wcg MAGI | Journal of Clinical Research Best Practices

Vol. 17, No. 9, September 2021 "Happy Trials to You"

Sample Size Re-Estimation

By David R. Bristol

Abstract

Before a comparative clinical research study starts, the number of study participants (sample size) necessary to provide the statistical power to accomplish its objectives can only be estimated. If the sample size is too small, the study will fail to resolve the study hypothesis. If the sample size is too large, the study will take longer than necessary, money will be wasted, and more study participants than necessary will be subjected to the study's risks and burdens.

With interim data, the sample size can be adjusted, but re-estimation is unlikely to be 100 percent accurate. The strategy of planning a study using an optimistic scenario to obtain a small sample size on the assumption that it can be increased later based on interim assessments may not provide the desired result.

Introduction

The sample size for a comparative clinical research study determines the power of a test. The correct sample size can only be estimated based on available data. This problem can be addressed by re-estimating sample size during a clinical study. Sample size re-estimation is a type of adaptive design that has been well discussed in the literature.

Adaptive designs are often presented as a way to conserve resources. For example, Tufts CSDD (2013) estimates that "early study terminations due to futility and sample size re-estimation applied across the portfolio could save sponsor organizations between $100 million and $200 million annually." These estimates have likely increased since 2013. Tufts CSDD (2013)

Glossary

Final Analysis (FA)	The statistical analysis after a study concludes
Interim analysis (IA)	A statistical analysis during a study
Interim Analysis Plan (IAP)	Plan for the interim statistical analysis
n_0	Sample size per arm at IA, ideally about 0.5 N_A
N_A	Assumed sample size per arm at the start of a study
n_{max}	Maximum possible re-estimated sample size per arm, per the protocol
n_R	A re-estimated sample size per arm
N_T	True unknown sample size per arm necessary to accomplish a study's objectives
Power	A statistical measure, based on sample size, of the ability of a test to have statistical significance
Sample size determination	Specifying a study's sample size prior to the start of the study
Sample size re-estimation (SSRE)	Revising a study's sample size during the study
Statistical Analysis Plan (SAP)	Plan for the final statistical analysis

does not distinguish between savings due to early termination vs. sample size re-estimation. Such presentations make adaptive designs very appealing but the potential for sample size re-estimation to result in a waste of resources as discussed below is rarely mentioned.

Sample size is an important aspect of study design. It is only as good as the assumed values upon which it is based. Studies are usually designed with fixed sample sizes based on imperfect information. Consider a two-arm clinical study to compare the means of two normal distributions with a common unknown variance. The correct sample size depends on a parameter, the clinically significant difference, for which the power is to be controlled, as discussed in Bristol (1989, 1995). The value of this parameter, typically based on clinical as well as historical and financial considerations, is not easily specified, as is the value of the variance, which is unknown and unknowable.

Sample Size Re-Estimation (SSRE)

The concept of sample size re-estimation (SSRE), based on estimation of an unknown variance using interim data, goes back at least to Stein (1945). Renewed interest in the 1990s, such as Wittes and Brattain (1990), Shih (1992) and Bristol (1993), among many others, was a motivation for FDA (2010). With further interest and following a plethora of publications, FDA (2019), a revision of FDA (2010), was issued. Interest and research continue with more publications, books and software.

Studies using SSRE are designed with a Final Analysis (FA) and a planned Interim Analysis (IA), which must be specified in the protocol. In some cases, an amendment may be necessary so the IA still qualifies as "planned," even if not envisioned at the onset. The fixed sample size for the FA, here denoted N_A to note that it is based on assumed values, is determined using techniques appropriate for presence of an IA. This article discusses a two-arm, double-blind study with balanced randomization; other designs can be handled similarly. The IA will be performed after data for n_0 subjects are available for analysis. This sample size should be large enough to provide sufficient information and yet small enough to incorporate the information in a timely manner, so n_0 is often set at 0.5 N_A. However, the strategy of planning a study using an optimistic scenario to obtain a small sample size on the assumption that it can be increased later based on interim assessments may not provide the desired result.

The sample sizes discussed in this article are per arm, so the target enrollment is $2N_A$, and the IA is performed when data from $2n_0$ subjects are available for analysis. The latter is assumed to be n_0 subjects from each arm, since the IA is to be performed while the study is blinded. In the following discussion, topics specified to be presented in the protocol could be presented in the Statistical Analysis Plan (SAP) or Interim Analysis Plan (IAP) with additional details.

Unblinded results are usually used for SSRE, but blinded techniques have also been proposed, such as by Bristol and Shurzinske (2001). Blinded SSRE is usually performed using the clinically significant difference and the variance estimated using the interim data to re-estimate the sample size.

One approach to unblinded SSRE is to use the same technique as used for N_A, except the estimates of the treatment difference and the variance observed at the IA are used. The rationale is that the observations at the IA are based on the data and thus may be better than the assumed values. Other approaches to unblinded SSRE using various weighted techniques have been proposed.

The re-estimated sample size, denoted n_R in this discussion, must satisfy $n_R > n_0$. Although rare, $n_R < n_0$ may occur and should be addressed in the protocol. For example, if $n_R < n_0$, the

study could be stopped with n_0 as the revised sample size; alternatively, a larger value could be used.

A maximum re-estimated sample size (n_{max}) must be specified in the protocol, typically based on budgetary concerns. The re-estimated sample size must satisfy $n_R<n_{max}$. This maximum must exceed N_A and is often chosen as a multiple of N_A.

If the calculated value of n_R satisfies $n_R>n_{max}$, the study sponsor has a decision to make. It has already decided that n_R cannot exceed n_{max}, but continuing with n_{max} will have insufficient power. If n_R is not much larger than n_{max}, then n_R might be a reasonable option. Another alternative is to stop the study for futility after the IA if $n_R>n_{max}$. A common procedure is to conduct test for futility or superiority and then perform SSRE if neither test is rejected. This initial testing lessens the possibility of $n_R<n_0$ or $n_R>n_{max}$. If used, this procedure must be planned and specified in the protocol, as should the value of n_{max} and the action to be taken if $n_R>n_{max}$.

Many protocols specify that the study will continue with the planned sample size N_A if n_R is less than N_A. This apparent waste of resources is based on guidance in FDA (2010) that these "methods should be used only for increases in the sample size, not for decreases." Although the rule in FDA (2010) is still often used, this requirement is not in FDA (2019) and should be avoided as a waste of resources.

An optimistic N_A may be desirable for financial reasons or to garner support for the study, especially if the rule in FDA (2010) is applied. If necessary, sample size can be increased based on the IA. However, the increase may be larger than necessary, as shown below.

An Interim Analysis May Lead to a Waste of Resources

The above discussion seems to encourage the use of SSRE to conserve resources, but the opposite can also be true, which is rarely mentioned. The true treatment difference and a true variance are unknown and unknowable parameters. (Many nonstatisticians have difficulty differentiating between parameters, estimates and assumed values of the parameters. Experiments are conducted to estimate parameters and to test corresponding hypotheses.) There is a true unknown sample size (N_T) corresponding to the true unknown parameters. Both N_A and n_R are estimates of N_T, with N_A using assumed values of the parameters and n_R using estimates from the IA. n_R varies randomly and may be smaller or larger than N_T. If it is larger, a significant result will be observed with high probability, but at excessive cost. It is often assumed that a study's success depends on using n_R instead of N_A because N_A was too small, but that success may result from $n_R > N_T$, which is impossible to know because N_T is unknown. In other words, increasing the sample size based on an IA may waste resources.

Simulations

Simulation can assess the performance of SSRE based on a distribution of true parameter values. The author performed various SSRE simulations. While the details are not presented, results are presented for three cases. The value of $N_T=393$ was determined using the ratio of the true difference to the true standard deviation equal to 0.2. Various values of N_A, n_0 and n_{max}, which are specified by design, were used.

> If N_A =250 and n_0 =100, n_R exceeds N_T in about 50 percent of the cases and 750 in about 40 percent of the cases.
>
> If N_A =800 and n_0 =500, n_R =500 in about 65 percent of the cases and exceeds 1,200 in about 10 percent of the cases.

If N_A =393 and n_0 =200, n_R exceeds 1,200 in about 20 percent of the cases.

SSRE is typically designed for the first case, with an optimistic value of N_A used and the hope that SSRE will provide an appropriate larger value. However, in this simulation, SSRE often resulted in a value that was too large, sometimes more than twice as large as necessary.

In the second case, the sample size for the IA, and thus n_R, was larger than the true value.

In the third case, the sample size using the assumed values is equal to the true value, yet n_R may be more than three times larger than the true value. This case is the most surprising and is evidence of a possible inefficiency of SSRE.

Conclusions

SSRE has become popular as a way to conserves resources. However, SSRE has a large associated risk that the re-estimated sample size may be larger than necessary. This over-estimation may result in numerous subjects unnecessarily enrolled. Fortunately, this waste of resources will increase the power beyond the planned value, so clinical studies will still be successful. However, some SSRE increases may not be large enough, wasting resources without achieving the necessary power. These inefficiencies cannot be identified because of the dependence on the unknown correct sample size. These observations regarding possible SSRE inefficiencies are not intended to discourage the use of SSRE, but to inform the reader of potential issues.

References

Bristol, D.R., "Sample Sizes for Constructing Confidence Intervals and Testing Hypotheses," *Statistics in Medicine 8*, 803-811, (1989).

Bristol, D.R., "Sample Size Determination Using an Interim Analysis," *Journal of Biopharmaceutical Statistics 3(2)*, 159-166, (1993).

Bristol, D.R., "Delta: The True Clinically Significant Difference to Be Detected," *Drug Information Journal 29*, 33-36, (1995).

Bristol, D.R. and Shurzinske, L., Blinded Sample Size Adjustment, *Drug Information Journal 35*, 1123-1130, (2001).

FDA, "Guidance for Industry: Adaptive Design Clinical Trials for Drugs and Biologics," February 2010.

FDA, "Adaptive Designs for Clinical Trials of Drugs and Biologics: Guidance for Industry," November 2019.

Shih W.J. (1992). Sample Size Re-estimation in Clinical Trials in *Biopharmaceutical Sequential Statistical Applications* (K. Peace, ed.) New York, N.Y.: Marcel Dekker, 285-301.

Stein C. (1945). A Two-Sample Test for a Linear Hypothesis Whose Power is Independent of the Variance, Annals of Mathematical Statistics 16, 43-58.

Tufts CSDD Impact Reports (2013) The Adoption and Impact of Adaptive Trial Designs (csdd.tufts.edu, Accessed 1 July 2013).

Wittes J. and Brittain E. (1990). The Role of Internal Pilot Studies in Increasing the Efficiency of Clinical Trials, Statistics in Medicine 9, 65-72.

Author

David R. Bristol, PhD, is president, Statistical Consulting Services. Contact him at 336.293.7771 or david@statistical-consulting-services.com.

