Circulation

JOURNAL OF THE AMERICAN HEART ASSOCIATION

American Heart Association®

Learn and Live℠

Optimizing Trial Design: Sequential, Adaptive, and Enrichment Strategies

Cyrus Mehta, Ping Gao, Deepak L. Bhatt, Robert A. Harrington, Simona Skerjanec and James H. Ware

Circulation 2009;119;597-605

DOI: 10.1161/CIRCULATIONAHA.108.809707

Circulation is published by the American Heart Association. 7272 Greenville Avenue, Dallas, TX 72514

Optimizing Trial Design

Sequential, Adaptive, and Enrichment Strategies

Cyrus Mehta, PhD; Ping Gao, PhD; Deepak L. Bhatt, MD, MPH; Robert A. Harrington, MD;
Simona Skerjanec, PharmD; James H. Ware, PhD

Despite substantial progress in the prevention of cardiovascular disease and its ischemic complications, it remains the single largest killer in the United States. New treatment options are needed, particularly to respond to the challenges of an aging population and rising rates of obesity and diabetes. Development of novel therapeutic strategies for the management of acute cardiovascular disease is especially challenging. Specific problems include relatively low event rates, diverse patient populations, lack of reliable surrogate end points, and small treatment effects subject to substantial uncertainty. Because the clinical development process is enormously expensive and time consuming, there is considerable interest in statistical methods that use accumulating data from a clinical trial to inform and modify its design. Such redesign might include changes in target sample size and even changes in the target population. This article discusses developments in adaptive design of interest to cardiovascular research.

To illustrate the methods we discuss, we focus on the development of novel therapies for the management of acute coronary syndromes. However, the ideas we discuss have much wider application. We begin by discussing the traditional approach to determination of a fixed sample size for a clinical trial. We then describe group sequential designs and the benefit they provide in the conduct of trials. In the section on Adaptive Sample Size Reestimation, we discuss designs that are adaptive in the sense that they allow an adjustment of the target sample size based on the accumulating data from the trial. In the section on Adaptive Sample Size Reestimation With Enrichment, we discuss enrichment designs that shift the focus to a patient subgroup when the accumulating data suggest greatest benefit for that subgroup. Additional details about population enrichment are provided in the online-only Data Supplement.

The Setting

We consider therapies intended to reduce the risk of acute ischemic complications in patients undergoing percutaneous coronary intervention. For specificity, we consider a placebo-controlled randomized trial with a composite primary end point including death, myocardial infarction, or ischemia-driven revascularization during the first 48 hours after randomization. We assume, based on prior knowledge, that the placebo event rate is in the range of 7% to 10%. The investigational drug is assumed, if effective, to reduce the event rate by 20%, but the evidence to support this assumption is limited. The actual risk reduction could be larger but could also easily be as low as 15%, a treatment effect that would still be of clinical interest given the severity and importance of the outcomes, but that would require a substantial increase in sample size. As is often true when clinical trials are planned, there is substantial uncertainty about both the placebo-group event rate and the treatment effect. In the following sections, we describe how 4 design strategies perform in such situations. We assume throughout that patients are randomized in equal proportions to the experimental and placebo arms. Because the primary objective is to determine whether the new treatment is superior to placebo, our focus will be on 1-sided hypothesis testing.

Fixed Sample Size Designs

The simplest and most common method for determining sample size in the presence of uncertainty is to take best estimates, sometimes based on limited information, of both the placebo group event rate and the treatment effect, apply one of the standard sample size formulas, and perform a fixed sample size trial with that target enrollment. Suppose that π_c represents the event rate for the placebo arm, π_e the event rate for the experimental arm, and $\rho=(\pi_c/\pi_e)$ the relative risk in the treatment and control groups. We are interested in testing the null hypothesis, H_0, that the event rates do not differ in the treatment and placebo arms ($\rho=1$). If we employ a 1-sided level-α test, the combined sample size N (both arms) needed to achieve power $1-\beta$ to reject the null hypothesis when ρ is <1 is

$$(1) \qquad N=2\left[\frac{1-\rho\pi_c}{\rho\pi_c}+\frac{1-\pi_c}{\pi_c}\right]\left[\frac{z_\alpha+z_\beta}{\mathrm{In}(\rho)}\right]^2$$

where z_α and z_β are the appropriate percentiles of the standard normal distribution. For example, when $\pi_c=8\%$, the com-

From the Harvard School of Public Health, Boston, Mass (C.M., J.H.W.); Cytel Statistical Software and Services, Cambridge, Mass (C.M.); The Medicines Company, Parsippany, NJ (P.G., S.S.); Veterans Affairs Boston Healthcare System and Brigham and Women's Hospital, Boston, Mass (D.L.B.); and Duke Clinical Research Institute, Durham, NC (R.A.H.).

The online-only Data Supplement is available with this article at http://circ.ahajournals.org/cgi/content/full/119/4/597/DC1.

Correspondence to Cyrus Mehta, PhD, Cytel Statistical Software and Services, 675 Massachusetts Ave, Cambridge, MA 02139. E-mail mehta@cytel.com

(*Circulation*. 2009;119:597-605.)

***Circulation* is available at http://circ.ahajournals.org** **DOI: 10.1161/CIRCULATIONAHA.108.809707**

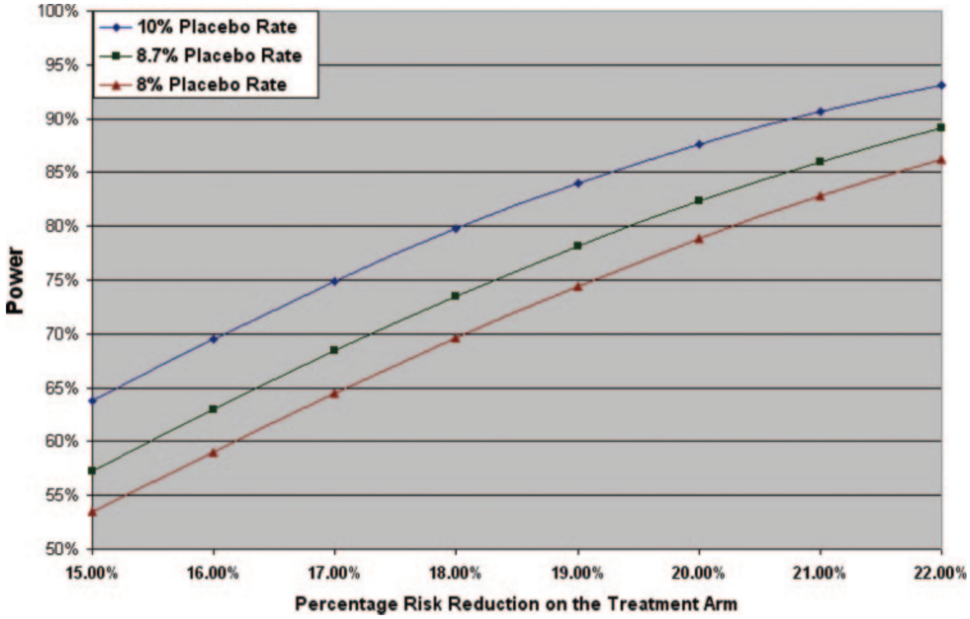


Figure 1. Power curves of a fixed sample size design for 3 different placebo-group event rates and N=8000.

bined sample size needed to detect a 20% risk reduction (ie, a relative risk of ρ=0.8) with 80% power with a 1-sided level-0.025 test is calculated from Equation 1 as N=8236.

This approach has 2 important limitations: The power varies as a function of both the placebo group event rate and the magnitude of the treatment effect. For instance, if π_c=7% instead of 8%, the sample size needed to detect a 20% risk reduction with 80% power increases from N=8236 to N=9503. If, in addition, the actual risk reduction is only 18%, the required sample size increases even further to N=11 841. Figure 1 displays the power of an 8000-patient study for 3 different placebo-group event rates and risk reductions between 15% and 22%, in which the risk reduction is 100×(1−ρ)%. The power of the study varies from 53% to 93%. In subsequent sections, we focus on the middle curve, corresponding to an 8000-patient fixed sample trial with an 8.7% placebo-group event rate. For this event rate and a relative risk of ρ=0.8, 8000 patients yield power of 82%.

When we design a trial with a fixed sample size based on an assumed placebo-group event rate and treatment effect, we are at risk of mounting a trial that is underpowered for the actual situation. When the end point is measured as the time to the event rather than the event rate, the dependence of power on prior knowledge of the placebo-group event rate can be eliminated by continuing the trial to a prespecified number of events, D. The number of events needed for a 1-side level-α test to detect a hazard ratio ρ with 1−β power is then given by

$$(2) \qquad D=4\left[\frac{z_\alpha+z_\beta}{\text{In}(\rho)}\right]^2$$

(See, for example, Rao and Schoenfeld.[1]) Unlike Equation 1, Equation 2 does not depend on the control group event rate.

When the end point is binary, that is, the occurrence or nonoccurrence of the event rather than the time to the event, the event-driven approach does not eliminate dependence of sample size on the control group event rate. Event-driven trials are, however, a special case of a more general approach for trial design known as the information-based design. In an information-based design, the statistical information rather than the sample size is fixed in advance. The trial continues to enroll patients until the desired amount of statistical information is attained. Prior specification of π_c at the design stage is thereby eliminated. The information-based approach is described in greater detail in the online-only Data Supplement and elsewhere.[2]

Neither event-driven trials nor their generalization to information-based trials is helpful when the treatment effect is different than anticipated at the planning stage. If it is smaller than anticipated, the study can fail even though the investigational drug is beneficial. If it is larger than anticipated, the study can enroll more patients than would have been needed to demonstrate superiority of the new drug. In the remainder of this article we discuss more flexible group sequential and adaptive designs, as well as enrichment strategies, that address both the uncertainty about the placebo-group event rate and the uncertainty about the treatment effect.

Group Sequential Designs

By the 1960s, investigators recognized both the risk of ignoring trends and the risk of responding too quickly to trends observed in comparisons of treatment groups during the course of an ongoing trial. Seminal articles by Pocock[3] and O'Brien and Fleming[4] described statistical methods for determining critical values that could be used to analyze accumulating data at a fixed number of prespecified occasions. These group sequential designs permit early stopping for efficacy based on interim analyses of the accumulating data without inflating the false-positive rate, also known as the type 1 error rate. The O'Brien-Fleming group sequential designs have been used widely because they are more

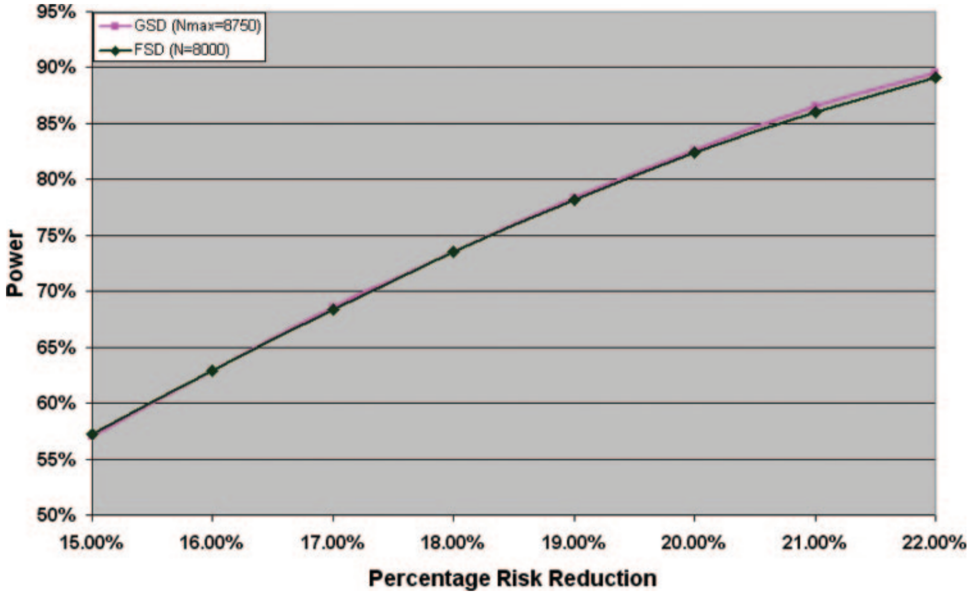


Figure 2. Power curves for the fixed sample-size design with N=8000 and the group sequential design with N_{max}=8750.

conservative than other designs in early analyses and therefore reduce the risk of stopping early on the basis of a strong trend based on limited experience. Lan and DeMets[5] introduced the idea of an α spending function that allows Data and Safety Monitoring Boards to modify the timing of interim analyses based on logistical and other considerations, while preserving α, the allowable type 1 error. Formally, an α spending function is an increasing function of sample size that specifies the cumulative type 1 error that may be expended at any interim analysis time point. It assumes the value 0 at sample size 0 and the value α at the maximum sample size, thereby ensuring that the overall type 1 error of the trial cannot exceed α.

Group sequential designs provide a framework that allows Data and Safety Monitoring Boards to stop trials early when a sufficiently strong trend is observed without compromising the statistical strength of the evidence. One can also build in futility stopping rules for terminating a trial if interim results suggest that a positive outcome would be unlikely[6] if the trial were continued. To achieve the same power for a given treatment effect, studies with an interim analysis plan must have a larger maximum sample size than the fixed sample size design, but the expected sample size, accounting for the gain achieved by early stopping, will typically be smaller for sequential designs than for fixed sample size designs.

To illustrate, consider a group sequential design with interim analyses after 50% and 70% of patients are enrolled, making a total of 3 analyses (including the final analysis). Using standard group sequential software,[7,8] one can show that an O'Brien-Fleming design with an overall one sided type 1 error of 0.025 would achieve statistical significance at the first look if the nominal 1-sided probability value was <0.0015, would achieve statistical significance at the second look if the nominal 1-sided probability value was <0.0069, and would achieve statistical significance at the third and final look if the corresponding nominal 1-sided probability value was <0.0227. In contrast, a single (final)-look level-0.025 design would achieve statistical significance if its nominal 1-sided probability value at the end of the study was 0.025.

Now suppose that we add a futility stopping rule whereby the trial will be terminated early if the interim results are trending in the wrong direction. Any reasonable futility rule can be imposed because early stopping for futility cannot inflate the type 1 error rate. Stopping for futility will, however, decrease power because it increases the chances of a false-negative outcome. To offset this power loss, one must increase the sample size of the group sequential design. Suppose, for example, that we decide to terminate for futility at the first interim analysis if the estimated risk of an event is 1% greater for the experimental treatment than for placebo, and at the second interim analysis if the conditional power (or probability of a successful outcome at the end of the trial given the current data and assuming that the current estimate of the treatment effect is the true effect in the population) is $<20\%$. Suppose also that the placebo-group rate is 8.7%. In that case, Figure 2 shows that, to match the power curve of the 8000-patient fixed sample size design, the maximum sample size of the 3-look group sequential design must be increased to 8750 patients.

Having equated the power curves of the fixed sample and group sequential designs, we can compare the 2 designs with respect to average sample size. Figure 3 displays average sample sizes for different values of percent risk reduction over the relevant range. The average sample size for the fixed sample design is constant throughout with N=8000. Although the group sequential design requires an upfront commitment of N_{max}=8750 patients, its average sample size is substantially lower because of the possibility of early stopping. The average sample size of the group sequential design ranges between 6400 and 7000 patients, achieving smaller expected sample size than the fixed sample size design by ≈1000 patients when the risk reduction with the investigational drug is ≈15% and by ≈1600 patients when the risk reduction is ≈22%.

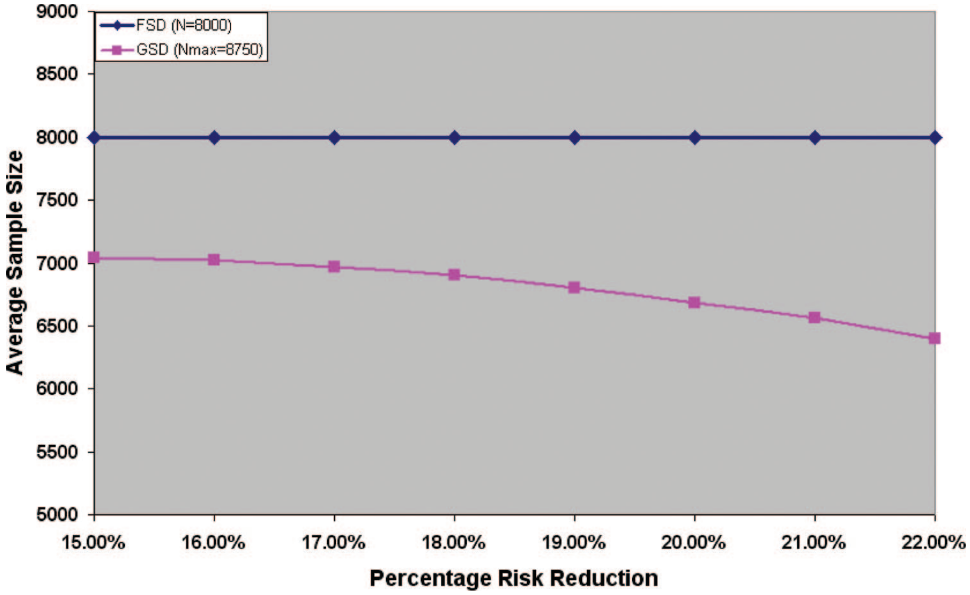


Figure 3. Average sample sizes for the fixed sample size and group sequential designs for different values of the percent risk reduction.

Group sequential designs can produce efficiency gains relative to fixed sample size designs, and they also protect patients when the treatment effect is larger than expected because stronger trends will lead to earlier termination. They do not, however, help in the situation in which the investigational drug is beneficial but with an effect size smaller than that assumed in the initial study design; in this circumstance, the only useful action is a substantial increase in the sample size, a decision that is typically based on financial and logistical considerations.

Adaptive Sample Size Reestimation

Many modern trials in cardiovascular disease require large international networks of investigators and substantial initial investments. The coordination of these trials is complex and resource intensive. Moreover, these phase 3 (or confirmatory) trials come at the end of an already long and costly drug development process. Although the study is initially designed with a particular effect size in mind (say 20%), the drug may be of clinical interest when the effect size is somewhat smaller (say 15%), and investigators may wish to increase the sample size rather than carry on with an underpowered trial when the accumulating data suggest that the effect size is smaller than anticipated. Until recently, statisticians believed that use of accumulated data to modify the trial design would compromise the statistical integrity of the trial. In an important article, however, Müller and Schäfer[9] showed that, at any interim analysis, a group sequential design can be modified to change the number and timing of subsequent analyses without altering the unconditional type 1 error rate, provided that the conditional type 1 error of the modified design is identical to the conditional type 1 error of the original design. The conditional type 1 error of either design is defined as the probability of rejecting the null hypothesis when it is in fact true, given the accumulated data at that analysis point and continuation of the trial according to the original or modified design.

Intuitively, if the conditional type 1 error of the study is preserved by the modified design for every realization of the data, it must be preserved for the study as a whole, and therefore the overall type 1 error is preserved. Thus, if the original interim analysis plan had a type 1 error rate of 0.025, a plan that allows for redesign of the sequential boundary but satisfies the Müller and Schäfer condition will also have a type 1 error rate of 0.025.

The implications of this observation are enormous. If, at an interim analysis, the accumulated data yield a point estimate of effect size that is half the size of the effect employed in the initial study, the investigators have the option of increasing the sample size to achieve desired power against this smaller effect. Of course, this may be impractical or the effect may not be of clinical interest. Still, this represents an important new dimension in trial design that is critical to consider given the complexity and cost of performing large international clinical trials.

To illustrate the Müller and Schäfer principle, consider once again the 8750 patient 3-look group sequential design described in the previous section. This design is displayed schematically in Figure 4. The *x* axis shows the sample size at each of the three looks, and the *y* axis shows the corresponding efficacy and futility boundaries on the standardized normal scale. To be specific, let $\hat{\rho}_j$ denote the estimate of the relative risk, ρ, and N_j be the combined sample size at look j for j=1, 2, 3. Then the standardized test statistic at look j is computed as

$$Z_j = \frac{\ln(\hat{\rho}_j)}{se[\ln(\hat{\rho}_j)]}$$

where $se[\ln(\hat{\rho}_j)]$ is the standard error of $\ln(\hat{\rho}_j)$ and is estimated by the formula

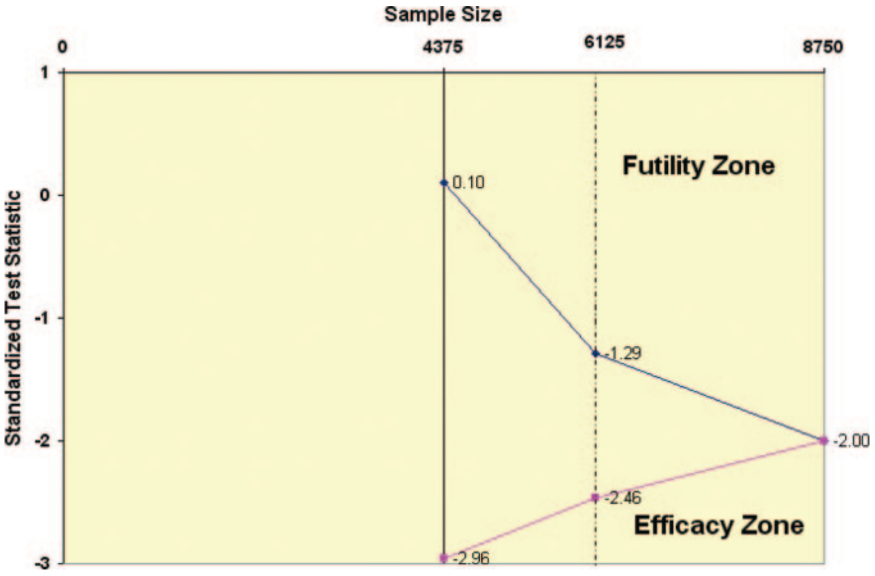


Figure 4. Stopping boundary for the group sequential design with N=8750 and type 1 error=0.025 (1-sided).

$$se[\ln(\hat{\rho}_j)]=\frac{2(1-\hat{\pi}_{cj})}{N_j\hat{\pi}_{cj}}+\frac{2(1-\hat{\pi}_{ej})}{N_j\hat{\pi}_{ej}}$$

Statisticians refer to Z_j as the *Wald statistic*. Under the null hypothesis that the treatment and placebo arms have equal risk, the Wald statistic has a standard normal distribution.

Figure 4 shows that at the first look, when $N_1=4375$, the trial stops for efficacy if $Z_1\leq-2.96$ and for futility if $Z_1\geq0.1$. At the second look, when $N_2=6125$, the corresponding efficacy and futility bounds for Z_2 are -2.46 and -1.29, respectively. At the final look, when $N_3=8750$, the null hypothesis is rejected in favor of the active treatment if $Z_3\leq-2.0$. Otherwise, the null hypothesis is accepted. The unconditional type 1 error, or probability that the Wald statistic will cross one of the efficacy boundaries under the null hypothesis, is 0.025 for this group sequential design.

Example

Now consider a hypothetical trial in which the data at the first and second looks are displayed in the Table. Because the test statistics, $Z_1=-1.6728$ and $Z_2=-1.8816$, do not exceed either the efficacy or futility boundaries, the trial can continue beyond the second look. The estimate of the relative risk at the second look is $\hat{\rho}_2$. That is, the estimated percent risk reduction is only 15.07%. The Table also reveals that $\hat{\pi}_{c2}$, the estimated event rate for the control arm, is 266/3026, or ≈8.7% at the second look. Because these estimates are based on a large sample size, they are good estimates of the true parameter values ρ and π_c, respectively. For these values, the power curve displayed in Figure 2 shows that the unconditional power of this study is only ≈58%. We can also calculate the conditional power or probability that at the final look the Wald statistic, Z_3, will drop below -2.0, conditional on the observed value of $Z_2=-1.8816$ and assuming that $\rho=\hat{\rho}_2$ and $\pi_c=\hat{\pi}_{c2}$. This probability is only 67%. Thus, by either assessment the trial is underpowered.

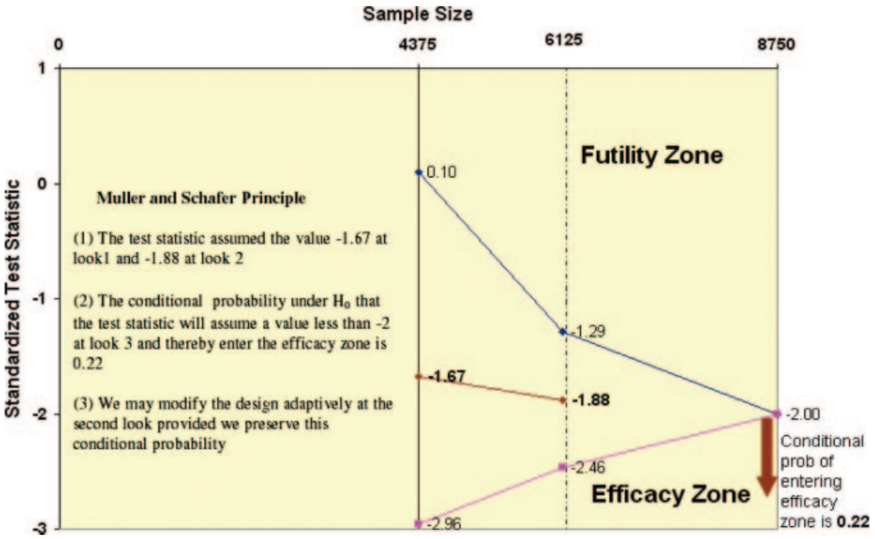
We can increase power by increasing the sample size. Suppose we wish to increase the conditional power from 67% to 80%. With the use of Equation 6 of Gao et al,[10] the final sample size must be increased from 8750 patients to 10 678 patients to achieve this power. However, because the sample size modification is based on the data observed at the second interim analysis, we must invoke the Müller and Schäfer criterion at the final analysis to protect the overall type 1 error. This means that the conditional type 1 error, or conditional probability of a statistically significant outcome under the null hypothesis, given Z_2, must remain the same for the altered and unaltered trials. The conditional type 1 error of the unaltered trial is depicted in Figure 5 as the probability, under the null hypothesis, that the sample path will drop below -2.0 and thereby enter the efficacy zone at the final look. We have calculated this probability to be 0.22.

Figure 6 depicts the altered trial with a sample size of 10 678 patients. For the conditional type 1 error of the altered trial to also be 0.22, the final critical value must be increased from -2.0 to -1.93.[10] Because the original design had an unconditional type 1 error of 0.025, an adaptive trial in which the conditional type 1 error is preserved will also have an unconditional type 1 error of 0.025.

When an adaptive trial is to be implemented, we recommend that the decision rules for making the adaptive change be prespecified and evaluated by simulation. Accordingly, suppose we decide in advance that we will, if necessary, increase the sample size of the original 3-look group sequen-

Table. Data at Looks 1 and 2 of a 3-Look Group Sequential Trial

Look j	Sample Size N_j	Control Event Rate $\hat{\pi}_{cj}$	Experimental Event Rate $\hat{\pi}_{ej}$	Relative Risk $\hat{\rho}_j$	Log Relative Risk $\ln(\hat{\rho}_j)$	Standard Error $se[\ln(\hat{\rho}_j)]$	Wald Statistic Z_j
1	4375	190/2187 (8.68%)	160/2188 (7.31%)	0.8417	−0.1731	0.103	−1.6728
2	6125	266/3062 (8.69%)	226/3063 (7.38%)	0.8493	−0.1633	0.0868	−1.8816



Muller and Schafer Principle

(1) The test statistic assumed the value -1.67 at look1 and -1.88 at look 2

(2) The conditional probability under H_0 that the test statistic will assume a value less than -2 at look 3 and thereby enter the efficacy zone is 0.22

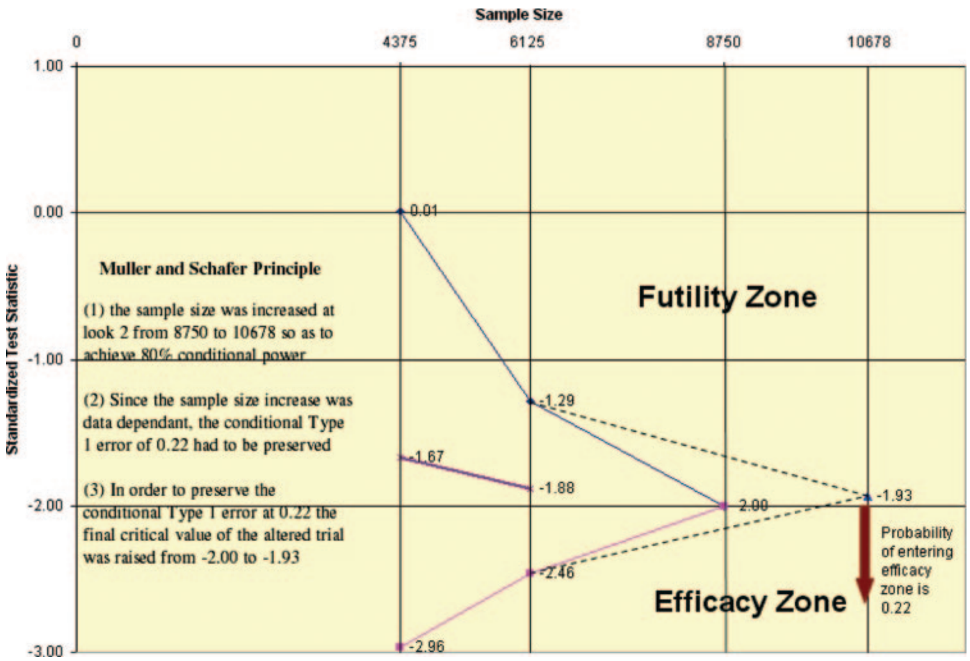
(3) We may modify the design adaptively at the second look provided we preserve this conditional probability

Figure 5. Status of the group sequential design at the second interim analysis. The conditional type 1 error is 0.22.

tial design at the second interim analysis so as to achieve 80% conditional power against an alternative hypothesis equal to the estimated effect size at that analysis. Specifically, after completing the second interim analysis and concluding that the trial should not be stopped for either superiority or futility, we perform a sample size reestimation (subject to a cap) based on the observed effect size. If the conditional power of the study can be increased to 80% for that effect size with a revised total sample size of <15 000, the sample size will be increased to that number. If the power is <80% for a sample size of 15 000, the sample size will be left at the originally specified 8750. This latter feature is not required but establishes a cap on the maximum sample size. The cap reflects consideration of the clinical importance of an assumed effect size, the operational challenges of additional patient recruitment, and the financial limitations of the required enrollment.

The power curve for this adaptive group sequential design can be evaluated by simulation and is displayed in Figure 7 alongside the power curve for the nonadaptive group sequential design with a maximum sample size of 8750 patients and the 8000-patient fixed sample design. The adaptive group sequential design produces a power gain of 3% to 4% across the range of effect sizes of interest.

Figure 8 displays the expected sample sizes for the 3 designs. Both the group sequential and adaptive group sequential designs achieve a substantial reduction in average sample size compared with the fixed sample design. Moreover, even though it was permissible for the sample size of the adaptive group sequential design to increase to as much as 15 000 patients, the average sample size is in the range of 6600 to 7400 patients because of the presence of the efficacy and futility boundaries. We also note from Figure 8 that the



Muller and Schafer Principle

(1) the sample size was increased at look 2 from 8750 to 10678 so as to achieve 80% conditional power

(2) Since the sample size increase was data dependant, the conditional Type 1 error of 0.22 had to be preserved

(3) In order to preserve the conditional Type I error at 0.22 the final critical value of the altered trial was raised from -2.00 to -1.93

Figure 6. Boundaries for the modified group sequential design that preserves the conditional type 1 error at 0.22 and also has power of 0.80 against the effect size observed at the second interim analysis.

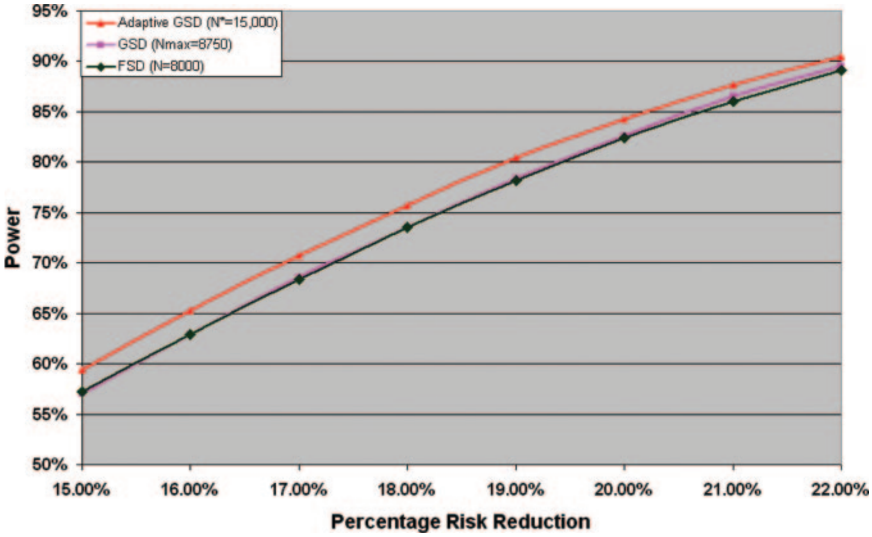


Figure 7. Power curves for the fixed sample size, group sequential, and adaptive designs.

3% to 4% gain in power relative to the nonadaptive group sequential design is achieved at the cost of between 200 and 400 additional patients on average.

Adaptive sample size reestimation will not be appropriate for every situation. It could result in recommendations for substantially larger trials, possibly to pursue effect sizes of limited clinical interest. Thus, it is a tool to be applied judiciously. However, in settings in which a sponsor has made a substantial investment and the investigators find that the apparent effect size, while smaller than initially planned, is still clinically important, it can provide another path to a positive result without compromising the integrity of the trial.

Adaptive Sample Size Reestimation With Enrichment

In some clinical settings, there is reason to believe that the treatment effect might vary among patient subgroups. For example, an intervention might provide greater benefit to patients with comorbid conditions based on the pathophysiology of their disease and the mode of action of the pharmacological agent studied. In such situations, one might like to have an option to enrich the patient population, that is, to selectively enroll the remainder of the patients in the subgroup(s) in which greatest benefit is observed at the interim look.

Example

Let G_0 denote the entire population under study. Suppose that investigators have identified 2 subgroups of patients, G_1, and G_2, such that G_1 is a subset of G_0 and G_2 is a subset of G_1. Let H_i denote the null hypothesis that the treatment and control arms are equally efficacious in population G_i, for $i=0, 1, 2$. We now describe a design that allows both sample size reestimation and enrichment after an interim analysis.

We assume as before that the initial target sample size is 8750 patients and that interim analyses will be performed at 50% and 70% of that target. We also assume that the sample size reestimation at the 70% mark will first be attempted without enrichment, as described in the section Adaptive Sample Size Reestimation, so as to achieve 80% conditional power with the G_0 population. However, if the sample size reestimation for G_0 produces a revised sample size >15 000,

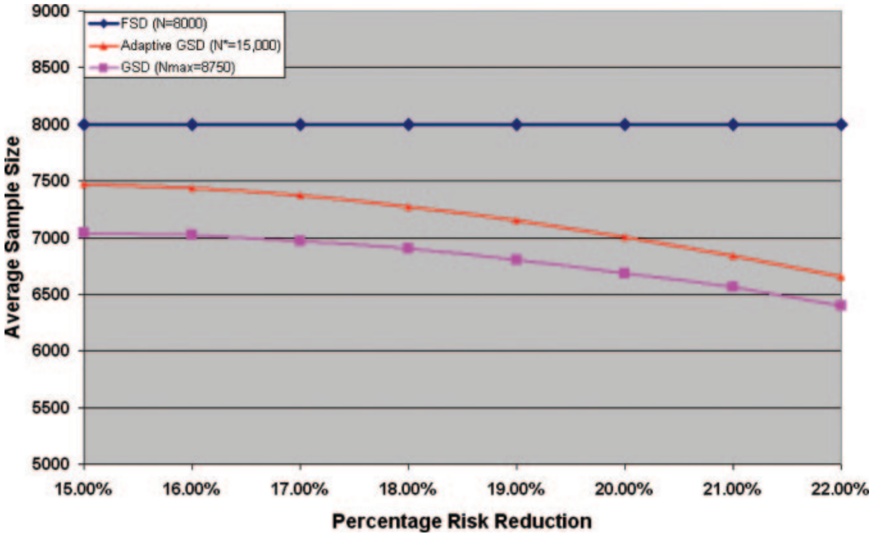


Figure 8. Average sample sizes for the fixed sample size, group sequential, and adaptive designs.

we will consider enrichment. The enrichment strategy proceeds as follows: We first estimate the number of additional patients needed to test the null hypothesis H_1 with 80% conditional power, assuming that the observed effect size in G_1 is the true effect size. That is, we consider testing for a treatment effect in the G_1 subpopulation. If that sample size plus the number of patients already enrolled is <15 000, the trial will continue until the additional number of patients is enrolled, but future eligibility will be restricted to patients belonging to subgroup G_1. If that enrichment strategy does not yield a sample size <15 000, the same calculation will be performed with the estimated effect size for subgroup G_2 and, if that reestimation yields a sample size <15 000, the trial will continue, enrolling only members of G_2. If neither of the sample size calculations for the patient subgroups yields a sample size <15 000, the trial will be continued with the original eligibility criteria and sample size target, provided that the conditional power with 8750 patients is at least 20%. Otherwise, the trial will be terminated for futility.

A question arises as to how to test for a treatment effect at the end of the trial given the possibility of a sample size increase and population enrichment at the second interim look. To preserve the type 1 error of the study as a whole, we employ a closed testing procedure that guarantees strong control of the type 1 error rate.[11] Specifically, if enrichment is implemented, the testing procedure involves 2 hypothesis tests. Suppose for specificity that the data lead to an enrichment strategy with the G_1 subpopulation. In that instance, test 1 is a test of the null hypothesis in the entire study population, consisting of patients enrolled from G_0 before the initiation of the enrichment strategy and patients enrolled from G_1 thereafter. This test is based on the original adaptive design, so the group sequential boundaries and the Müller and Schäfer procedure ensure that test 1 will be a level-α test. However, if test 1 rejects the null hypothesis, we can conclude only that the new treatment is superior to placebo in either the G_o population or the G_1 subpopulation. Thus, we then perform test 2, a conventional level-α test of the null hypothesis H_1 in patients from the G_1 subpopulation enrolled after the second interim analysis, that is, after enrichment began. If both hypotheses are rejected, we conclude that treatment is superior to placebo in the subpopulation G_1. A similar procedure is followed if enrichment is limited to subpopulation G_2. Details of this testing procedure and the demonstration that it provides strong control of type 1 error are provided in the online-only Data Supplement.

Simulation Experiments to Investigate Population Enrichment

Investigators conducting a large clinical trial comparing a novel antithrombotic agent with clopidogrel in patients undergoing percutaneous coronary intervention partitioned the overall patient population into 4 mutually exclusive subpopulations in which the event rates were expected to differ. The 4 subpopulations were as follows:

- High-risk patients naive to clopidogrel (≈30% of study population)
- High-risk patients pretreated with clopidogrel (≈30% of study population)
- Low-risk patients naive to clopidogrel (≈20% of study population)
- Low-risk patients pretreated with clopidogrel (≈20% of study population)

In this classification, high-risk patients are those with diabetes or an elevated troponin level. These subpopulations were selected before any unblinding of the data on the following basis:

1. They were clinically relevant and easily identifiable in a clinical setting.
2. Current understanding of the underlying pathophysiology supported the hypothesis that the risk reduction with the study drug relative to clopidogrel might differ in the 4 subpopulations.[12]
3. Parameters to identify the subpopulations were consistently collected.

For purposes of population enrichment, the aforementioned 4 subpopulations were combined into 3 nested groups G_0, G_1, and G_2 as follows:

G_0, the entire group
G_1, the subgroup of high-risk patients (60% of G_0)
G_2, the subgroup of high-risk patients naive to clopidogrel (50% of G_1)

To evaluate the adaptive enrichment strategy for selecting G_0, G_1, or G_2 at the second interim look, we generated events from the 4 mutually exclusive subpopulations, consolidated them according to the definitions of the 3 nested groups, and implemented the group sequential, sample size increase, and enrichment rules outlined in the earlier sections of this article. Because there are infinitely many ways to specify response rates for the treatment and control arms, we limited our investigation to 2 scenarios: a case of heterogeneous risk reduction across the 4 subpopulations and a case of homogeneous risk reduction across the 4 subpopulations. As before, we assume a 3-look group sequential trial with a planned maximum sample size of 8750 patients, interim looks after 50% and 70% of enrollment, and the possibility of increasing the sample size to 15 000 patients at the second interim analysis. The rules for enrichment and the subsequent hypothesis tests for strong control of type 1 error are as specified in the previous section. We provide a brief summary of the findings here. The results are discussed in greater detail in the online-only Data Supplement.

Scenario 1: Heterogeneous Risk Reduction

We considered a scenario in which the average risk reduction for the G_0 population was 17.4% but the 2 nested subgroups benefited to a greater extent than the average patient, with a risk reduction of 22.6% for G_1 and 27.1% for G_2. In this heterogeneous setting, the overall chance of claiming treatment efficacy for either G_0, G_1, or G_2 is boosted by enrichment from 69% to 81%. This 12% gain in power is achieved at the cost of an average sample size increase of 840 patients and a 3% drop in the chance of claiming efficacy for G_0 alone, relative to a design with no enrichment.

Scenario 2: Homogeneous Risk Reduction

We considered a scenario in which the G_0, G_1, and G_2 populations all achieved the same risk reduction of 17.4%. In this homogeneous setting, the enrichment design produces a 5% gain in power, from 69% to 73%. This 5% gain in power is achieved at the cost of an average sample size increase of 366 patients and a 1% drop in the chance of claiming efficacy for G_0 alone relative to a design with no enrichment.

In any application of the enrichment strategy, simulation results should be explored extensively because they enable the trial leadership to weigh the tradeoffs between an overall power gain, an increase in average sample size, and the possibility that, at the end of the trial, efficacy may only be claimed for a subset of the G_0 population.

Conclusions

As we have demonstrated, adaptive and enrichment designs for phase 3 trials are feasible and methodologically sound. Considerable work and practical experience, however, will be needed to characterize how best to use these designs in clinical research.

Although we have described the enrichment component of the study design in terms of prespecified subgroups, the 2-stage testing procedure has the advantage of allowing the group responsible for redesign to choose any patient subgroup on the basis of analysis of the accumulating data or external information. Because the null hypothesis of equal efficacy in the enriched subgroup is tested only in patients enrolled after the second interim analysis, the null distribution of that test does not depend on the procedure or deliberations used to define the subgroup. Narrowing patient eligibility to a subgroup is not without cost, of course, because more restrictive eligibility criteria make it more difficult to enroll patients and will usually require an extension of the enrollment period. Of course, any potential labeled indication for a drug or device might then be restricted to the enriched subgroup. Thus, enrichment designs must be studied carefully before they are applied in any specific trial.

Disclosures

Dr Bhatt discloses the following relationships: research grants (directly to the institution): Bristol-Myers Squibb, Eisai, Ethicon, Heartscape, Sanofi Aventis, The Medicines Company; consultant/advisory board (donated to nonprofits for >2 years): Arena, Astellas, AstraZeneca, Bayer, Bristol-Myers Squibb, Cardax, Centocor, Cogentus, Daiichi-Sankyo, Eisai, Eli Lilly, GlaxoSmithKline, Johnson & Johnson, McNeil, Medtronic, Millennium, Molecular Insights, Otsuka, Paringenix, PDL, Philips, Portola, Sanofi Aventis, Schering Plough, Scios, Takeda, The Medicines Company, Vertex. Dr Harrington discloses the following relationships: research grants to DCRI from The Medicines Company. A full listing of his disclosures is available at http://www.dcri.duke.edu/research/coi.jsp. Dr Ware discloses the following relationships: consultant/advisory board (compensation paid directly to the Harvard School of Public Health): Biogen Idec, The Medicines Company, Schering Plough. Dr Mehta is the President of Cytel Inc, a software and services company whose products and services are used by pharmaceutical and biotechnology companies. He does not derive any personal consulting income from these activities.

References

1. Rao SR, Schoenfeld DA. Survival methods. *Circulation*. 2007;115:109–113.
2. Mehta CR, Tsiatis AA. Flexible sample size considerations using information based interim monitoring. *DIA J*. 2001;35:1095–1112.
3. Pocock SJ. Group sequential methods in the design and analysis of clinical trials. *Biometrika*. 1977;64:191–199.
4. O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. *Biometrics*. 1979;35:549–556.
5. Lan KKG, DeMets DL. Discrete sequential boundaries for clinical trials. *Biometrika*. 1983;70:659–663.
6. Ware JH, Muller JE, Braunwald E. The futility index. *Am J Med*. 1985;78:635–643.
7. *East: Software for Advanced Clinical Trial Design, Simulation, and Monitoring*. Version 5.2. Cambridge, Mass: Cytel Statistical Software and Services; 2008.
8. *PEST 4: Planning & Evaluation of Sequential Trials*. Reading, UK: Medical and Pharmaceutical Statistics Research Unit, University of Reading, UK; 2000.
9. Müller HH, Schäfer H. Adaptive group sequential designs for clinical trials: combining the advantages of adaptive and of classical group sequential approaches. *Biometrics*. 2001;57:886–891.
10. Gao P, Ware JH, Mehta CR. Sample size re-estimation for adaptive sequential design in clinical trials. *J Biopharm Stat*. 2008;18:1184–1196.
11. Marcus R, Peritz E, Gabriel KR. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*. 1976;63:655–660.
12. Meadows TA, Bhatt DL. Clinical aspects of platelet inhibitors and thrombus formation. *Circ Res*. 2007;100:1261–1275.

Key Words: statistics ■ clinical trials ■ methods ■ design ■ angioplasty