minimum clinical effect size necessary to change practice
- Possible sources of variation in estimate of primary outcome
- Rate of protocol nonadherence

Once an investigator has addressed this list, it is possible to estimate an ideal sample size. In fact, more often, the statistician can provide calculations for sample size under multiple different assumptions (eg, best case scenario for effect size and worst case scenario). The final step is, if necessary, to reduce from the ideal sample size to what is realistically possible. To make this decision, additional information is required. Specifically, how many subjects can be recruited in a reasonable timeframe and the cost to achieve this goal. To determine the sample size of a trial, the entire study team (both statistical and clinical) should have a meaningful and productive discussion on ideal and realistic scenarios.

Acknowledgments
I thank the anonymous reviewers for their thorough and constructive comments to improve the discussions in this manuscript.

Sources of Funding
This work was supported by National Institutes of Health and National Institute of Neurological Disorders and Stroke U01 NS059041 and U01 NS087748; and National Institute of Diabetes and Digestive and Kidney Diseases U01 DK058369.

Disclosures
None.

References

Key Words: sample size
What Information Will a Statistician Need to Help Me With a Sample Size Calculation?
Caitlyn Ellerbe

*Stroke*. 2015;46:e159-e161; originally published online June 11, 2015;
doi: 10.1161/STROKEAHA.115.007982

*Stroke* is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231
Copyright © 2015 American Heart Association, Inc. All rights reserved.
Print ISSN: 0039-2499. Online ISSN: 1524-4628

The online version of this article, along with updated information and services, is located on the
World Wide Web at:
http://stroke.ahajournals.org/content/46/7/e159

**Permissions**: Requests for permissions to reproduce figures, tables, or portions of articles originally published in *Stroke* can be obtained via RightsLink, a service of the Copyright Clearance Center, not the Editorial Office. Once the online version of the published article for which permission is being requested is located, click Request Permissions in the middle column of the Web page under Services. Further information about this process is available in the Permissions and Rights Question and Answer document.

**Reprints**: Information about reprints can be found online at:
http://www.lww.com/reprints

**Subscriptions**: Information about subscribing to *Stroke* is online at:
http://stroke.ahajournals.org//subscriptions/