STATISTICS IN MEDICINE
Statist. Med. 2000; **00**:1–6 *Prepared using* simauth.cls *[Version: 2002/09/18 v1.11]*

Adaptive Increase in Sample Size when Interim Results are Promising: A Practical Guide with Examples

Cyrus R. Mehta[1,2], Stuart J. Pocock[3]

[1] *Cytel Corporation,* [2] *Harvard School of Public Health,* [3] *London School of Hygiene and Tropical Medicine*

SUMMARY

This paper discusses the benefits and limitations of adaptive sample size re-estimation for phase 3 confirmatory clinical trials. Comparisons are made with more traditional fixed sample and group sequential designs. It is seen that the real benefit of the adaptive approach arises through the ability to invest sample size resources into the trial in stages. The trial starts with a small up-front sample size commitment. Additional sample size resources are committed to the trial only if promising results are obtained at an interim analysis. This strategy is shown through examples of actual trials, one in neurology and one in cardiology, to be more advantageous than the fixed sample or group sequential approaches in certain settings. A major factor that has generated controversy and inhibited more widespread use of these methods has been their reliance on non-standard tests and p-values for preserving the type-1 error. If, however, the sample size is only increased when interim results are promising, one can dispense with these non-standard methods of inference. Therefore, in the spirit of making adaptive increases in trial size more widely appealing and readily implementable we here define those promising circumstances in which a conventional final inference can be performed while preserving the overall type-1 error. Methodological, regulatory and operational issues are examined. Copyright © 2000 John Wiley & Sons, Ltd.

KEY WORDS: sample size re-estimation, two-stage designs, flexible clinical trials, conditional power, adaptive design, real examples.

1. Introduction

The sample size of a clinical trial is usually determined by statistical power calculations. Often, however, there is uncertainty and debate as to what magnitude of alternative hypothesis treatment difference and between-patient variability it is appropriate to pre-specify. Hence there exist methods to facilitate re-estimation of the sample size in the light of interim results from the ongoing trial. Statistical methods based only on interim estimates of between-patient variability, see Mehta and Tsiatis [1], are well understood and are not the focus of this article. We are concerned here with adaptive sample size re-estimation in confirmatory phase 3 trials based on unblinded interim estimates of the primary effect size. This controversial topic generates two areas of concern: 1) the need for a robust **statistical methodology** for sample size re-estimation and its consequences for making inferences from the final trial data, and 2) the **practical organization** of such an adaptive approach paying due regard to the confidentiality of interim data and the need to preserve the integrity of the trial's conduct throughout. Accordingly, any proposed adaptive re-estimation needs meticulous planning both in methodology and practicalities, which all needs documenting in a pre-defined Adaptive Charter. This is particularly important in pivotal phase 3 trials for regulatory approval.

*Correspondence to: Cyrus R. Mehta, Cytel Corporation, 675 Massachusetts Avenue, Cambridge, MA 02139. USA
E-mail: mehta@cytel.com

Received December 2009
Revised June 2010

Amongst the many methodological articles on adaptive sample size re-estimation, a focus has been on the need to preserve the overall type I error at the time of the final analysis. Bauer and Köhne [2] and Lehmecher and Wassmer [3] approached this problem by combining the p-values, obtained before and after the adaptation, with pre-specified weights. Cui, Hung and Wang [4] proposed an equivalent approach in which the Wald statistics are combined with pre-specified weights. Müller and Schäfer [5] showed that the overall type-1 error can be preserved unconditionally under any general adaptive change provided the conditional type-1 error that would have been obtained had there been no adaptation is preserved. This specializes to the method of Cui, Hung and Wang [4] if the only adaptive change is a sample size re-estimation. All the above adaptive methods of increasing the sample size require a non-standard final analysis in which the outcomes of patients enrolled before the unblinded interim analysis are treated differently from those of patients enrolled after the unblinded interim analysis. This could lead to practical difficulties in presenting trial findings in journals and regulatory submissions, because there is an attractive straightforward need to report conventional hypothesis tests and p-values, uncluttered by complex statistical adjustments. Chen, DeMets and Lan [6] took the first step in this direction by showing that if one increases the sample size only when interim results are promising, a conventional hypothesis test can be performed without inflating the type-1 error. Gao, Ware and Mehta [7] extended this idea to a broader range of promising zones in which such conventional tests may be performed. Posch, Bauer and Brannath [8] touched on the same issue indirectly by identifying so-called inconsistency regions in the sample space of the first stage statistic where it is possible for the final analysis based on the adaptive test to reject the null hypothesis while the conventional test will accept. Brannath, Koenig and Bauer [9] extended these regions to the more complicated case of comparing two treatments with a common control.

This paper makes the work of Gao, Ware and Mehta [7] more accessible to practioners by presenting it in the context of two-stage designs and tabulating explicit cut-off values for the promising region under different adaptive rules based on conditional power. The method is then applied to two actual case studies where it highlights the real benefit that these adaptive methods bring to the design of confirmatory phase 3 trials. The standard approach for powering such studies is to first estimate the underlying treatment effect for the primary endpoint from past studies in similar settings and use this estimate to compute the sample size needed for adequate power. There is usually considerable uncertainty associated with this estimate, however. This may be due to limited experience with the new compound, small sample sizes in the previous studies, changes in quality of care, lack of comparability between the old and new patient populations, use of a different primary endpoint, and numerous other factors. It is sometimes suggested that rather than attempt to estimate the treatment effect prior to launching the study, it might be more useful to specify the smallest treatment effect that is considered clinically meaningful and power the study to detect this effect. This approach too has its drawbacks. Even assuming that it is possible to identify the smallest clinically meaningful effect, the sample size required to detect that effect might be unacceptably large thereby rendering the trial infeasible. Designing for the smallest clinically meaningful effect has the additional drawback that if the true treatment effect is larger, the trial might be substantially overpowered. One option is to utilize the group sequential design whereby the trial will be terminated early with a smaller sample size if in truth the treatment effect is larger than the smallest clinically meaningful difference. An alternative approach that could be used either in place of, or in conjunction with the group sequential approach is the adaptive approach, wherein one starts out with a smaller sample size commitment which may be increased at an interim analysis if the results obtained are reasonably promising. In this method the risk of an underpowered study is reduced since there is an opportunity to review the initial sample size and adjust it based on data from the study itself. Just as important, since the additional sample size investment is only called for if the interim results are promising, it greatly increases the chances of eventual success and reduces the risk of an expensive failed study. The two case studies compare the group sequential and adaptive strategies and highlight the real benefits that the latter brings to the design in settings where it is simply infeasible commit a large sample size up-front. These benefits have not been highlighted in this manner before.

In Section 2 we review the statistical methodology pertinent to adaptive sample size re-estimation. In Section 3 we discuss the applicability of this methodology to actual trials, paying attention to practical concerns. In Section 4 we illustrate its use with two real examples, highlighting the importance of both analytic results and

simulation studies. Section 5 deals with regulatory and operational issues. Finally, in Section 6, we disscuss the wider context of using such an approach routinely in phase 3 trials.

2. Statistical Methods

The methods of this paper are applicable to two-arm, multi-stage group sequential trials of normal, binomial and survival endpoints, in which the sample size adaptation is performed at the penultimate stage. In order to simplify the exposition, however, we shall confine our discussion to two-stage trials. Subjects enter the trial one at a time and are randomized to either the control or experimental arms. Subject responses are assumed to be independent, identically distributed random variables. We will assume normality with equal variances to aid theoretical developments, though the principles relate more broadly. Specifically, let $X_{cj} \sim N(\mu_c, \sigma)$ and $X_{ej} \sim N(\mu_e, \sigma)$ denote the responses of subject j if randomized to the control or experimental arms, respectively. Define the treatment difference $\delta = \mu_e - \mu_c$. We are interested in a level-α hypothesis test of the null hypothesis

$$H_0\colon \delta = 0$$

versus the one sided alternative hypothesis that $\delta > 0$. To this end suppose we design a two-stage trial with balanced treatment assignment, an interim analysis after n_1 patients and a final analysis after n_2 patients. We assume initially that the interim analysis has just one purpose: to decide whether to increase the final sample size to be greater than the original n_2. The extension to early stopping for reasons of either efficacy or futility will be discussed in Section 3.3. The sample size n_2 required for a one-sided level-α fixed-sample test to attain $1-\beta$ power at $\delta = \delta_1$ is

$$n_2 = 4\sigma^2 \left[\frac{z_\alpha + z_\beta}{\delta_1}\right]^2 \tag{1}$$

where $z_u = \Phi^{-1}(1-u)$. To maximize the information for adaptive decision making one would like n_1 to be as near as possible to n_2 but, in practice, the follow-up needed to observe the primary outcome in each patient relative to the speed of trial recruitment, and also the time needed to prepare for any potential sample size increase, will both necessitate a somewhat earlier choice of n_1.

Throughout this paper we shall distinguish all incremental quantities from cumulative ones with a "tilde". Thus $\tilde{n}_2$ represents the incremental sample size between stage 1 and stage 2, whereas $n_2 = n_1 + \tilde{n}_2$ represents the cumulative sample size up to and including stage 2. Let $\hat{\delta}_1$, $\hat{\delta}_2$ and $\hat{\tilde{\delta}}_2$ be the maximum likelihood estimates of δ based on corresponding n_1, n_2 and $\tilde{n}_2$ subjects. Denote the cumulative Wald statistics at stages 1 and 2 by

$$Z_j = \frac{\hat{\delta}_j}{\mathrm{se}(\hat{\delta}_j)}\ , \quad j = 1, 2, \tag{2}$$

and the incremental Wald statistic at stage 2 by

$$\tilde{Z}_2 = \frac{\hat{\tilde{\delta}}_j}{\mathrm{se}(\hat{\tilde{\delta}}_j)}\ , \tag{3}$$

where, for balanced designs, the standard errors in the above expressions are given by $\mathrm{se}(\hat{\delta}_j) = 2\hat{\sigma}_j/\sqrt{n_j}$ and $\mathrm{se}(\hat{\tilde{\delta}}_j) = 2\hat{\tilde{\sigma}}_j/\sqrt{\tilde{n}_j}$. A level-$\alpha$ test will reject H_0 at the final analysis if $Z_2 \geq z_\alpha$ and the trial is not adaptive.

2.1. Sample Size Re-Estimation Using a Weighted Statistic

When an unblinded interim analyis is performed at the end of stage 1, as a consequence the cumulative and incremental sample sizes might be increased to n_2^* and $\tilde{n}_2^*$, respectively, resulting in corresponding Wald statistics Z_2^* and $\tilde{Z}_2^*$, as defined by equations (2) and (3) with n_2 replaced by n_2^* and $\tilde{n}_2$ replaced by $\tilde{n}_2^*$. A conventional final analysis would reject H_0 if $Z_2^* \geq z_\alpha$. However, Cui, Hung and Wang [4] have shown, by

simulation in the more general group sequential setting, that when the sample size is increased in a data dependent manner from n_2 to n_2^*, it is possible to have $P_0(Z_2^* \geq z_\alpha) > \alpha$, thus resulting in an inflation of type-1 error if the conventional final analysis is adopted. A similar result was shown analytically by Proschan and Hunsberger [10].

Cui, Hung and Wang [4] have shown that, regardless of the rule for altering the stage 2 sample size, if the conventional Wald statistic Z_2^* is replaced by the CHW statistic

$$Z_{2,\text{chw}}^* = \sqrt{\frac{n_1}{n_2}} Z_1 + \sqrt{\frac{\tilde{n}_2}{n_2}} \tilde{Z}_2^* \tag{4}$$

then the type-1 error is preserved. That is, $P_0(Z_{2,\text{chw}}^* \geq z_\alpha) = \alpha$.

Now although the CHW statistic (4) guarantees preservation of the type-1 error, it has the undesirable property of downweighting the contribution of the $\tilde{n}_2^* > \tilde{n}_2$ patients evaluated after the interim analysis. (Notice that $\tilde{Z}_2^*$ is weighted in proportion to $\sqrt{\tilde{n}_2}$ and not $\sqrt{\tilde{n}_2^*}$ in equation (4)). This contradicts the common sense premise that "all patients are equal".

2.2. Sample Size Re-Estimation Using the Conventional Statistic

Chen, DeMets and Lan [6] showed that if one increases the sample size only when the interim result is promising, the type-1 error is not inflated by use of the conventional Wald statistic Z_2^*. Specifically, let

$$\text{CP}_\delta(z_1, \tilde{n}_2) = P_\delta(Z_2 \geq z_\alpha | z_1) \tag{5}$$

denote the conditional power, or conditional probability of rejecting H_0 at the final analysis given $Z_1 = z_1$. Since δ is unknown, it is usual and convenient to replace it with $\hat{\delta}_1$ in equation (5). That is, conditional power is calculated assuming that the estimated treatment difference at interim analysis is the true effect. We shall see later that for our purposes this is preferable to evaluating conditional power at the planned value δ_1, as recommended by, for example, Bauer and Koenig [11].

It can then be shown that

$$\text{CP}_{\hat{\delta}_1}(z_1, \tilde{n}_2) = 1 - \Phi\left(\frac{z_\alpha \sqrt{n_2} - z_1 \sqrt{n_1}}{\sqrt{\tilde{n}_2}} - \frac{z_1 \sqrt{\tilde{n}_2}}{\sqrt{n_1}} \right) . \tag{6}$$

Chen, DeMets and Lan [6] have shown that, provided $\text{CP}_{\hat{\delta}_1}(z_1, \tilde{n}_2) \geq 0.5$, one may increase the stage-2 sample size from $\tilde{n}_2$ to $\tilde{n}_2^*$ and nevertheless use the conventional statistic Z_2^* for the final analysis, and the type-1 error will not be inflated. Gao, Ware and Mehta [7] gave further insight into this finding and extended it to situations where $\text{CP}_{\hat{\delta}}(z_1, \tilde{n}_2)$ was somewhat less than 50%. Exactly how far below 50% conditional power depends on specific circumstances which we will quantify in Section 3.2.

3. A Simple Approach to Adaptive Sample Size Increase

If adaptive sample size re-estimation methods are to be widely and wisely used in phase 3 trials, we need to encourage simple, transparent and robust approaches that appeal to experienced trialists and sponsors without requiring them to have an in-depth understanding of technical statistical methodology. It is with that goal in mind that we elucidate our adaptive approach.

3.1. Defining the Adaptive Algorithm: General Principles

The key idea is to evaluate conditional power at the interim look. If it is either too low or too high, we do not alter the sample size. If however, conditional power falls in a range that we deem promising, then the sample size may be increased, subject to a pre-determined upper limit, so as to boost the conditional power up to some

target level. In this article, we will be targetting the value $1-\beta$ for the boosted conditional power (typically 80% or 90%), though other values could just as well be selected. Note that this conditional power is for the observed interim estimate $\hat{\delta}_1$ and not the originally planned δ_1. Accordingly let $n_{\max}$ be a pre-specified maximum allowable sample size for the trial, typically determined by the sponsor's budgetary limits and feasibility of recruitment. In order to determine the new sample size $n_2^* \leq n_{\max}$ we partition the sample space of attainable $\mathrm{CP}_{\hat{\delta}_1}(z_1, \tilde{n}_2)$ values into three zones – Favorable, Promising and Unfavorable. We propose the following guidelines for sample size increase, depending on the zone into which $\mathrm{CP}_{\hat{\delta}_1}(z_1, \tilde{n}_2)$ falls at the interim look.

Unfavorable: $\mathrm{CP}_{\hat{\delta}_1}(z_1, \tilde{n}_2) < \mathrm{CP}_{\min}$ defines the unfavorable zone, where $\mathrm{CP}_{\min}$ is a fairly low probability, either pre-declared (e.g. 30% or 50%, say) or computed by an algorithm such as that described in Section 3.2. Implicit here is that the interim result is so disappointing that it is not worth increasing the sample size to retrieve conditional power. In this zone one continues to the original n_2.

Promising: $\mathrm{CP}_{\min} \leq \mathrm{CP}_{\hat{\delta}_1}(z_1, \tilde{n}_2) < 1-\beta$ defines the promising zone. In this zone the interim result is not disappointing, but it is not good enough for the conditional power to equal or exceed the unconditional power $1-\beta$ specified at the design stage. In that case the sample size is increased by just the right amount to recover the targeted power $1-\beta$, subject to not exceeding $n_{\max}$. Specifically, the new sample size is increased to

$$n_2^*(z_1) = \min(n_2'(z_1), n_{\max}) \ , \tag{7}$$

where $\tilde{n}_2'(z_1)$ satisfies the condition

$$\mathrm{CP}_{\hat{\delta}_1}(z_1, \tilde{n}_2') = 1-\beta \ . \tag{8}$$

A simplification of equation (5) of Gao, Ware and Mehta [7] shows that (8) is satisfied by the function

$$\tilde{n}_2'(z_1) = [\frac{n_1}{z_1^2}] \left[\frac{z_\alpha \sqrt{n_2} - z_1 \sqrt{n_1}}{\sqrt{n_2 - n_1}} + z_\beta \right]^2 \ . \tag{9}$$

Favorable $\mathrm{CP}_{\hat{\delta}_1}(z_1, \tilde{n}_2) \geq 1-\beta$ defines the favorable zone. In this zone the interim results are sufficiently favorable that the trial continues to the original n_2 without the need to adaptively increase the trial size. Note that this favorable zone includes all values of $\hat{\delta}_1 \geq \delta_1$, but also extends downwards to values of $\hat{\delta}_1$ slightly smaller than δ_1, the value of δ at which the trial was originally powered.

We require that the chosen value of $\mathrm{CP}_{\min}$ must not be so small as to result in an inflated overall type-1 error using a conventional final analysis. We shall show in Section 3.2 that the smallest $\mathrm{CP}_{\min}$ satisfying this requirement depends on the ratios $n_{\max}/n_2$ and n_1/n_2, and on the targeted power $1-\beta$. Table I displays smallest $\mathrm{CP}_{\min}$ values for some typical two-stage designs. It should be noted that the range $\mathrm{CP}_{\min} \leq \mathrm{CP}_{\hat{\delta}_1}(z_{1,\tilde{n}_2}) < 1-\beta$ represents the outer limits of the promising zone. The user is free to restrict sample size increases to a subset of this range if desired, though that will somewhat reduce somewhat the trial's overall power.

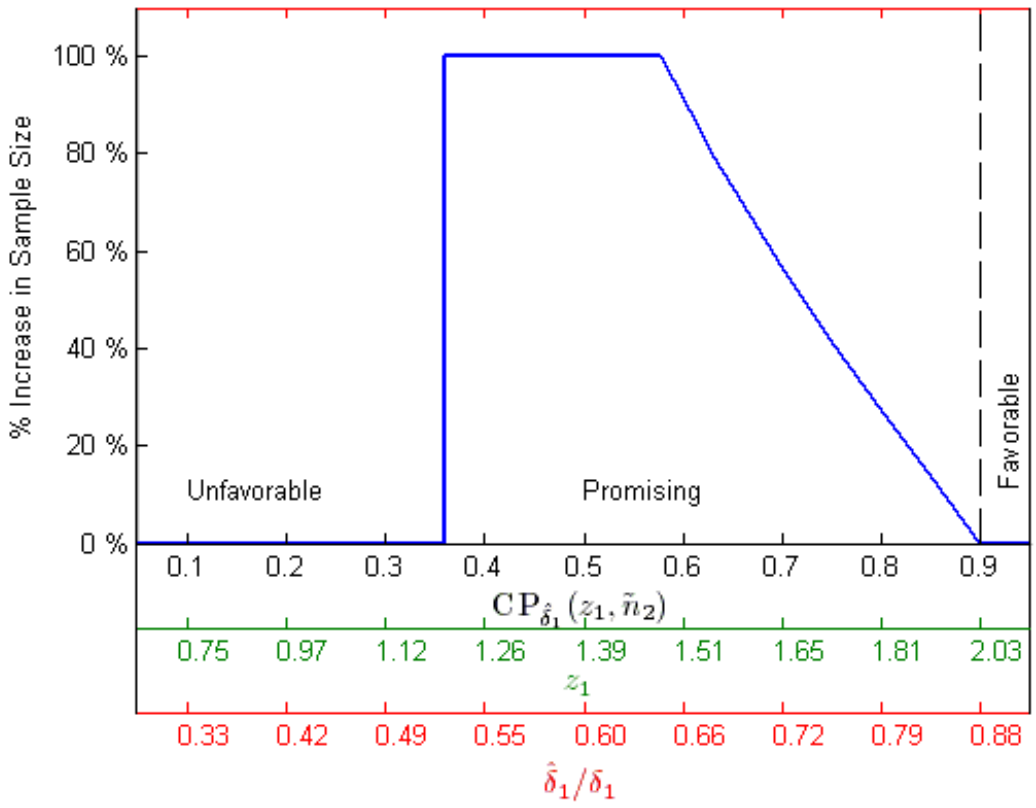
A schematic representation of the adaptive algorithm for sample size increase is displayed in Figure 1 for the situation where $n_{\max}/n_2 = 2$, $n_1/n_2 = 0.5$, and $1-\beta = 0.9$. Note that up to a doubling in trial size may well be a realistic maximum in many adaptive trials. For this design, Table I shows that $\mathrm{CP}_{\min} = 0.36$. Hence the promising zone is defined by $0.36 \leq \mathrm{CP}_{\hat{\delta}_1}(z_{1,\tilde{n}_2}) < 0.9$.

While the determination of the favorable, promising and unfavorable zones relates to conditional power, these zones can also be expressed in terms of z_1 via equation (6) and in terms of $\hat{\delta}_1/\delta_1$ through the relationship

$$\frac{\hat{\delta}_1}{\delta_1} = \left[\frac{z_1}{z_\alpha + z_\beta} \right] \sqrt{\frac{n_1}{n_2}} \ .$$

For instance, in Figure 1 the promising zone is also defined by $1.206 \leq z_1 < 2.027$, and by $0.526 \leq \hat{\delta}_1/\delta_1 < 0.884$.

Figure 1. Partitioning of Interim Result into Zones(†) and % Sample Size Increase in Each Zone: An Illustrative Example where $n_{\max}/n_2 = 2$, $n_1/n_2 = 0.5$, one-sided $\alpha = 0.025$, and $1 - \beta = 0.9$.



(†) The promising zone is defined by $0.36 \leq \mathrm{CP}_{\hat{\delta}_1}(z_1, \tilde{n}_2) < 0.9$ on the Conditional Power scale, by $1.206 \leq z_1 < 2.027$ on the z_1 scale, and by $0.526 \leq \hat{\delta}_1/\delta_1 < 0.884$ on the $\hat{\delta}_1/\delta_1$ scale.

3.2. Justifying the Use of a Conventional Final Analysis

We have stated in Section 2.2 that if the sample size is increased in a data dependent manner from n_2 to n_2^*, the CHW test ($Z^*_{2,\mathrm{chw}} \geq z_\alpha$) always protects the type-1 error whereas the conventional test $Z_2^* \geq z_\alpha$ does not carry this universal guarantee. If, however, the sample size is only increased when the interim conditional power falls within a certain "promising zone" then the conventional test will indeed protect the type-1 error. Let us now determine what that promising zone needs to be.

Lemma 1: Suppose that, upon observing $Z_1 = z_1$ at the interim look, the incremental sample size is altered from $\tilde{n}_2$ to $\tilde{n}_2^*$. Then, regardless of the formula used to compute $\tilde{n}_2^*$,

$$P_0(Z_2^* \geq b(z_1, \tilde{n}_2^*)) = \alpha \ , \tag{10}$$

where

$$b(z_1, \tilde{n}_2^*) = (n_2^*)^{-0.5} \left[\sqrt{\frac{\tilde{n}_2^*}{\tilde{n}_2}} (z_\alpha \sqrt{n_2} - z_1 \sqrt{n_1}) + z_1 \sqrt{n_1} \right] \ . \tag{11}$$

The proof is given in Gao, Ware and Mehta [7]. Note that if the sample size is not altered, then $\tilde{n}_2^* = \tilde{n}_2$, $Z_2^* = Z_2$, and $b(z_1, \tilde{n}_2^*) = z_\alpha$.

The lemma implies that if the sample size is altered at the interim analysis, one may nevertheless use the conventional test statistic Z_2^* for a level-α test at the final analysis provided the critical boundary z_α is replaced by $b(z_1, \tilde{n}_2^*)$. However, in keeping with our philosophy of having a simple, straightforward final analysis we would prefer to use the test $Z_2^* \geq z_\alpha$ rather than $Z_2^* \geq b(z_1, \tilde{n}_2^*)$ for rejecting H_0. To make this possible without inflating the type-1 error we define the promising zone as the set

$$\mathcal{P} = \{\mathrm{CP}_{\hat{\delta}_1}(z_1, \tilde{n}_2) \colon b(z_1, \tilde{n}_2^*(z_1)) \leq z_\alpha\} \ , \tag{12}$$

where the altered sample size $\tilde{n}_2^*(z_1)$ is made to depend explicitly on z_1 via equation (7). Note that the promising zone is identified ahead of time, before any data are unblinded. We then pre-specify that we will increase the incremental sample size from $\tilde{n}_2$ to $\tilde{n}_2^*(z_1)$ only if the observed value $Z_1 = z_1$ at the interim analysis is such that $\mathrm{CP}_{\hat{\delta}_1}(z_1, \tilde{n}_2) \in \mathcal{P}$. Otherwise the incremental sample size will remain unaltered at $\tilde{n}_2$. It then follows that

$$\alpha = P_0(Z_2^* \geq b(z_1, \tilde{n}_2^*(z_1)) \geq P_0(Z_2^* \geq z_\alpha) \ , \tag{13}$$

thus ensuring that the type-1 error will be preserved conservatively if we use the conventional test $Z_2^* \geq z_\alpha$ for the final analysis.

This adaptive strategy requires us to identify the promising zone $\mathcal{P}$ prior to unblinding the interim results, as follows. For any given value of $\mathrm{CP}_{\hat{\delta}_1}(z_1, \tilde{n}_2) \in (0, 1)$ one can find the required value of z_1 using equation (6). One then computes $n_2^*(z_1)$ using equations (7) to (9). Then substitute these values of z_1 and $n_2^*(z_1)$ into equation (11) to obtain $b(z_1, \tilde{n}_2^*)$. By repeating these calculations across the entire range of $\mathrm{CP}_{\hat{\delta}_1}(z_1, \tilde{n}_2) \in (0, 1)$, one can plot $\mathrm{CP}_{\hat{\delta}_1}(z_1, \tilde{n}_2)$ versus $b(z_1, \tilde{n}_2^*(z_1))$. Then $\mathcal{P}$ is characterized by the region of the curve that lies below z_α.

Figure 2 displays plots of $b(z_1, \tilde{n}_2^*(z_1))$ versus $\mathrm{CP}_{\hat{\delta}_1}(z_1, \tilde{n}_2)$ for ratios of $n_{\max}/n_2 = (1, 1.5, 2, 3, \infty)$. All plots are for the case $\alpha = 0.025$ (1-sided), $1 - \beta = 0.9$ and $n_1/n_2 = 0.5$.

Note that the curve in Figure 2 representing $n_{\max}/n_2 = 1$ is actually the line $z_\alpha = 1.96$. The curves for the other $n_{\max}/n_2$ ratios intersect the $z_\alpha = 1.96$ line in two places. The left intersection point, representing the start of the promising zone, is $\mathrm{CP}_{\min}$. It is different for each $n_{\max}/n_2$ curve, being about 0.31 for the $n_{\max}/n_2 = \infty$ curve, and increasing to 0.5 as $n_{\max}/n_2$ approaches 1. The right intersection point, representing the end of the promising zone, is equal to $1 - \beta$ for all the curves. The promising zone is the region of $\mathrm{CP}_{\hat{\delta}_1}(z_1, \tilde{n}_2)$ where the $n_{\max}/n_2$ curve dips below the $z_\alpha = 1.96$ line. Since the right intersection point is always $1 - \beta$, each promising zone is completely specified by its left intersection point, which is the $\mathrm{CP}_{\min}$ cut-off value.

Table I provides $\mathrm{CP}_{\min}$ cut-off values for some typical two-stage adaptive designs. For example, if $n_{\max}/n_2 = 2$, $n_1/n_2 = 0.5$ and the target conditional power value is $1 - \beta = 0.9$, then the promising zone is $0.36 \leq \mathrm{CP}_{\hat{\delta}_1}(z_1, \tilde{n}_2) < 0.9$. In this range it is permissible to increase the sample size so as to achieve the target conditional power of 0.9, subject to the cap $n_{\max} = 2n_2$, and nevertheless use conventional test $Z_2^* \geq 1.96$ for the final analysis. Although equation (13) suggests that the use of the conventional test in this setting might be conservative, the example in Section 4.1 shows that the conservatism is negligible.

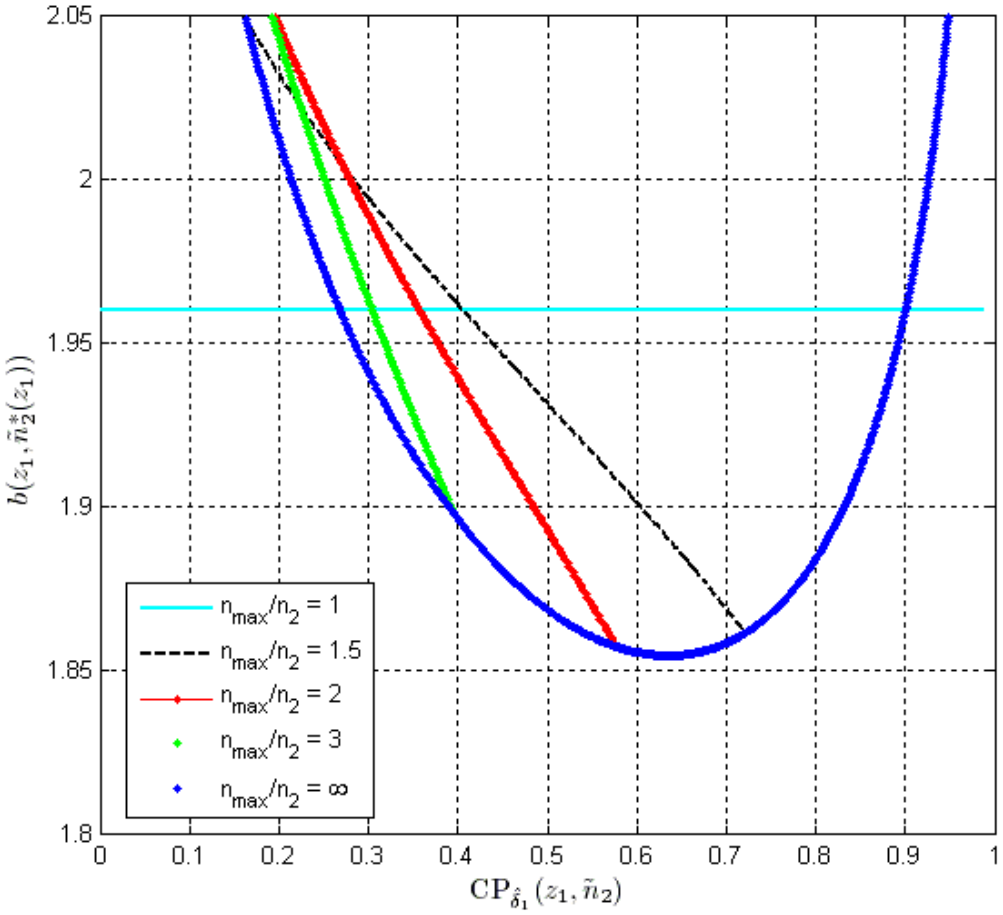
3.3. Early Stopping for Efficacy or Futility at Stage 1

The methods discussed above readily generalize to two-stage designs with possible early stopping at Stage 1. Using standard group-sequential methods (for example, Jennison and Turnbull [12]) we can construct a two-stage, one-sided, level-α, group sequential trial with futility boundaries (a_1, a_2), efficacy boundaries (b_1, b_2) and cumulative sample sizes (n_1, n_2) such that $a_2 = b_2$, and

$$P_0(Z_1 \geq b_1) + P_0(a_1 < Z_1 < b_1, Z_2 \geq b_2) = \alpha \ , \tag{14}$$

$$P_{\delta_1}(Z_1 \leq a_1) + P_{\delta_1}(a_1 < Z_1 < b_1, Z_2 \leq a_2) = \beta \ . \tag{15}$$

Figure 2. Plots of $b(z_1, \tilde{n}_2^*(z_1))$ versus $\mathrm{CP}_{\hat{\delta}_1}(z_1, \tilde{n}_2)$ for ratios of $n_{\max}/n_2 = (1, 1.5, 2, 3, \infty)$ given $\alpha = 0.025$ (1-sided), $1 - \beta = 0.9$ and $n_1/n_2 = 0.5$



Then the trial would stop early for efficacy if $Z_1 \geq b_1$ and for futility if $Z_1 \leq a_1$. By forcing $a_2 = b_2$ in the construction of the boundaries, one is assured of an unambiguous decision at the end of the trial; reject H_0 if $Z_2 \geq b_2$, otherwise accept H_0. Equation (14) ensures that the type-1 error of this group sequential trial equals α, while equation (15) ensures that its power remains $1 - \beta$.

All the results of Section 3 will carry over provided we replace z_α by b_2 in every formula that contains z_α.

4. Two Case Studies

We illustrate the methods described above by applying them to the design of two phase 3 clinical trials, one for a continuous primary efficacy endpoint and the other for a binomial primary efficacy endpoint. These two case

Table I. CP_{min} Cut-Off Values for Some Typical Two-Stage Adaptive Designs with no Early Stopping either for Efficacy or Futility

Sample Size Ratios		CP_{min} Values for Targeted Conditional Powers	
Maximum Allowed (n_{max}/n_2)	At Interim Look (n_1/n_2)	80%	90%
1.5	0.25	0.42	0.42
1.5	0.5	0.41	0.41
1.5	0.75	0.38	0.38
2	0.25	0.37	0.37
2	0.5	0.36	0.36
2	0.75	0.33	0.33
3	0.25	0.32	0.32
3	0.5	0.31	0.31
3	0.75	0.30	0.27
∞	0.25	0.32	0.28
∞	0.5	0.31	0.27
∞	0.75	0.30	0.25

studies refer to actual trials. For reasons of confidentiality some details, such as the identities of the trial sponsors and the names of the test drugs are omitted.

4.1. A Schizophrenia Trial

Consider a randomized phase 3 trial of a test drug versus an active comparator for patients with negative symptoms schizophrenia. The primary efficacy endpoint is the improvement from baseline to week 26 in the Negative Symptoms Assessment (NSA) a standardized score, to be analyzed as a quantitive variable. The planned trial size, $n_2 = 442$, was based on having 80% power to detect a mean score difference $\delta_1 = 2$ with a one sided level-0.025 test, assuming a score standard deviation $\sigma = 7.5$ in each treatment group. The true value of δ could be less than 2; a value as low as 1.6 would still be considered clinically meaningful, but would require a trial size of 690 subjects. The operating characteristics of Plan 1 (fixed sample size of 442) and Plan 2 (fixed sample size of 690) are displayed side by side in Table II for values of δ between 1.6 and 2.0. If resources were

Table II. Operating Characteristics of Plan 1 and Plan 2

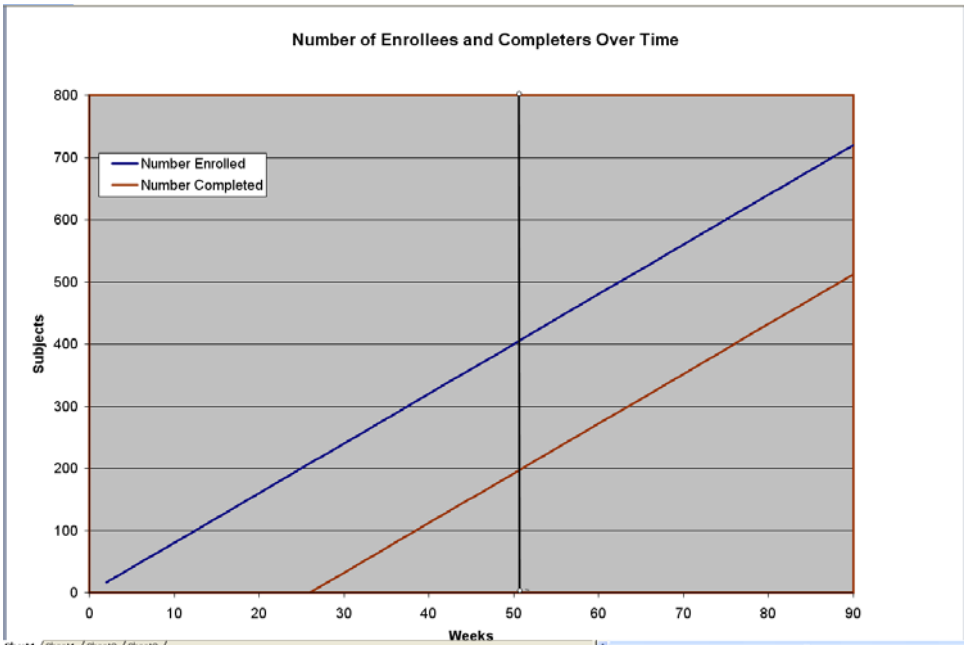
	Plan 1		Plan 2	
δ	Sample Size	Power	Sample Size	Power
1.6	442	61%	690	80%
1.7	442	66%	690	85%
1.8	442	71%	690	88%
1.9	442	76%	690	91%
2.0	442	80%	690	93%

plentiful, Plan 2 would clearly be the preferred option. The sponsor must, however, allocate scarce resources over a number of studies and in any case is not in favor of designing an overpowered trial. This leads naturally to considering a design that might be more flexible with respect to sample size than either of the above two single-look fixed sample designs. We will consider two types of flexible designs; group sequential and adaptive.

Group Sequential Design

When sample size flexibility is desired for late-stage trials, it is often appropriate to first explore the group sequential option. Let Plan 3 be a group sequential design with one interim look and 80% power to detect $\delta = 1.6$ such that if in fact $\delta = 2$, the trial will stop early. While this would appear to be an attractive option, it is important to consider not just the saving in study duration but also the saving in the actual number of subjects randomized to the study. Since the efficacy endpoint for this trial will only be observed at week 26, the actual saving in sample size will be affected by the enrollment rate. In the current study it is anticipated that subjects will enroll at an average rate of 8 per week. The number of subjects enrolled and the number of completers over time are displayed graphically in Figure 3 Observe that there is a 26-week horizontal

Figure 3. Impact of Enrollment Rate and Length of Follow-Up on Trial Completion



separation between the two parallel lines depicting, respectively, the graph for enrollment and the graph for study completion. This 26-week gap must be taken into consideration when evaluating the savings achieved by utilizing a group sequential design.

The two major design parameters to be specified for Plan 3 are the timing of the interim analysis and the amount of type-1 error to be spent. We will assume that data must be available for at least 200 completers before the trial can be terminated for efficacy so that an adequate safety profile may be developed for the study drugs. Therefore a suitable time point for the interim analysis is week 52, when we will have enrolled 416 subjects with complete 26-week data on 208 subjects and partial data on the remaining 208 subjects. These 208 subjects with partial data are referred to as the "overruns". It is generally held that the type-1 error should be spent conservatively in the early stages of a trial so as to ensure that results based on premature termination will be compelling and have the capacity to change medical practice (see Pocock, 2005). Suppose then that we use the $\gamma(-4)$ error spending function proposed by Hwang, Shih and DeCani citehwang-1990 to obtain the early stopping boundary. The boundary thus produced resembles the conservative O'Brien-Fleming [13] boundary. Its nominal critical point for early stopping is 3.067 standard deviations. The one sided p-value corresponding to this early stopping boundary is $1 - \Phi(3.067) = 0.0011$ which, if met, would indeed be compelling enough to justify premature termination. In order to maintain 80% power to detect $\delta = 1.6$ with this spending function we require the maximum sample size of Plan 3 to be 694, a slight inflation from the corresponding fixed sample Plan 2. Plan 3, however has a lower expected sample size because of the possibility

of early stopping. The power and expected sample size values of the group sequential Plan 3 (with and without adjustment for the overruns) are displayed in Table III for different choices of δ alongside corresponding power and sample size values for the fixed sample Plan 2. It is seen from Table III that Plan 3 offers a modest benefit

Table III. Operating Characteristics of Plan3 (Group Sequential) and Plan2 (Fixed Sample)

	Plan3 (Group Sequential)				Plan2 (Fixed Sample)	
	Probability of	Expected Sample Size				
δ	Early Stopping	No Overruns	With Overruns	Power	SampSiz	Power
1.6	6.6%	663	676	80%	690	80%
1.7	7.9%	657	672	85%	690	85%
1.8	9.3%	650	668	88%	690	88%
1.9	11.0%	641	663	91%	690	91%
2.0	13.0%	632	658	94%	690	94%

relative to Plan 2. After accounting for the overruns, the expected sample sizes under Plan 3 range between 658 and 676 for corresponding values of δ between 2 and 1.6, as compared to a fixed sample size of 690 under Plan2. This amounts to a sample size saving of between 2% and 5% for the group sequential method. In terms of power, Plan 2 and Plan 3 are practically identical. Thus for the current trial a group sequential design with conservative error spending offers no substantial advantage over a conventional single look design with a fixed sample size. One is still faced with the dilemma of committing excessive sample size resources up front in order to ensure adequate power at $\delta = 1.6$, with limited prospects of saving on sample size in the event that $\delta = 2$.

Adaptive Design The above difficulties lead us to consider whether Plan 1, which was intended to detect $\delta = 2$ with 80% power and hence does not have such a large up-front sample size commitment, might be improved so as to provide some insurance against substantial power loss in the event that $\delta = 1.6$. The adaptive approach is suited to this purpose. Accordingly it is planned to undertake a 442 person trial with an interim analysis after observing $n_1 = 208$ completers, and to possibly increase the sample size if the interim results are promising. There is no intention to stop early at $n_1 = 208$, and the sponsor has set a maximal doubling of sample size to $n_{\max} = 884$.

The adaptive plan focuses on the estimate of conditional power $\mathrm{CP}_{\hat{\delta}_1}(z_1, \tilde{n}_2)$ at the interim analysis $n_1 = 208$ with the following choices, in which the lower cut-off value for the promising zone, $\mathrm{CP}_{\min} = 0.365$, was obtained by a conservative extrapolation from Table I.

Favorable: If $\mathrm{CP}_{\hat{\delta}_1}(z_1, \tilde{n}_2) \geq 0.8$ (or equivalently, $\hat{\delta}_1 \geq 1.835$), continue to $n_2 = 442$.

Promising: If $0.365 \leq \mathrm{CP}_{\hat{\delta}_1}(z_1, \tilde{n}_2) < 0.8$ (or equivalently $1.219 \leq \hat{\delta}_1 < 1.835$), increase sample size to $n_2^* = \min(n_2', 884)$, where n_2' is such that $\mathrm{CP}_{\hat{\delta}_1}(z_1, \tilde{n}_2') = 0.8$.

Unfavorable: If $\mathrm{CP}_{\hat{\delta}_1}(z_1, \tilde{n}_2) < 0.365$ (or equivalently $\hat{\delta}_1 < 1.219$), continue to $n_2 = 442$.

Table IV displays power and expected sample sizes for selected values of δ between 1.6 and 2.0, based on 100,000 simulations of the adaptive design. For comparative purposes, corresponding power and sample size values for the conventional fixed sample design are also displayed. The power of the adaptive design has increased by 4% at $\delta = 1.6$ and by 3% at $\delta = 2$ compared to the fixed sample design. These power gains were obtained at the cost of corresponding average sample size increases of between 10% and 12%. The gains in power appear to be fairly modest, largely because, as shown in Table V, the chance that a sample size increase will actually happen is only about 25%. However, the adaptive design offers a significant benefit in terms of risk reduction, not reflected in Table IV. Table V reveals that on those promising occasions when an increase in sample size does happen, the gain in conditional power compared to a fixed sample design is substantial. For

Table IV. Operating Characteristics of Fixed Sample and Adaptive Designs

Value of δ	Fixed Sample Design		Plan 4 (Adaptive)	
	Power	Expected SampleSize	Power	Expected Sample Size
1.6	61%	442	65%	499
1.7	66%	442	71%	498
1.8	71%	442	75%	497
1.9	76%	442	79%	494
2.0	80%	442	83%	491
All Plan 4 results are based on 100,000 simulated trials				

example, for true $\delta = 1.6$ the conditional power rises from 62% to 82%. This is a major attraction of the adaptive design. The additional sample size resources are requested only if the interim results are promising, in which case the study team can justify the extra investment since the chances of success are high. If the interim results are either unfavorable or favorable, there is no extra investment and the sponsor is no worse off than would be the case under the non-adaptive Plan 1.

Table V. Operating Characteristics of the Fixed Sample and Adaptive Designs, Conditional on Interim Outcome

δ	Interim Outcome	Probability of Interim Outcome	Power Conditional on Interim Outcome		Expected Sample Size	
			Fixed	Adaptive	Fixed	Adaptive
1.6	Unfavorable	36%	30%	30%	442	442
	Promising	23%	62%	82%	442	687
	Favorable	41%	87%	87%	442	442
1.7	Unfavorable	32%	34%	34%	442	442
	Promising	23%	67%	85%	442	685
	Favorable	45%	89%	89%	442	442
1.8	Unfavorable	29%	38%	38%	442	442
	Promising	23%	70%	88%	442	682
	Favorable	49%	91%	91%	442	442
1.9	Unfavorable	26%	43%	43%	442	442
	Promising	22%	74%	90%	442	679
	Favorable	52%	93%	93%	442	442
2.0	Unfavorable	23%	47%	47%	442	442
	Promising	21%	77%	92%	442	678
	Favorable	56%	95%	95%	442	442
All results are based on 100,000 simulated trials						

We have proven in Section 3.2 that the use of the conventional Wald statistic rather than the weighted CHW statistic preserves they type-1 error conservatively. In fact the extent of the conservatism in this example is very small; it is seen via simulation that the overall type-1 error is actually 0.024 and leads to negligible power loss.

4.2. A Placebo Controlled Acute Coronary Syndromes Trial

This example concerns a large phase 3 randomized trial of an active drug versus placebo in patients with acute coronary syndromes undergoing percutaneous coronary angioplasty (see Mehta et. al. [14] for details). The

primary endpoint is a composite of death, myocardial infarction and ischemia-driven revascularisation within 48 hours of randomisation. We assume on the basis of prior knowledge that the event rate for the placebo arm is 8.7%. The investigational drug is expected to reduce the event rate by at least 20%. The investigators are planning to randomize a total of 8000 subjects in equal proportions to the two arms of the study. It is easy to show that a conventional fixed sample design enrolling a total of 8000 subjects will have 83% power to detect a 20% risk reduction with a one-sided level-0.025 test of significance. The actual risk reduction is expected to be larger, but could also be as low as 15%, a treatment effect that would still be of clinical interest given the severity and importance of the outcomes. In addition, there is some uncertainty about the magnitude of the placebo event rate. For these reasons the investigators wish to build into the trial design some flexibility for adjusting the sample size. Two options under consideration are, a group sequential design with the possibility of early stopping in case the risk reduction is large, and an adaptive design with the possibility of increasing the sample size in case the risk reduction is small. In the remainder of this section we shall discuss these two options and show how they may be combined into a single design that captures the benefits of both.

Group Sequential Design

We first transform the fixed sample design into an 8000 person group sequential design with two interim looks, one after 4000 subjects are enrolled (50% of total information) and the second after 5600 subjects are enrolled (70% of total information). Early stopping efficacy boundaries are derived from the Lan and DeMets [15] O'Brien-Fleming type error spending function. Let us denote this group sequential design as GSD1. The operating characteristics of GSD1 are displayed in Table VI. The first column of Table VI is a list of potential risk reductions, defined as $100 \times (1-\rho)\%$ where $\rho = \pi_t/\pi_c$, π_t is the event rate for the treatment arm, and π_c is the event rate for the control arm. The remaining columns display early stopping probabilities, power and expected sample size. Since the endpoint is observed within 48 hours, the problem of overruns that we encountered in the schizophrenia trial is negligible and may be ignored.

Table VI. Operating Characteristics of GSD1, a Three-Look 8000-Person Group Sequential Design

Risk Reduction $100 \times (1-\rho)$	Probability of Crossing Efficacy Boundary			Overall Power	Expected Sample Size
	At Look 1 ($N = 4000$)	At Look 2 ($N = 5600$)	At Final Look ($N = 8000$)		
15%	0.074	0.183	0.309	57%	7264
17%	0.109	0.235	0.335	68%	7002
20%	0.181	0.310	0.330	82%	6535
23%	0.279	0.362	0.275	92%	6017
25%	0.357	0.376	0.222	96%	5671

Table VI shows that GSD1 is well powered, with large savings of expected sample size for risk reductions of 20% or more. It is thus a satisfactory design if, as is initially believed, the magnitude of the risk reduction is in the range 20% to 25%. This design does not, however, offer as good protection against a false negative conclusion for smaller risk reductions. In particular, even though 15% is still a clinically meaningful risk reduction, GSD1offers only 57% power to detect this treatment effect.

One possibility then is to increase the up-front sample size commitment of the group sequential design so that it has 80% power if the risk reduction is 15%. This leads to GSD2, a three-look group sequential design with a maximum sample size commitment of 13,853 subjects, one interim look after 6926 subjects (50% of total information) and a second interim look after 9697 subjects (70% of total information). GSD2 has 80% power to detect a risk reduction of 15% with a one-sided level-0.025 test. Table VII displays operating characteristics of GSD2 for risk reductions between 15%, and 25%. Notice that by attempting to provide adequate power at 15% risk reduction, the low end of clinically meaningful treatment effects, we have significantly over-powered the trial for values of risk reduction in the expected range, 20% to 25% . This need not be a cause for concern, however, because the trial stops early with high probability for risk reductions in the 20% to 25% range, resulting in much smaller expected sample sizes. Thus if the sponsor is in a position to commit nearly 14,000

Table VII. Operating Characteristics of GSD2, a Three-Look 13,853-Person Grp Sequential Design

Risk Reduction $100 \times (1-\rho)$	Probability of Crossing Efficacy Boundary			Overall Power	Expected Sample Size
	At Look 1 ($N = 6926$)	At Look 2 ($N = 9697$)	At Final Look ($N = 13,853$)		
15%	0.167	0. 298	0.335	80%	11,456
17%	0.246	0.349	0.296	89%	10,699
20%	0.395	0.375	0.196	97%	9558
23%	0.565	0.329	0.099	99.3%	8574
25%	0.675	0.269	0.054	99.8%	8061

subjects at the very begining of the trial, GSD2 would be a very satisfactory design. In this case, however, the sponsor was unable to find investors who would risk such a large sample size commitment up-front. But it was possible to to obtain funding for GSD1 with a commitment to invest further, conditional on promising interim results. This is a situation faced by many small pharmaceutical and biotechnology companies who are unable to design a trial for the smallest clinically meaningful treatment effect up front, but can find the resources provided the investment is made conditionally, with a smaller up-front investment followed by an additional investment conditional on the interim results being promising.

Adaptive Group Sequential Design

We convert the three-look group sequential design GSD1 into an adaptive group sequential design by inserting into it the option to increase the sample size at look 2, when 5600 subjects have been enrolled. Denote the modified design by A-GSD1. The rules governing the sample size increase for A-GSD1 are similar to the rules specified in Section 4.1 for the schizophrenia trial, but tailored to the needs of the current trial. The idea is to identify unfavorable, promising and favorable zones for the interim results at look 2, based on the attained conditional power. The sample size should only be increased if the interim results fall in the promising zone. Subject to an upper limit, the sample size should be increased by just the right amount to boost the current conditional power to some desired level (say 80%). The following are the design specifications for A-GSD1:

1. The starting design is GSD1 with a sample size of 8000 subjects, one interim look after enrolling 4000 subjects and a second interim look after enrolling 5600 subjects. The efficacy stopping boundaries at these two interim looks are derived from the Lan and DeMets [15] error spending function of the O'Brien-Fleming type
2. At the second interim analysis, with data available on 5600 subjects, the conditional power is computed using the estimated value $\hat{\rho}$ as though it were the true relative risk ρ. If the conditional power is no greater than 33% the outcome is deemed to be unfavorable. If the conditional power is between 33% and 80%, the outcome is deemed to be promising. If the conditional power is at least 80%, the outcome is deemed to be favorable
3. If the interim outcome is promising, the sample size is re-computed so as to achieve 80% conditional power at the estimated value $\hat{\rho}$. The original sample size is then updated to the re-computed sample size, subject to the constraint in item 4 shown below
4. If the re-computed sample size is less than 8000, the original sample size of 8000 subjects is used. If the re-computed sample size exceeds 16,000, the sample size is curtailed at 16,000 subjects

Some features of this adaptive strategy are worth pointing out. First, we selected a cut-off of 33% for the start of the promising zone because it satisfies the condition for $\mathrm{CP}_{\min}$ discussed in Section 3.2. Thus we may use the conventional Wald statistic Z_2^* for testing the null hypothesis at the final analysis even if the sample size is increased. The simulation results for A-GSD1 are all based on using Z_2^* for the final analysis. Second, the sample size is re-computed on the basis of data from 5600 subjects from the trial itself. Therefore the estimate of ρ available at the interim analysis is substantially more reliable than the estimate that was used at the start

of the trial to compute an initial sample size of 8000 subjects. The latter estimate is typically derived from smaller pilot studies or from other phase 3 studies in which the patient population might not be exactly the same as that of the current trial. Finally, a sample size increase is only requested if the interim results are promising, in which case the trial sponsor should be willing to invest the additional resources needed to power the trial adequately. In contrast GSD2 increases the sample size substantially at the very begining of the trial, before any data are available to determine if the large sample size is justified.

Operating Characteristics of Adaptive Group Sequential Design
Table VIII displays the power and expected sample size of the adaptive group sequential design A-GSD1. For comparative purposes corresponding power and sample size values of GSD1 are also provided. If there is a 15%

Table VIII. Operating Characteristics of GSD1 (Group Sequential) and A-GSD1 (Adaptive Group Sequential) Designs

Risk Reduction	GSD1 (Group Sequential)		A-GSD1 (Adaptive Group Sequential)	
$100 \times (1-\rho)$	Power	Expected Sample Size	Power	Expected Sample Size
15%	57%	7264	62%	8288
17%	68%	7002	72%	7957
20%	82%	6535	86%	7313
23%	92%	6017	93%	6580
25%	96%	5671	97%	6052
All results for A-GSD1 are based on 100,000 simulated trials				

risk reduction, A-GSD1 has 6% more power than GSD1 but utilizes an additional 1093 subjects on average. It is seen that as the risk reduction parameter increases the power advantage and additional sample size requirement of A-GSD1 are reduced relative to GSD1.

The power and sample size entries in Table VIII were computed unconditionally, and for that reason do not reveal the real benefit that design A-GSD1 offers compared to design GSD1. As discussed previously in the schizophrenia example, the real benefit of an adaptive design is the opportunity it provides to invest in the trial in stages with the second stage investment forthcoming only if promising results are obtained at the first stage. Accordingly Table IX displays the operating characteristics of both GSD1 and A-GSD1 conditional on the zone into which the conditional power falls at the second interim analysis. The table reveals substantial gains in power for A-GSD1 compared to GSD1 at all values of risk reduction if the second interim outcome falls in the promising zone, thereby leading to an increase in the sample size. Outside this zone the two designs have the same operating characterisics since the sample size does not change. If the second interim outcome falls in the unfavorable zone, the trial appears to be headed for failure and an additional sample size investment would be risky. If the second interim outcome falls in the favorable zone, the trial is headed for success without the need to increase the sample size. Thus the adaptive design provides the opportunity to increase the sample size only when the results of the second interim analysis fall in the promising zone. This is precisely when the trial can most benefit from a sample size increase.

Adding a Futility Boundary One concern with design A-GSD1 is that it lacks a futility boundary. There is thus the risk of proceeding to the end, possibly with a sample size increase, when the magnitude of the risk reduction is small and unlikely to result in a successful trial. In particular, suppose that the null hypothesis is true. In that case we can show by simulation that the power (i.e., the type-1 error) is 2.3% and the expected sample size under A-GSD1 is 8234 subjects. It might thus be desirable to include some type of futility stopping rule for the trial. In this trial the investigators proposed the following futility stopping rules at the two interim analysis time points:

1. Stop for futility at the first interim analysis ($N = 4000$) if the estimated event rate for the experimental arm is at least 1% higher than the estimated event rate for the control arm

Table IX. Operating Characteristics of GSD1 (Group Sequential) and A-GSD1 (Adaptive Group Sequential) Designs Conditional on Second Interim Outcome

Risk Reduction $100 \times (1-\rho)$	Second Interim Outcome	Probability of Interim Outcome	Power Conditional on Second Interim Outcome		Expected Sample Size	
			GSD1	A-GSD1	GSD1	A-GSD1
15%	Unfavorable	37%	16%	16%	8000	8000
	Promising	23%	57%	81%	8000	12445
	Favorable	40%	94%	94%	6152	6161
17%	Unfavorable	29%	21%	21%	8000	8000
	Promising	22 %	64%	87%	8000	12326
	Favorable	49 %	96%	96%	5992	6000
20%	Unfavorable	17%	31%	31%	8000	8000
	Promising	19%	73%	93%	8000	12164
	Favorable	64%	98%	98 %	5721	5727
23%	Unfavorable	9%	42%	42%	8000	8000
	Promising	14%	81%	96%	8000	12008
	Favorable	77%	99%	99%	5440	5449
25%	Unfavorable	5%	49%	49%	8000	8000
	Promising	10%	85%	98%	8000	11875
	Favorable	85%	99.5%	99.5%	5250	5247
All results are based on 100,000 simulated trials						

2. Stop for futility at the second interim analysis ($N = 5600$) if the conditional power, based on the estimated risk ratio $\hat{\rho}$, is no greater than 20%

The impact of the futility boundary on the unconditional operating characteristics of the A-GSD1 design are displayed in Table X. The inclusion of the futility boundary has resulted in a dramatic saving of nearly 3000 subjects, on average, at the null hypothesis of no risk reduction. Furthermore, notwithstanding a small power loss of 2-4%, the trial continues to have well over 80% power for risk reductions of 20% or more. The trial suffers a power loss of 4% if the magnitude of the risk reduction is 15%, the low end of the range of clinical interest. In this situation, however, the unconditional power is inadequate (only 62%) even without a futility boundary. To fully appreciate the impact of the futility boundary on power and expected sample size, it is

Table X. Operating Characteristics of the A-GSD1 Design with and without a Futility Boundary

Risk Reduction $100 \times (1-\rho)$	A-GSD1 with No Futility Boundary		A-GSD1 with Futility Boundary	
	Power	Expected Sample Size	Power	Expected Sample Size
0%	2.3%	8242	2.1%	5346
15%	62%	8288	58%	7482
20%	86%	7313	82%	6983
25%	97%	6052	95%	5959
All results are based on 100,000 simulated trials				

necessary to study the operating characteristics of the trial conditional on the results of the second interim analysis. These results are displayed in Table XI. It is seen that the presence of the futility boundary does not cause any loss of power for trials that enter the promising or favorable zones at the second interim analysis.

Table XI. Operating Characteristics of A-GSD1 Design with and without a Futility Boundary, Conditional on the Second Interim Outcome

Risk Reduction $100 \times (1-\rho)$	Second Interim Outcome	Probability of Interim Outcome	Power Conditional on Second Interim Outcome		Expected Sample Size	
			No Fut	With Fut	No Fut	With Fut
0%	Unfavorable	93%	0.5%	0.2%	8000	4874
	Promising	6%	13%	14%	13225	13115
	Favorable	2%	61%	65%	6976	6913
15%	Unfavorable	37%	16%	6%	8000	5805
	Promising	23 %	81%	81%	12445	12286
	Favorable	40 %	94%	95%	6161	6103
20%	Unfavorable	17%	31%	13%	8000	6037
	Promising	19%	93%	93%	12164	11976
	Favorable	64%	98%	98%	5727	5698
25%	Unfavorable	5%	49%	21%	8000	6257
	Promising	10%	98%	97%	11875	11850
	Favorable	85%	99.5%	99.5%	5247	5224
All results are based on 100,000 simulated trials						

Additionally the presence of the futility boundary causes the average sample size to be reduced substantially in the unfavorable zone while remaining the same in the other two zones. In effect the futility boundary terminates a proportion of trials that enter the unfavorable zone thereby preventing them from proceeding to conclusion. It has no impact on trials that enter the promising or favorable zones.

5. Regulatory and Operational Issues

The statistical methodology that permits sample size re-estimation based on an interim estimate of the treatment effect has been available for about ten years and has gradually made its way into actual confirmatory clinical trials. Both examples discussed in the current paper are based on real trials whose designs were accepted by the FDA, and subsequently activated. The recently released FDA Guidance for Industry on Adaptive Design [16] is an indispensible document for sponsors who are considering the adaptive route for their confirmatory trials. It classifies adaptive approaches into those that are generally well-understood and those that are less well understood. Group sequential designs and blinded sample size re-estimation fall into the category of well-understood methods while sample size re-estimation using unblinded estimates of the treatment effect falls into the category of less well understood methods. This is not to suggest that the latter methods are disallowed. Rather, the guidance document says that these methods should be used when other better understood methods are unable to meet the primary study objectives. In our context this implies that one should be able to provide a valid reason for including the option for unblinded sample size re-estimation at an interim look in preference to using a conventional design that fixes the maximum sample size at the outset. The usual reason is that even after a thorough review of all relevant prior data there is still some uncertainty regarding the true treatment effect and population variability. Sometimes a group sequential design powered to detect a small but clinically meaningful treatment effect will resolve this difficulty. We have seen, however, that over-runs, large up-front commitments, and conservative boundaries for early efficacy stopping are deterrents to the group sequential approach.

The reason for the FDA's classification of of unblinded sample size re-estimation into the "less well understood" category has nothing to do with the statistics. The validity of the statistical methodology is accepted. The FDA's real concern is the possibility of operational bias. Operational bias might be introduced

into the study if the interim results were somehow revealed to the investigators and led to selective withdrawal of patients from the study before they had completed their full course of treatment. The only way to prevent this type of operational bias is by creating good operating procedures that are built into an interim analysis charter and are strictly followed. Such procedures are already in place for group sequential designs, and can be modified for the specifics of the adaptive setting. To this end it is customary to set up an independent statistical center (ISC) for creating the interim analysis report and an independent interim analysis review committee (IARC) whose task it is to review the interim analysis report and implement the sample size re-assessment in accordance with the interim analysis charter. If the study already has a functioning data monitoring committee (DMC), that committee or a subset of that committee could fulfill the role of the IARC. The IARC charter would however differ from that of a traditional DMC. This charter should describe how the interim data are to be transferred to the ISC, what the interim analysis report should contain, who may have access to that report, the precise rules for altering the sample size, and the procedure to be followed in making the sample size recommendation to the sponsor. Access to the charter must be restricted to the ISC, the IARC, and only those employees of the sponsor organization who were involved in the trial design. Everyone with access to the charter should be required to sign a confidentiality agreement in order that the precise rules governing sample size re-estimation may not be disclosed to the outside community. It should be recognized that, as in good data monitoring practice, any statistical adaptive algorithm should be seen as guideline to aid decision-making rather than as a definitive rule. Other criteria, such as secondary endpoints, safety issues, feasible recruitment rates and budgetery limits, can impact on both independent recommendations by IARC and decisions by sponsors. It is sensible to mention such judgemental issues in the Charter, so that necessary flexibility is not perceived as "breaking the rules".

Even with good operating procedures in place to prevent premature disclosure of interim results, the mere fact that a sample size increase was implemented cannot be hidden from the sites enrolling subjects. Concerns have been raised that this knowledge alone could modify investigator behavior. It is difficult, however, to anticipate how investigators would interpret this knowledge. Some might feel that the chances of the trial succeeding have improved while others might take it as a sign that the initial estimate of treatment effect was too optimistic. A discussion about the rationale for the adaptive design, its potential for reducing the risk of running an underpowered study, and the need to maintain the same pattern of patient enrollment throughout the study might be an important agenda item at an investigator meeting. It would likewise be important to educate the Institutional Review Boards about the adaptive nature of the design, stressing that while the actual sample size is not known at the outset, the maximum sample size if an adaptation takes place has been fixed. In this sense the uncertainty about the sample size of an adaptive design is similar to the uncertainty about the sample size of a group sequential design or of any event driven trial in which the number of events but not the sample size is known in advance.

Additional discussion concerning these operational issues is provided in a White Paper published by the PhRMA Adaptive Working Group [17]. The FDA has adopted a "wait-and-see" attitude in order to gain more experience with the risks and benefits of unblinded sample size re-estimation. However, a well prepared submission that addresses both the statistical and operational issues and backs them up with simulations and a detailed charter stands an excellent chance of regulatory acceptance.

6. Concluding Remarks

While there has been increasing enthusiasm for adaptive designs amongst methodologists, those with a specialist understanding, and some trial sponsors, delivered applications in major phase 3 trials have been relatively few to date. This is partly because investigators and sponsors who do not have any in-depth knowledge of adaptive designs, and do not fully understand what value they add to conventional designs, are wary of getting involved in their statistical and organizational complexities. We have demonstrated through case studies that unblinded sample size re-estimation is a valuable design option, to be used alongside other more established fixed sample and group sequential design options. In addition we have identified those promising circumstances under which one may increase the sample size and still perform a conventional final

analysis without resorting to adjustments or weightings that would be a mystery to most trialists.

In the early years following publication of methods for unblinded sample size re-estimation, concerns were raised about their efficiency relative to conventional group sequential methods. These concerns arose because in these publications preservation of the type-1 error was achieved by use of a weighted statistic like (4) with pre-specified weights. This violates the sufficiency principle. Tsiatis and Mehta [18] demonstrated that for any adaptive design with sample size modification requiring the use of the weighted statistic (4) one could construct a group sequential design utilizing the usual sufficient statistic that would stop earlier with higher probability of rejecting the null hypothesis if $\delta > 0$ and also stop earlier with higher probability of accepting the null hypothesis if $\delta < 0$. Jennison and Turnbull [19] demonstrated a similar result empirically by creating a group sequential design with the same power function as an adaptive design and then demonstrating by simulation that it would have a smaller expected sample size. These results are of great theoretical interest but of limited practical value for sponsors of industry trials. They produce appreciable efficiency gains only if there are no over-runs, a large number of interim analyses, a large up-front sample size commitment and aggressive early-stopping boundaries. Sponsors are usually unwilling or unable to impose these conditions on their trial designs. Furthermore, the analyses of Tsiatis and Mehta [18] and Jennison and Turnbull [19] do not capture the essential appeal of the adaptive approach which is to invest limited resources initially and to invest more only after seeing interim results that are promising. Most adaptive designs have only two stages, increase the sample by at most a factor of two, and that too, only if the results fall in a promising zone. It has yet to be demonstrated that there is any appreciable loss of efficiency due to the use of the weighted statistic in these settings.

It has been argued [11] that using conditional power assuming $\hat{\delta}_1$ is the truth, for purposes of sample size re-estimation, has its limitations given the lack of precision of the estimate. The point, however, is not to estimate δ with precision but rather to identify if the interim results are promising. The conditional power estimate is one way of accomplishing this. One can, if preferred, represent the promising zone in terms of z_1, $\hat{\delta}_1$ or $\mathrm{CP}_\delta(z_1, \tilde{n}_2)$ since these statistics are all transformations of $\mathrm{CP}_{\hat{\delta}_1}(z_1, \tilde{n}_2)$. For example, the X-axis of Figure 1 reveals that a conditional power of 36% (the start of the promising zone for the schizophrenia trial) corresponds to an estimate $\hat{\delta}_1$ that is about 50% of the δ_1 for which the trial was powered. Ultimately the effectiveness of the promising zone must be judged in terms of its impact on the operating characteristics of the design. The case studies have shown that this approach works well in practice.

An important issue that has not been discussed at all is how to provide a valid point estimate and confidence interval for the combined data in which an adaptive sample size change was made. This is a more difficult problem and research is still on-going. We refer the reader to articles by Mehta, Bauer, Posch and Brannath [20], and Brannath, Mehta and Posch [21] for recent results.

In conclusion, while we would like to see adaptive sample size re-estimation play a larger role in major phase 3 trials, each case requires careful attention to both organizational and statistical issues. An adaptive design might not always be the right choice. The more established fixed sample and group sequential designs should always be evaluated alongside an adaptive design. Simulations play a crucial role in understanding the operating characteristics of an adaptive design and deciding whether it is an appropriate choice for the trial under consideration. There should be a tangible, quantifiable benefit arising from the decision to take the adaptive route.

REFERENCES

1. Mehta CR, Tsiatis AA. Flexible sample size considerations using information-based interim monitoring. *Drug Information Journal* —2001—; **35**:1095–1112.
2. Bauer P, Kohne K. Evaluation of experiments with adaptive interim analyses. *Biometrics* —1994—; **50**:1029–41.
3. Lehmacher W, Wassmer G. Adaptive sample size calculations in group sequential trials. *Biometrics* —1999—; **55**:1286–90.
4. Cui L, Hung HM, Wang SJ. Modification of sample size in group sequential clinical trials. *Biometrics* —1999—; **55**:853–7.
5. Muller HH, Schafer H. Adaptive group sequential designs for clinical trials: combining the advantages of adaptive and of classical group sequential approaches. *Biometrics* —2001—; **57**:886–91.

Prepared using simauth.cls

6. Chen YH, DeMets DL, Lan KK. Increasing the sample size when the unblinded interim result is promising. *Stat Med* —2004—; **23**:1023–38.
7. Gao P, Ware JH, Mehta C. Sample size re-estimation for adaptive sequential design in clinical trials. *J Biopharm Stat* —2008—; **18**:1184–96.
8. Posch M, Bauer P, Brannath W. Issues in designing flexible trials. *Stat Med* —2003—; **22**:953–69.
9. Brannath W, Koenig F, Bauer P. Multiplicity and flexibility in clinical trials. *Pharm Stat* —2007—; **6**:205–16.
10. Proschan MA, Hunsberger SA. Designed extension of studies based on conditional power. *Biometrics* —1995—; **51**:1315–24.
11. Bauer P, Koenig F. The reassessment of trial perspectives from interim data–a critical view. *Stat Med* —2006—; **25**:23–36.
12. Jennison C, Turnbull BW. *Group Sequential Methods with Applications to Clinical Trials*. New York: Chapman & Hall/CRC, —2000—.
13. O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. *Biometrics* —1979—; **35**:549–556.
14. Mehta C, Gao P, Bhatt DL, Harrington RA, Skerjanec S, Ware JH. Optimizing trial design: sequential, adaptive, and enrichment strategies. *Circulation* —2009—; **119**:597–605.
15. Lan KKG, Demets DL. Discrete sequential boundaries for clinical trials. *Biometrika* —1983—; **70**:659–663.
16. FDA. Guidance for industry: Adaptive design clinical trials for drugs and biologics, —2010—.
17. PhRMA. White paper of the phrma adaptive working group. *DIA Journal* —2007—; .
18. Tsiatis A, Mehta C. On the inefficiency of the adaptive design for monitoring clinical trials. *Biometrika* —2003—; **90**:367–378.
19. Jennison C, Turnbull BW. Mid-course sample size modification in clinical trials based on the observed treatment effect. *Stat Med* —2003—; **22**:971–93.
20. Mehta CR, Bauer P, Posch M, Brannath W. Repeated confidence intervals for adaptive group sequential trials. *Stat Med* —2007—; **26**:5422–33.
21. Brannath W, Mehta CR, Posch M. Exact confidence bounds following adaptive group sequential tests. *Biometrics* —2009—; **65**:539–46.