Nonconventional Clinical Trial Designs

Approaches to Provide More Precise Estimates of Treatment Effects With a Smaller Sample Size, but at a Cost

George Howard, DrPH

Abstract—Statistical sciences have recently made advancements that allow improved precision or reduced sample size in clinical research studies. Herein, we review 4 of the more promising: (1) improvements in approaches for dose selection trials, (2) approaches for sample size adjustment, (3) selection of study endpoint and associated statistical methods, and (4) frequentist versus Bayesian statistical methods. Whereas each of these holds the opportunity for more efficient trials, each are associated with the need for more stringent assumptions or increased complexity in the interpretation of results. The opportunities for these promising approaches, and their associated “costs,” are reviewed. (Stroke. 2007;38[part 2]:804-808.)

Key Words: adaptive trial design ■ Bayesian statistics ■ clinical trials ■ frequentist statistics

B iostatistics and clinical trial methodology continues to evolve and offer opportunities for more efficient clinical trials, directly translating into increased statistical power, smaller sample sizes, or both. These advancements, however, come at the cost of additional statistical assumptions, logistical challenges to the conduct of the trials, or increased complexity in the interpretation of results. There are many recent statistical advances, and the selection of which to include herein is clearly a subjective decision. Four areas warrant discussion in the stroke arena: (1) improvements in dose selection trials, (2) approaches for adapting the sample size during the conduct of the trial, (3) selection of scales for end points, and (4) Bayesian statistical methods.

Improvements in Approaches for Dose Selection Trials

Selection of the appropriate dose is central to the success of compounds in Phase III trials. The fundamental assumption underlying dosing studies is that both efficacy and the risk of “safety end points” (toxicity) increases with dose, and therefore the “optimal” dose is the highest potential dose that has a low rate of safety events. For purposes of acute stroke trials, safety is often defined as having an acceptable rate of hemorrhage or death: for example, a hemorrhage/death rate of 10% or less.

Until recently, the “3+3” algorithm was the most commonly used algorithm for selection of the appropriate dose for acute stroke treatment compounds. This algorithm evaluates the first 3 patients at the lowest step of potential doses and adjusts the dose for subsequent patients as a function of the number of observed “safety events”: (1) if there are no safety events, the dose is increased to the next level for the next 3 patients; (2) if there is exactly 1 safety event, then an additional 3 patients are evaluated at the current dose level, then: (a) the dose is increased if there are no additional safety events (ie, 1 of 6 patients with safety events); (b) the dose is decreased or the process terminated if there are 1 or more events in the second cohort or patients (ie, 2 or more safety events in 6 patients.); (c) if there are 2 or more events then the dose is decreased or the process terminated.

This process is repeated until a dose is revisited twice, and then the procedure is terminated. Importantly, this algorithm was proposed by cancer researchers to find the dose for which 50% of the patients would experience a safety event. This is in contrast to acute stroke trials where a rate of “safety events” above 10% might be considered excessive.

The “3+3” algorithm is poorly suited to select a dose when the acceptable safety event rate is low. The root of the issues arises from the dichotomous outcome of safety events (ie, event versus no event) and fortunately the relatively low acceptable frequency of these events. Specifically, after 3 patients with no events the algorithm will step up to the next dose despite 95% confidence limits from 0.0% to 70.8%, in essence a decision based on virtually no information. Likewise, assume there is a single event in the initial 3 patients at a dose, and that the required 3 additional patients are performed without an event. In this case, there will be 6 patients with 1 event, where the 95% confidence limits on the estimated proportion would range 0.4% to 64.1%. Saying that one is 95% sure that the estimate is between 0.4% and 64.1% is fundamentally providing no meaningful information re-
Regarding the likelihood of an adverse event; however, the algorithm still steps up to the next dose. Put simply, even with 6 individuals in the dose strata there is virtually no information provided by the study. As a corollary, Figure 1 shows the likelihood of “stepping up” to the next higher dose as a function of the true probability of a safety event for individual patients. As shown, there is still as high as a 50% chance of “stepping up” if the true rate of a safety event is as high as 30%, and the chance of stepping to the next higher dose level only falls substantially below 20% if the true rate of a safety event for patients is 50% or higher (as in the context where the algorithm was originally proposed.) This implies that the dose selected by the “3/H11001” algorithm is likely to be substantially higher than the true “safe” dose, which could easily contribute to a conclusion in a Phase III trial that the drug is not efficacious where the drug may be beneficial at lower dose levels.

The most promising innovation to address this shortcoming is the Continual Reassessment Method (CRM). Briefly, the CRM assumes a parametric monotonic increasing risk function relating dose to safety event risk. Commonly used functional forms are hyperbolic tangent models (as in Figure 2) or a logistic model. There have been several alternative approaches proposed to implement the CRM; however, all share the goal of solving for the parameters of the assumed underlying risk function, and then selecting the dose associated with the acceptable level for safety events. The steps of the CRM algorithm are described as follows:

1. An initial estimate of the parameters determining the underlying risk function is calculated. A Bayesian approach can provide the initial estimate from previous information (such as animal data), or an empirical approach can be used for a frequentist approach.
2. Using the estimated risk function, the dose associated with the targeted event rate is calculated.
3. A small number of patients are treated at the estimated dose.
4. The risk function is re-estimated using the data from these new patients, and an estimate of the dose is revised.
5. There are several alternative stopping rules for the algorithm. For example, Piantadosi has suggested stopping if there is less than a 10% difference between successive drug doses. If the algorithm continues, the process returns to Step 3.

An advantage of the CRM approach is that data from all treated patients contribute to the estimated underlying function, thereby providing improved precision over approaches where estimates of the event rates based only on patients at a specific level (as in the “3/H11001” algorithm). In addition, the CRM naturally allocates patients to dose levels in the neighborhood of the desired dose level, not “wasting” patients treated at doses far below or above the targeted dose. Finally, unlike the “3/H11001” algorithm, the CRM algorithm provides an unbiased estimate of the targeted dose, implying that, on average, the correct dose will be used in Phase III trials.

There are costs associated with these gains, though. Perhaps the greatest is the required assumptions of the underlying risk function, where choosing the incorrect function can result in providing an incorrect estimate of the targeted dose. Although there are approaches that provide more flexibility in the risk function (for example, using a 2-parameter logistic function rather than a 1-parameter logistic function), these approaches decrease the precision of estimates at a fixed sample size.

Approaches for Sample Size Adjustment
Selection of target sample size is a critical factor in the design of any randomized trial. Although calculations are straightforward, the underlying assumptions are frequently based on incomplete data. Relatively small errors in the assumptions can design a trial that is either substantially under- or overpowered.

One of the more rapidly evolving clinical trial methodological areas is that of adaptive clinical trials, where the study design is adjusted based on data collected as part of the initial conduct of the trial. Methods for adaptive trials are sometimes confused with the much more mature methodological area of group sequential trials. Shih eloquently relates that group sequential trial methodology has the goal of “saving lives or saving resources” whereas the adaptive clinical trial approach has the goal of “saving the study.”
There are 2 approaches of adaptive trial methodology. The first examines estimates of parameters that were assumed during the design phase of the project (such as the variance of the outcome measure or the overall event rate) that can have a major impact on the projected sample size.\textsuperscript{11–15} The second approach focuses on the estimated treatment difference between groups from an interim analysis.\textsuperscript{16–17} Adaptive trial methodology permits the use of these 2 types of information to adjust the sample size (or other study parameters) while protecting the overall \( \alpha \) level.

For a simple example, assume that a trial has incident stroke as a primary end point, and that during the design phase it was assumed that 5% of study patients will experience a stroke over a 2-year follow-up. However, suppose that when half the patients reach the 2-year end point, only 4% of the patients are experiencing a stroke. Because the power of the study is affected by the number of events (not the number of patients), without an adjustment to the sample size the study will be underpowered. Alternatively, assume that a study is performed with the primary end point of the mean Rankin Scale at 90 days, and during the design phase the standard deviation of this score was assumed to be \( \pm 0.5 \). However, after half the patients reach this point the standard deviation is observed to be 0.45. Because of this smaller variance in the outcome, the study is overpowered and will not require nearly as many patients. In neither of these examples was the sample size adapted based on a treatment difference but rather a “design (or nuisance) parameter.” A more complex situation is the case where a study was found to detect a 30\% treatment effect between the placebo and active treatment groups; however, after half the patients have been studied a difference of only 20\% is noted. If that treatment difference persists, the study will be underpowered with adapting the sample size; however, in this case the sample size could be very carefully adapted based on information directly incorporating the interim treatment efficacy.

One of the most frequently used applications of adaptive trial design is in “two stage” studies,\textsuperscript{10,18} where a study is begun with a prespecified sample size for an interim analysis but not a prespecified total sample size. In adaptive trial design, at the time of the interim analysis: (1) if a “large” treatment effect is observed, the study can be terminated with the rejection of the null hypothesis; (2) if a “small” treatment effect is observed the study is terminated while declaring the treatment futile; or (3) if there is an intermediate treatment effect, the study is continued where the number of patients in the second stage is determined as a function of the observed treatment effect from the first stage (the study sample size is not specified until after the interim analysis). Methods to implement this “two stage” approach have been developed for trials with continuous\textsuperscript{17} and time-to-event\textsuperscript{19} outcomes.

Adaptive trial methodology comes at a cost, with the greatest “cost” being the uncertainty of study sample size. Under the current clinical trial review process at the National Institutes of Health (NIH), it may be difficult to obtain funding for a clinical trial where the sample size is an underdetermined random variable that can have considerable leeway for variation. Methods have been developed to restrict the range of variation in the study sample size; however, the value of the adaptive design approach decreases with increasing restrictions on the ability to adapt.

### Selection of Study End Point and Associated Statistical Methods

The recently reported SAINT trial has generated considerable excitement and discussion regarding using “shift analysis” to assess treatment efficacy: for example, using outcomes such as the Rankin Scale.\textsuperscript{20} By using techniques including ordinal logistic regression, the SAINT investigators are to be congratulated for breaking a barrier that had limited the analysis of Rankin Scale data to a dichotomous scale. This breakthrough provided a substantial gain in power or reduction in sample size. Power is generally increased as the outcome moves from a dichotomous to an ordinal scale and from an ordinal to a continuous scale. The Rankin Scale is an excellent example where alternative metrics can be reasonably considered. These same concepts would apply to any other ordinal outcome such as Glasgow Coma Scale. Specifically, Rankin Scale data can be considered as: (1) a dichotomous outcome where patients are divided into 2 groups on the basis of their outcome scores and data are analyzed most frequently using \( \chi^2 \) analysis or ordinary logistic regression; or (2) an ordinal outcome where the Rankin Scale scores from 0 to 5 are maintained but there is no assumption that the difference in outcome between adjacent scores is similar and data can be analyzed using methods such as the proportional odds models; or (3) a continuous outcome where the Rankin Scale is considered as a general 0 to 5 scale and treatment differences are described by differences in mean scores and are analyzed most frequently using general linear models approaches (eg, \( t \) test, analysis of variance, or regression techniques) sometimes after ranking or “RIDIT Scoring” to reduce the assumption associated with distance measures.

This increased power comes at a cost, primarily the need for an increasing number of assumptions regarding the nature of the outcome data and increasing complexity in the interpretation of the results. Considering the Rankin Scale score as a dichotomous outcome has major advantages. Specifically, virtually no assumptions are required and the interpretation of the data are quite straightforward (ie, just differences in the percent below the threshold); however, information is lost and statistical power is reduced when data are categorized.

As noted by the SAINT investigators, considering the Rankin Scale as an ordinal outcome has the major advantage of capturing information lost by categorization. In the example of the Rankin Scale, this loss is potentially clinically important because data from the Duke Stroke PORT suggests that the increase in quality adjusted life-years is nearly linear across the entire response scale\textsuperscript{21} (see Figure 3). Capturing this information pays the reward of providing additional precision and power (ie, smaller sample size). The most common analysis approach for such data (and used by the SAINT investigators) is the proportional odds logistic regression model. This model requires the assumption that the impact of treatment on the odds of having an outcome of 0 relative to an outcome of 1 or greater is the same as the impact of treatment on the odds of having an outcome of 1 or less relative to a 2 or greater, and this is the same as having...
an outcome of 2 or less relative to 3 or greater, and so on. This approach has the additional advantage of providing a clinically interpretable outcome, the estimate of the odds across the scale of disease severity. However, this also requires the assumption that a treatment has a similar effect across the entire spectrum of disease severity. Fortunately, there is a statistical test of the “proportional odds” assumption, but unfortunately this test has a notoriously low power to detect deviations from the assumption. Furthermore, if the majority of the patients fall into only 1 or 2 of the Rankin Scale scores, the advantage of the approach is reduced and the power to detect deviations from the proportional odds estimate is even further reduced. Importantly, the assumption of proportional odds underlies both analytic technique and the clinical interpretation of the results. Alternative ordinal data analysis, such as the Cochran-Mantel-Haenszel approach, also used by the SAINT investigators, makes a similar (albeit implicit) assumption of consistency of effects across strata. As such, if this assumption is not correct, then the resulting estimate does not appropriately describe the effect of the treatment.

As the number of categories of ordinal data increases, the distinction between ordinal and interval data begins to blur. Generally, if there are 4 or 5 levels of an ordinal outcome it can reasonably be considered as an interval measure (technically, that the asymptotic normal properties required by general linear models begin to be reasonable assumptions). Fortunately, linear models are a “robust” statistical procedure, meaning that even moderate violations of the assumptions are not all that important. Considering the Rankin score as continuous data captures even more information than considering it as an ordinal outcome, and as such will generally have greater precision and statistical power. However, this increased power comes at the price of making even more assumptions, specifically that there is an equal “distance” between the outcome categories (in this case, linear relationship with quality adjusted life-years may imply this is a reasonable assumption). In addition, the outcome of a “mean Rankin Score” may not be as clinically interpretable.

The increase in assumptions and challenges in interpretation does not mean that considering outcomes as an ordinal or continuous outcome should not be done, just that the advantages and disadvantages need to be weighed and the best choice for specific trials selected.

**Frequentist Versus Bayesian Statistical Methods**

One of the most confusing statistical topics to the clinical investigator is the long-standing “debate” between Bayesian and frequentist statistical approaches. In a frequentist approach, the difference between treatment groups is assumed to be an unknown and fixed parameter (ie, the difference in proportions responding for dichotomous outcomes, the difference mean score for continuous outcomes, or the ratio hazards for time-to-event outcomes). In contrast, a Bayesian approach considers the parameters not as unknown constants, but characterized by a distribution of potential differences between treatment groups. There may be information describing the differences between groups before the trial is done (ie, prior distribution), and this information is updated based on additional information on the differences between the groups obtained from the trial (ie, the posterior distribution). The advantage of the Bayesian approach is the opportunity to include information that is known from other studies before conducting the trial. Much like individuals “changing their opinion” to reflect new knowledge, the approach allows for changing the description of likely differences between treatments to reflect the outcome of the trial (ie, the “posterior” distribution).

There are strong advocates of both approaches. Individuals supporting Bayesian methods have argued that the approach allows for the appropriate “blending” of information from Phase I, II and III trials, allowing for a smooth transition of learning as data accumulates from all sources. At the same time, frequentists express caution for the use of these techniques in Phase III trials. In contrasting these approaches: (1) it is important to recognize that both approaches are only tools for assessing treatment differences, and because there is no tool that is universally “best” for all jobs, both approaches have much to offer; (2) it is quite important to recognize that whereas there are philosophical differences in the underlying assumptions, differences in the performance of the approaches and the domains where they can be applied are not as large as are sometimes portrayed. For example and as noted in our previous editorial, there are numerous approaches for adaptive design available using both Bayesian and frequentist approaches.

Although there are not “magic bullets” to universally provide improved clinical trials in either the Bayesian or frequentist approach, the differences in these approaches should be considered in the spectrum of tools that are useful in different scenarios.

The challenge to having alternative approaches is that it is not always clear which to use. Questions that appear straightforward on the surface, such as “Should we use the information that is available from previous studies?” may be deceptively complex. It seems obvious that it would be wasteful not to use all information that is available in making a decision; however, if some of the information is wrong (biased), then using it will lead to incorrect conclusions. Like so many other areas in life, perhaps the best answer to whether to use...
Bayesian or frequentist approaches is “it depends.” Although this position is hardly fulfilling to the individual who is seeking a uniform best approach for trial design, it avoids the oversimplification of what in truth is a complex decision.

Summary
There are exciting advancements in trial design and analysis that permit decisions to be made based on more efficient methods that provide either better precision or reduced sample size (or both). In general, these advances come at the price of requiring additional assumptions (which if false may imply that incorrect decisions are made), increasing the complexity of interpretation, or introducing logistical hurdles in the implementation of the trial. Does this mean that these improved approaches should not be used? Obviously not, but the consideration of these approaches needs to be made after weighing the advantages and disadvantages of the alternatives, and the decision need not be the same for all trials (not even for trials that are reasonably similar). Perhaps the innovation required for trial design and conduct is thoughtful consideration of alternative approaches and an open-minded presentation of the results including the limitations introduced by the selected trial design approach.

Disclosures
Dr. Howard is on advisory boards for Bayer, CoAxia, PhotoThera, and KOS. Furthermore, Merck acted as the expert witness.

References
Nonconventional Clinical Trial Designs: Approaches to Provide More Precise Estimates of Treatment Effects With a Smaller Sample Size, but at a Cost
George Howard

*Stroke*. 2007;38:804-808
doi: 10.1161/01.STR.0000252679.07927.e5

*Stroke* is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231
Copyright © 2007 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/38/2/804

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/