

## Adaptive, group sequential and decision theoretic approaches to sample size determination

Cyrus R. Mehta<sup>1,2,\*</sup>,<sup>†</sup> and Nitin R. Patel<sup>1,3</sup>

<sup>1</sup>*Cytel Inc., 675, Massachusetts Avenue, Cambridge, MA, U.S.A.*

<sup>2</sup>*Harvard School of Public Health, U.S.A.*

<sup>3</sup>*Massachusetts Institute of Technology, U.S.A.*

### SUMMARY

This paper presents two adaptive methods for sample size re-estimation within a unified group sequential framework. The conceptual and practical distinction between these adaptive modifications and more traditional sample size changes due to revised estimates of nuisance parameters is highlighted. The motivation for the adaptive designs is discussed. Having established that adaptive sample size modifications can be made without inflating the type 1 error, the paper concludes with a novel decision theoretic approach for determining the magnitude of the sample size modification. Copyright © 2006 John Wiley & Sons, Ltd.

**KEY WORDS:** sample size re-estimation; two-stage designs; flexible clinical trials; Bayesian decision theory; utility functions

### 1. INTRODUCTION

Sample size is a key design input for any randomized clinical trial. Unfortunately, it is often computed in the face of inadequate knowledge about  $\sigma^2$  the inter-subject variance, and  $\delta$  the effect size. Economic pressures, possibly combined with competition for patients, then encourage trial investigators to make optimistic estimates of these two design parameters, a tendency that frequently results in underpowered studies. An underpowered trial is extremely undesirable, for it places human subjects at risk with a low probability of reaching a positive scientific conclusion and can result in abandoning an effective compound. Therefore, in recent years there has been a considerable amount of research on more flexible clinical trials where the sample size is re-estimated after the clinical trial is underway, on the basis of updated information about  $\sigma^2$

\*Correspondence to: Cyrus R. Mehta, Cytel Software Corporation, 675, Massachusetts Avenue, Cambridge, MA, U.S.A.

<sup>†</sup>E-mail: mehta@cytel.com

and  $\delta$ . The updated information may arise either from external sources, from interim results of the on-going trial, or from a combination of the two. The various methods of sample size re-estimation based on updated information about  $\sigma^2$  are well established, and accepted by the statistical community (see, for example, References [1, 2]). They will not be discussed here. In this paper, we are concerned primarily with so-called ‘adaptive trials’ in which the sample size is re-estimated from updated information about  $\delta$  after the study is activated. The appropriateness of such sample size re-estimation has generated some debate. Critics of this type of design revision argue that the same end—ensuring adequate power at the appropriate value of  $\delta$ —can be achieved more efficiently through a group sequential design [3, 4]. This is a valid argument in settings where one is prepared to pre-specify a minimum clinically meaningful value of  $\delta$ , commit a large maximum sample size to the trial up-front, and forgo the option to make data-driven design changes as the trial progresses. There may be situations, however, where the flexibility to learn from the interim data and adapt the future course of the trial offsets the improved efficiency of the group sequential approach. We shall provide examples of such situations. More generally, we shall question the appropriateness of using power as the primary criterion for sample size determination and shall turn to statistical decision theory for a radically different formulation of the problem.

This paper is divided into six sections. In Section 2 we motivate the problem through an anti-depression study that could be implemented by either the group sequential or adaptive approaches. In the group sequential setting control of the type 1 error is preserved through the well-established error spending function approach [5]. Methods for controlling the type 1 error after an adaptive sample size re-estimation have a more recent history. Two such methods are discussed in Section 3. Section 4 presents examples of situations in which a sponsor might be willing to trade-off some loss of efficiency in exchange for the flexibility to make mid-course corrections to an on-going study. Section 5 then attempts to address the sample size question through a Bayesian decision theoretic formulation. Some final remarks are presented in Section 6.

## 2. MOTIVATING EXAMPLE

We consider a recently completed two-arm randomized clinical trial of a new treatment *versus* an active control. (The names of the compounds being tested are not disclosed for confidentiality reasons.) The primary endpoint is the week-6 decline in the Hamilton Depression Index (HAM-D) relative to the baseline value at the time of randomization. Let  $\mu_c$  and  $\mu_e$  be the mean week-6 declines for the control and experimental groups, respectively. Define  $\delta = \mu_e - \mu_c$  to be the effect size, or mean HAM-D improvement with the new treatment relative to the active control. The patient-to-patient variance is assumed, from past experience with this population, to be  $\sigma^2 = 100$ . Sample size adjustments due to revised estimates of  $\sigma^2$  from accruing data are not the focus of the present paper, though we will summarize the ‘information-based’ approach to this problem in Section 6. Our main concern is with  $\delta$ , an unknown parameter that crucially affects the power of the study for any given sample size.

Technically, the Neyman–Pearson theory of statistical inference does not require us to provide an estimate of  $\delta$  at the design stage. Rather it requires that we specify  $\delta_1$ , a clinically meaningful treatment effect at which it would be desirable to provide high power for the trial. The total sample size  $N$  needed to achieve a power of  $100 \times (1 - \beta)$  per cent with a two-sided level- $\alpha$  test is then computed by the well-known formula

$$N = 4\sigma^2 \left[ \frac{z_{\alpha/2} + z_{\beta}}{\delta_1} \right]^2 \quad (1)$$

where  $z_u = \Phi^{-1}(1 - u)$  and  $\Phi(\cdot)$  is the standard normal distribution function. When investigators try to put this formula into practice, however, they sometimes face difficulties. There are many reasons for this. In some trials it is difficult to specify the value of  $\delta_1$  *a priori* because of limited experience either with the endpoint, the compound or the patient population. Sometimes a very small value of  $\delta_1$  might be clinically meaningful but would require too large an up-front sample size commitment to justify launching the study without an opportunity for subsequent sample size revision. Often, the choice of  $\delta_1$  is based on an initial assessment of the value of  $\delta$ . During the course of a long-term trial the standard of care in the control group might improve leading to a dilution in the benefit provided by the experimental regimen, and hence to a smaller value of  $\delta$ . This has happened with cardiac trials, HIV trials and melanoma trials where each subject receives background therapy that is optimized to the specifics of the patient's medical and biological profiles. For all these reasons and several others it would appear desirable to have some idea about the true value of  $\delta$  before choosing the final sample size for a study. A reasonable approach would be to run a two-stage trial where the data at stage 1, possibly combined with external evidence, may be used to estimate  $\delta$  and thereby determine the sample size at stage 2. This is the essence of the adaptive design, and it can also be extended to more than two stages.

Now let us consider the design of the depression trial. The value of  $\delta$  is unknown and prior data on this parameter are rather limited. However, the investigators expect that  $\delta$  lies between 2 and 4 units on the HAM-D scale. Moreover, any value in this range is considered to be both clinically meaningful as well as economically viable. A conservative option would therefore be to power the trial using  $\delta_1 = 2$ . If we substitute this value into equation (1), with  $\alpha = 0.05$  and  $\beta = 0.1$ , we obtain  $N = 1051$ . This is regarded as too large an up-front commitment. In fact, the sponsor will not permit such a large trial to be launched without additional data about the unknown parameter  $\delta$ . A more risky strategy would be to power the trial using  $\delta_1 = 4$ . Then only  $N = 263$  subjects are required to achieve 90 per cent power. However, the power would deteriorate rapidly if the true value of  $\delta$  was less than 4. For example, if the true value was  $\delta = 3$ , the trial would only have 68 per cent power unless additional patients were recruited. What is needed clearly is a flexible way to adjust the sample size as the trial proceeds and interim data become available. Two flexible approaches are available for this problem—the group sequential approach and the adaptive approach. These will be discussed in the following sections.

For future reference we will refer to the design utilizing  $N = 1051$  as Plan 1 and the design utilizing  $N = 263$  design as Plan 2. Both plans are implemented in the same manner;  $N$  subjects are enrolled and followed until week-6 HAM-D scores are available on all of them. Then the standardized test statistic

$$Z = \frac{\hat{\delta}}{\text{se}(\hat{\delta})} \quad (2)$$

is computed, where  $\hat{\delta}$  is the maximum likelihood estimate of  $\delta$  and  $\text{se}(\hat{\delta})$  is its standard error. If  $Z \geq 1.96$  statistical significance at the 0.05 level is achieved. The statistic (2) is called the Wald statistic.

### 2.1. The group sequential design

The group sequential methodology is well established both in the scientific literature and in clinical trials practice (for example, Reference [4]). It permits interim monitoring of the accruing data with possible early stopping if the observed value of the test statistic crosses a stopping boundary.

Figure 1 displays Plan 3, a group sequential design (created by the East [6] software package) with three equally spaced looks at the accruing data.

Plan 3, like its fixed-sample counterpart Plan 1, will achieve 90 per cent power with a two-sided level-0.05 test if  $\delta = 2$ . In this design, one would monitor the data sequentially up to three times, after 354, 709 and 1063 patients, respectively, have completed 6 weeks of therapy and are therefore evaluable for HAM-D response. At the  $i$ th monitoring time point,  $i = 1, 2, 3$ , one would compute the Wald statistic

$$Z_i = \frac{\hat{\delta}_i}{\text{se}(\hat{\delta}_i)}$$

and declare statistical significance if the  $i$ th stopping boundary,  $b_i$ , was crossed. Here  $\hat{\delta}_i$  is the maximum likelihood estimate of  $\delta$  computed from *all* data available up to and including the  $i$ th monitoring time point,  $\text{se}(\hat{\delta}_i)$  is its standard error and the three stopping boundaries are  $b_1 = 3.71$ ,  $b_2 = 2.51$  and  $b_3 = 1.99$ . These three boundary values were obtained by applying the  $\alpha$ -spending function methodology of Lan and DeMets [5] with the default spending function

$$\alpha(t) = 4 - 4\Phi\left(\frac{z_{\alpha/4}}{\sqrt{t}}\right) \tag{3}$$

The final stopping boundary for Plan 3, 1.99, is slightly larger than 1.96, the criterion for declaring statistical significance under Plan 1. In other words, if Plan 3 proceeds all the way to look 3, statistical significance can be declared only if the final  $p$ -value does not exceed  $2 \times (1 - \Phi(1.99)) = 0.046$  whereas the  $p$ -value criterion for declaring statistical significance under Plan 1 is  $2 \times (1 - \Phi(1.96)) = 0.05$ . The slightly stricter significance criterion that is imposed on

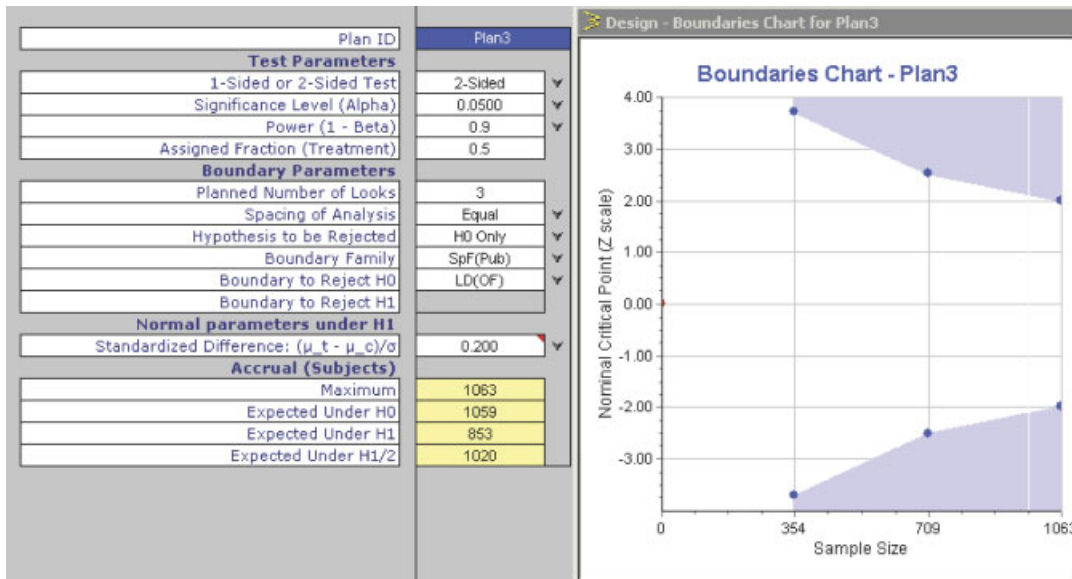


Figure 1. Three look group sequential Lan–DeMets design.

Plan 3 ensures that its type 1 error will be preserved despite the possibility of performing up to three hypothesis tests. The maximum sample size if Plan 3 proceeds all the way to look 3 without crossing an earlier boundary is 1063, a slight inflation over the Plan 1 sample size of 1051. This sample size, denoted by  $N_{\max}$ , is obtained by applying an inflation factor to equation (1):

$$N_{\max} = 4\sigma^2 \left[ \frac{z_{\alpha/2} + z_{\beta}}{\delta_1} \right]^2 \times \text{IF} \quad (4)$$

where the magnitude of the inflation factor IF depends on  $\alpha$ ,  $\beta$ , the number of interim looks and the spending function. For example, in a 3-look group sequential trial with  $\alpha = 0.05$ ,  $\beta = 0.1$ , and spending function given by equation (3),  $\text{IF} = 1.02$ .

Although the type 1 error ( $\alpha = 0.05$ ) and power (90 per cent at  $\delta = 2$ ) are identical for both Plan 1 and Plan 3, the latter has the advantage of permitting early stopping. For example, if  $\delta = 3$  the probability of crossing the boundary by the second look, after seeing data on 709 patients is 93 per cent, and the probability of crossing at the very first look, after seeing data on 354 patients is 19 per cent. If  $\delta = 4$ , the probability of crossing the boundary by the second look is 99.8 per cent. In this case, notwithstanding the large up-front commitment of 1063 patients, it is almost certain that the trial will stop early. Thus, the slight inflation in maximum sample size, from 1051 under Plan 1 to 1063 under Plan 3, seems to be a small price to pay in exchange for the possibility of a substantial saving in sample size if  $\delta > 2$ .

## 2.2. The adaptive design

The group sequential design has traditionally been utilized only for early termination decisions. Thus, Plan 3 was designed with a large up-front commitment of patients, but with the possibility of early stopping if fewer patients were needed. An adaptive design typically proceeds in the opposite direction. It starts out with a small initial commitment of patients but factors in the possibility that the sample size might be increased during the course of the trial. This can be achieved naturally within the group sequential framework by taking advantage of the interim monitoring to estimate  $\delta$  and re-assess whether the initial specification of sample size remains adequate. Plan 4, displayed below in Figure 2, is designed to achieve this objective.

In Plan 4 the trial is powered at  $\delta_1 = 4$ , the optimistic end of the range for  $\delta$ . It is thus the counterpart of the fixed-sample Plan 2. The maximum sample size for Plan 4 is 267 patients. The slight inflation relative to the fixed-sample size of 263 for Plan 2, reflects the fact that an interim look is taken under Plan 4. The interim look is taken 75 per cent of the way through the trial, after data are available on 200 patients. If the Wald statistic

$$Z_1 = \frac{\hat{\delta}_1}{\text{se}(\hat{\delta}_1)} \quad (5)$$

based on these 200 patients equals or exceeds the stage-1 stopping boundary  $b_1 = 2.34$ , the trial may be stopped and statistical significance declared. Otherwise, the trial proceeds until complete data are available for the remaining 67 patients. The Wald statistic

$$Z_2 = \frac{\hat{\delta}_2}{\text{se}(\hat{\delta}_2)} \quad (6)$$

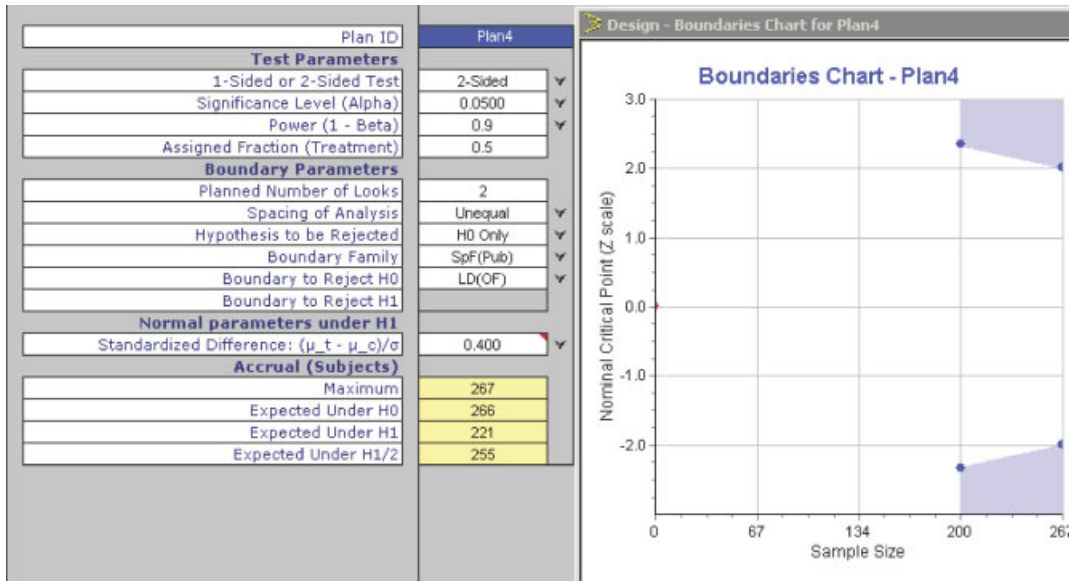


Figure 2. Two-stage adaptive design.

is then computed from data on *all* 267 patients and statistical significance is declared if this cumulative Wald statistic equals or exceeds the final stopping boundary  $b_2 = 2.01$ . This two-stage group sequential stopping boundary was computed using the Lan and DeMets [5]  $\alpha$ -spending function specified by equation (3).

Plan 4 becomes adaptive if we permit the total sample size to be altered after observing the interim results on the first 200 patients (assuming of course that the stage-1 stopping boundary is not crossed). Although both increases and decreases in sample size are permissible in an adaptive design, in this paper, we will only consider the situation where the sample size is increased. In an industry trial there is usually greater interest in increasing the sample size to avoid an underpowered study rather than decreasing the sample size to avoid an overpowered study. However, the methodology discussed here is applicable to both situations.

The magnitude of the sample size increase could be based solely on the interim data, solely on external information independent of the current study, or on a combination of these two factors. If the magnitude of the sample size increase at stage 2 is a function (either explicit or implicit) of the estimate of  $\delta$  obtained from the stage 1 data, there is a possibility that the overall type 1 error of the two-stage procedure will be inflated. How much can the type 1 error be inflated? Proschan and Hunsberger [7] have calculated that by suitable choice of the stage-2 sample size it is possible to inflate the overall type 1 error of a one-sided level- $\alpha$  two-stage procedure to as much as  $\alpha + \exp(-z_\alpha^2/2)/4$ . Thus, if the one-sided  $\alpha$  is 0.025 it is possible to increase the type 1 error to a maximum of 0.062. The inflation of type 1 error can be reduced by introducing a futility stopping rule at stage 1 (as in Reference [8]). For example, if we agree to terminate the trial for futility at stage 1 if  $\hat{\delta}_1 < 0$ , then the maximum type 1 error is 0.0495. Finally, some investigators (for example, Reference [9]) recommend that we increase the sample size at stage 2 by the smallest possible amount that will result in a conditional power (probability of achieving statistical significance at

the end of stage 2, conditional on the stage-1 results) of at least  $100 \times (1 - \beta)$  per cent. If we apply this rule, in conjunction with the requirement that the sample size should not be increased unless  $\hat{\delta}_1 > 0$ , the type 1 error is inflated by a minimal amount. Simulation results reveal that if this strategy is applied at stage 2 of plan Plan 4, the probability of crossing the upper stopping boundary under the null hypothesis increases from 0.025 to 0.032. For regulatory purposes, of course, it is essential that, regardless of the strategy implemented for increasing the sample size at stage 2, there be *no* inflation of type 1 error. We show how this may be accomplished in the next section.

### 3. PREVENTING INFLATION OF TYPE 1 ERROR IN AN ADAPTIVE DESIGN

In this section, we discuss two ways to adjust hypothesis tests in adaptive group sequential trials so as to preserve the type 1 error. For simplicity, we will only discuss the methods in the context of two-stage designs although the methods readily extend to more than two stages.

Before discussing the methods we develop some additional notation. Let  $n^{(i)}$ ,  $i = 1, 2$ , be the sample sizes proposed initially for the two stages of the trial, and let  $n^{(1)} + n^{(2)} = N_{\max}$ . For example, in Plan 4,  $n^{(1)} = 200$ ,  $n^{(2)} = 67$  and  $N_{\max} = 267$ . The data on the  $n^{(2)}$  stage-2 patients are assumed to be independent of the data on the  $n^{(1)}$  stage-1 patients. Let  $X_{je}^{(i)}$  denote the response of the  $j$ th patient entering the trial at stage- $i$  and randomized to the experimental arm. Let  $X_{jc}^{(i)}$  denote the corresponding response for the control arm. We assume these responses are normally distributed with respective means  $\mu_e$  and  $\mu_c$ , and with a common variance  $\sigma^2$ . As defined previously, the effect size is  $\delta = \mu_e - \mu_c$ . The maximum likelihood estimate of  $\delta$  based on the stage- $i$  data alone is

$$\hat{\delta}^{(i)} = \left[ \frac{\sum_{j=1}^{0.5n^{(i)}} X_{je}^{(i)}}{0.5n^{(i)}} \right] - \left[ \frac{\sum_{j=1}^{0.5n^{(i)}} X_{jc}^{(i)}}{0.5n^{(i)}} \right] \quad (7)$$

and its standard error is

$$\text{se}(\hat{\delta}^{(i)}) = \sqrt{\frac{4\sigma^2}{n^{(i)}}} \quad (8)$$

(We are assuming equal patient allocation to the two treatment arms and ignoring the minor difficulty that arises if  $n^{(i)}$  is odd.) The Wald statistic computed from the stage- $i$  data is

$$Z^{(i)} = \frac{\hat{\delta}^{(i)}}{\text{se}(\hat{\delta}^{(i)})} \quad (9)$$

In this notation, we attach superscripts to statistics if they are constructed from stage  $i$  data alone, and subscripts if they are constructed from the cumulative data available at stage  $i$ . Therefore, that the stage-1 statistics  $\hat{\delta}^{(1)}$ ,  $\text{se}(\hat{\delta}^{(1)})$  and  $Z^{(1)}$  are, respectively, identical to the statistics  $\hat{\delta}_1$ ,  $\text{se}(\hat{\delta}_1)$  and  $Z_1$  that appear in equation (5) of Section 2.2. On the other hand, the stage-2 statistics  $\hat{\delta}^{(2)}$ ,  $\text{se}(\hat{\delta}^{(2)})$  and  $Z^{(2)}$  are not the same as the cumulative statistics  $\hat{\delta}_2$ ,  $\text{se}(\hat{\delta}_2)$ , and  $Z_2$  that appear in equation (6) of Section 2.2.

Finally, the maximum likelihood estimate of  $\delta$  based on all  $N_{\max}$  subjects is

$$\hat{\delta}_2 = \left[ \frac{\sum_{j=1}^{0.5n^{(1)}} X_{je}^{(1)} + \sum_{j=1}^{0.5n^{(2)}} X_{je}}{0.5N_{\max}} \right] - \left[ \frac{\sum_{j=1}^{0.5n^{(1)}} X_{jc}^{(1)} + \sum_{j=1}^{0.5n^{(2)}} X_{jc}}{0.5N_{\max}} \right] \tag{10}$$

its standard error is

$$se(\hat{\delta}_2) = \sqrt{\frac{4\sigma^2}{N_{\max}}} \tag{11}$$

and the corresponding Wald statistic is

$$Z_2 = \frac{\hat{\delta}_2}{se(\hat{\delta}_2)} \tag{12}$$

### 3.1. Down-weighting the test statistic

This method was proposed by Cui *et al.* [9]. First we observe that if  $\sigma^2$  is known then the Wald statistic (12) based on all  $N_{\max}$  patients can be expressed as a weighted sum of the two independent Wald statistics for the two stages:

$$Z_2 = \sqrt{\frac{n^{(1)}}{n^{(1)} + n^{(2)}}} Z^{(1)} + \sqrt{\frac{n^{(2)}}{n^{(1)} + n^{(2)}}} Z^{(2)} \tag{13}$$

When  $\sigma^2$  is unknown we replace it in equation (8) by its unbiased estimate based on the  $n^{(i)}$  observations at stage- $i$ , and in equation (11) by its unbiased estimate based on all  $N_{\max}$  observations. In that case, equation (13) no longer holds. However, the distribution of the random variable  $Z_2$  is asymptotically normal with mean  $\delta\sqrt{N_{\max}}/(2\sigma)$  and variance 1 whether  $\sigma^2$  or its estimates from the relevant data sets are used in equation (13). Thus, hereafter we will proceed as though  $\sigma^2$  is known.

Suppose that the sample size at stage 2 is increased from  $n^{(2)}$  to  $n^{(2*)}$ . Denote the corresponding stage-2 statistics by  $\hat{\delta}^{(2*)}$ ,  $se(\hat{\delta}^{(2*)})$ , and  $Z^{(2*)}$ . Define

$$Z_2^* = \sqrt{\frac{n^{(1)}}{n^{(1)} + n^{(2*)}}} Z^{(1)} + \sqrt{\frac{n^{(2*)}}{n^{(1)} + n^{(2*)}}} Z^{(2*)} \tag{14}$$

If the new sample size  $n^{(2*)}$  at stage 2 is selected independent of the stage-1 statistics  $\hat{\delta}^{(1)}$  and  $Z^{(1)}$ , then, under  $H_0: \delta = 0$ , the random vectors  $(Z_1, Z_2^*)$  and  $(Z_1, Z_2)$  have identical multivariate normal distributions. Now the type 1 error for a two-stage group sequential design with stopping boundaries  $b_1$  and  $b_2$  at stages 1 and 2, respectively, is  $\alpha$  provided  $b_1$  and  $b_2$  satisfy the condition

$$P_0(|Z_1| \geq b_1) + P_0(|Z_1| < b_1, |Z_2| \geq b_2) = \alpha \tag{15}$$

where  $P_0(\cdot)$  denotes probability under the null hypothesis. When  $b_1$  and  $b_2$  are computed by the  $\alpha$ -spending method of Lan and DeMets [5], equation (15) is automatically satisfied. For example,

in Plan 4, equation (15) is satisfied with  $b_1 = 2.34$  and  $b_2 = 2.01$ . For the adaptive design, since the sample size at stage 2 has been increased from  $n^{(2)}$  to  $n^{(2*)}$ , we require

$$P_0(|Z_1| \geq b_1) + P_0(|Z_1| < b_1, |Z_2^*| \geq b_2) = \alpha \quad (16)$$

in order to preserve the type 1 error. But since the distributions of  $(Z_1, Z_2^*)$  and  $(Z_1, Z_2)$  are identical under  $H_0$ , equation (16) is indeed satisfied.

Next consider the case where  $n^{(2*)}$ , the augmented stage-2 sample size, does depend on the stage-1 results,  $\hat{\delta}^{(1)}$  and  $Z^{(1)}$ . In this case,  $(Z_1, Z_2^*)$  and  $(Z_1, Z_2)$  have different covariances even under  $H_0$  because the weight  $\sqrt{n^{(2*)}/(n^{(1)} + n^{(2*)})}$  assigned to  $Z^{(2*)}$  in equation (14) has been obtained in a data-dependent manner. Therefore, equation (16) no longer holds. Cui *et al.* [9] have observed, however, that the random vector  $(Z_1, Z_2^\dagger)$ , where

$$Z_2^\dagger = \sqrt{\frac{n^{(1)}}{n^{(1)} + n^{(2)}}} Z^{(1)} + \sqrt{\frac{n^{(2)}}{n^{(1)} + n^{(2)}}} Z^{(2*)} \quad (17)$$

does indeed have the same distribution as the statistic  $(Z_1, Z_2)$  under the null hypothesis. In other words the type 1 error is preserved even if the stage-2 sample size is increased in a data-dependent manner provided the weights assigned to the two components of  $Z_2^\dagger$  are the same as the weights assigned to the two components of  $Z_2$ .

Table I displays the weights assigned to the two components of  $Z_2^*$  and  $Z_2^\dagger$  for different choices of  $n^{(2*)}$  if Plan 4 is made adaptive with  $n^{(1)} = 200$  and  $n^{(2*)} \geq n^{(2)} = 67$ . Effectively, the Cui *et al.*'s [9] method preserves the type 1 error by down-weighting the contribution of the patients who enter the study at stage 2, relative to those who enter the study at stage 1. For example, if, as a result of observing  $\hat{\delta}^{(1)} = 1.8$ , we decide to increase the stage-2 sample size from 67 to 700 we would continue to use the weights (0.865, 0.501) that are appropriate for a two-stage group sequential test in which the stage-1 sample size is 200 and the stage-2 sample size is 67 instead of using the weights (0.471, 0.882) that are appropriate for a two-stage group sequential test in which the stage-1 sample size is 200 and the stage-2 sample size is 700. The last two columns of the table display the conditional power or probability of achieving statistical significance if we observe  $\hat{\delta}^{(1)} = 1.8$  at stage 1, assume that this is the true value of  $\delta$ , and then increase the stage-2 sample size to  $n^{(2*)}$ . Table I reveals that by down-weighting the stage-2 contribution we have incurred a slight loss of conditional power. This is the price to be paid for ensuring that the type 1

Table I. Impact of down-weighting on conditional power ( $\hat{\delta}^{(1)} = 1.8$ ).

$n^{(2*)}$	Weights for $Z_2^*$	Weights for $Z_2^\dagger$	Conditional power	
	$\left( \sqrt{\frac{n^{(1)}}{n^{(1)} + n^{(2*)}}}, \sqrt{\frac{n^{(2*)}}{n^{(1)} + n^{(2*)}}} \right)$	$\left( \sqrt{\frac{n^{(1)}}{n^{(1)} + n^{(2)}}}, \sqrt{\frac{n^{(2)}}{n^{(1)} + n^{(2)}}} \right)$	$Z^*$ (%)	$Z^\dagger$ (%)
67	(0.865, 0.501)	(0.865, 0.501)	14	14
500	(0.535, 0.845)	(0.865, 0.501)	67	58
700	(0.471, 0.882)	(0.865, 0.501)	78	71
900	(0.426, 0.905)	(0.865, 0.501)	86	81

error will be preserved despite a data-dependent increase in sample size. In technical terms,  $Z_2^*$  is the *sufficient statistic* for inference about  $\delta$ . By using  $Z_2^\dagger$  rather than  $Z_2^*$  we are violating the sufficiency principle (see, for example, Reference [10]). The resulting inference procedure is in a certain sense sub-optimal, as shown explicitly by Tsiatis and Mehta [3].

3.2. Preserving the conditional type 1 error

Müller and Schäfer [11] have developed a very general and flexible method for implementing adaptive group sequential designs. To illustrate their method consider again the two-stage adaptive design denoted in Section 2.2 as Plan 4. An interim look is taken at stage 1, after observing data on 200 evaluable subjects. Suppose the stage 1 estimate of  $\delta$  is  $\hat{\delta}_1 = 1.8$ . Assuming as before that  $\sigma = 10$ , the standard error is  $se(\hat{\delta}_1) = 2\sigma/\sqrt{200} = 1.414$  and thus the corresponding Wald statistic is  $z_1 = 1.27$ . Since  $|z_1| < b_1 = 2.34$ , the trial continues. Unless an adaptive change is made the study will be terminated after enrolling an additional 67 subjects. In order to declare statistical significance upon termination the Wald statistic  $|Z_2|$ , based on data from all 267 evaluable patients, must equal or exceed  $b_2 = 2.012$ . The Müller and Schäfer [11] method involves computing conditional rejection probabilities (CRP's), defined as the null probabilities of crossing the upper and lower final stopping boundaries, conditional on the observed  $z_1$ . Specifically, the upper CRP is

$$\mathcal{E}_u(z_1) = \Pr(Z_2 \geq b_2 | Z_1 = z_1, \delta = 0) \tag{18}$$

and the lower CRP is

$$\mathcal{E}_l(z_1) = \Pr(Z_2 \leq -b_2 | Z_1 = z_1, \delta = 0) \tag{19}$$

These two CRP's are the only items of information of that must be carried over from the first part of the study into the second part. Müller and Schäfer [11] have shown that the second part of the study may be an entire new trial with sample size, number and spacing of interim analyses, even the choice of error spending functions, all adaptively determined. The only requirement for the new trial is that it should have an asymmetric two-sided type 1 error with error probability  $\mathcal{E}_u$  in the positive direction and error probability  $\mathcal{E}_e$  in the negative direction. For Plan 4 the observed Wald statistic at the end of stage 1 with 200 evaluable subjects is  $z_1 = 1.27$ . The final stopping boundary with 67 additional subjects is  $b_2 = 2.012$ . Substituting these numbers into equations (18) and (19) we obtain  $\mathcal{E}_u(z_1) = 0.0346$  and  $\mathcal{E}_l(z_1) \approx 0$ . By the conditional rejection probability principle of Müller and Schäfer [11] we are free to design a new trial with any sample size and error spending function of our choosing provided we restrict the one-sided type 1 error of the new trial to 0.0346. Table II displays the values  $\mathcal{E}_u(z_1)$  for different values of  $z_1$  observed at the end of stage 1. For

Table II.  $\mathcal{E}_l(z_1)$  versus  $z_1$  for Plan 4 with  $\sigma = 10$ .

$\hat{\delta}_1$	$z_1$	$\mathcal{E}_u$
1.4	0.99	0.011
1.8	1.27	0.035
2.2	1.56	0.092
2.6	1.84	0.201
3.0	2.12	0.363

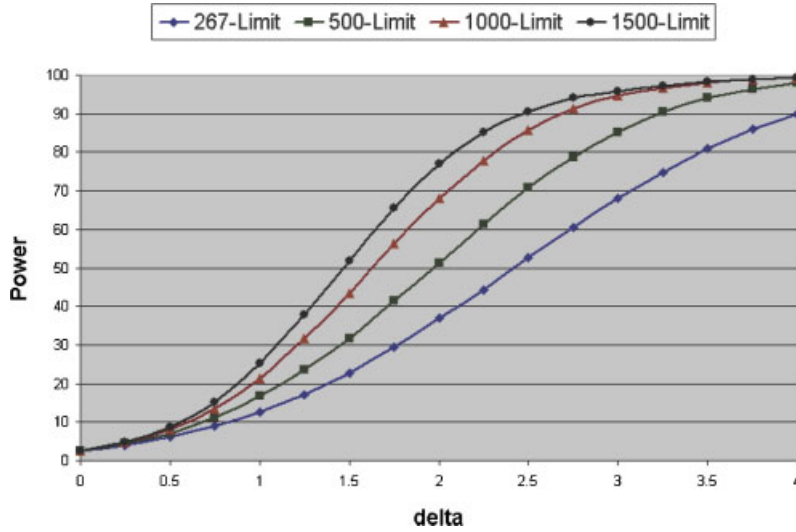


Figure 3. Unconditional power charts for different sample size limits.

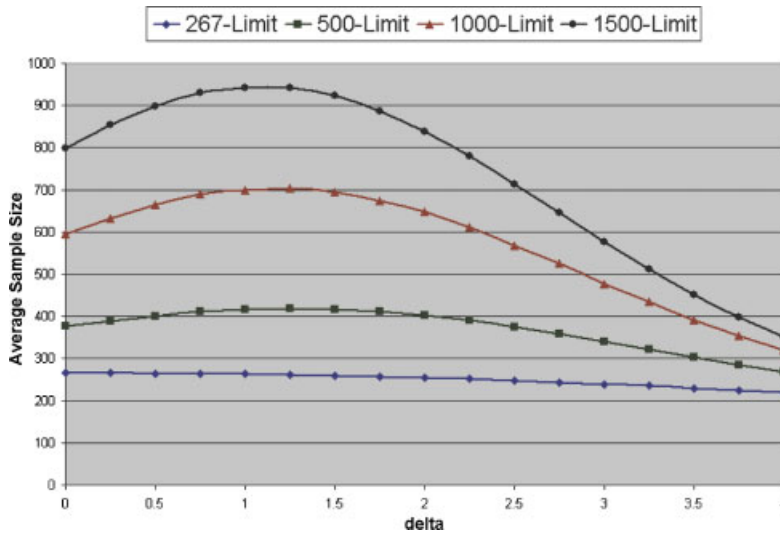


Figure 4. ASN charts for different sample size limits.

these choices of  $z_1$ ,  $\mathcal{E}_l(z_1) \approx 0$ . Notice how  $\mathcal{E}_u(z_1)$  increases with  $z_1$ . This shows that the more evidence one gathers against the null hypothesis at stage 1, the less evidence one requires to reject the null hypothesis in the new trial.

### 3.3. Unconditional power and ASN charts

Figure 3 displays unconditional power charts for the various two-stage adaptive designs applied to Plan 4. In these designs the sample size is increased so as to recover the conditional power after

observing the data at stage 1, subject to an upper limit on the total sample size of the trial. The final hypothesis test utilizes the down-weighted test statistic of Cui *et al.* [9] to preserve the type 1 error. The chart labelled '267-limit' is the power chart for the two-stage group sequential design with no adaptive increase in sample size. As one would expect the unconditional power increases as the upper limit on the total sample size is increased. Figure 4 shows that these increases in unconditional power come at the cost of fairly large increases in expected sample size for the adaptive designs relative to the non-adaptive group sequential design. The power and ASN charts for the other adaptive procedure are very similar to the ones shown here.

#### 4. TRADING OFF FLEXIBILITY FOR EFFICIENCY: SOME EXAMPLES

Recent articles by Jennison and Turnbull [4] and Tsiatis and Mehta [3] suggest that adaptive designs are less efficient than classical group sequential designs. Jennison and Turnbull [4] selected some typical two-stage adaptive designs and constructed corresponding multi-stage group sequential designs with uniformly higher power and uniformly lower expected sample size over the entire space of the alternative hypothesis. Tsiatis and Mehta [3] showed more generally that for any given adaptive design one can always construct a standard group sequential likelihood ratio test that, for any parameter value in the space of alternatives, will reject the null hypothesis earlier with higher probability, and, for any parameter value not in the space of alternatives, will accept the null hypothesis earlier with higher probability. These results would argue against using adaptive designs for sample size re-estimation. What then is their appeal?

The main reason that adaptive designs have generated so much interest is their greater flexibility compared to more classical fixed sample or group sequential designs. This flexibility is often so attractive that trial sponsors are willing to accept some loss of efficiency in exchange for it. In the remainder of this section we provide some examples that illustrate this trade-off. To preserve confidentiality in a competitive environment we have omitted the names of the sponsors and compounds being tested.

1. *Limited experience.* A major pharmaceutical company wishes to launch a clinical trial for testing a new compound in schizophrenia patients with negative symptoms. No previous negative symptoms trial has been carried out in a predominantly negative symptoms population. The two endpoints of interest are the negative symptoms assessment score (NSA) and the quality of life score (QLS). Both measures have been little used. There is neither consensus on which measure should be made the primary endpoint, nor any agreement regarding the magnitude of improvement that could be considered clinically meaningful. Under the circumstances it would be appear advisable to consider a two-stage adaptive design in which the study is re-designed after the data from stage 1 have been examined. One could, of course, run two separate studies for this purpose. In that case, there would be a gap of several months between closing the first trial and opening the second. Also, since one would not be permitted to combine the data from the two trials for the final inference, there is a loss of efficiency. The adaptive methodology discussed in Section 3 might be more attractive because it can save time and because it utilizes the data from both stages for the final inference.
2. *Avoidance of ambiguous results.* Many clinical trials end without a clear conclusion. For example, if a registration trial ends up with a  $p$ -value of 0.07 for the primary endpoint then, from a regulatory perspective, the trial has failed to demonstrate efficacy. There exists the

possibility, however, that the reason for the failure is that the trial was underpowered. The sponsor is now faced with the expensive and time consuming option of running a second trial on a larger group of subjects. Clearly, the alternative option of increasing the sample size adaptively is more appealing and efficient than starting a new trial.

Good timing is essential for this adaptive strategy to be successful. The interim look should be taken as near the currently planned end of the trial as possible so that the interim results are unlikely to change unless the trial is extended. In the past regulators have expressed concern when design changes were proposed near the end of a trial.

3. *Improved standard of care.* In a large cardiovascular trial patients are randomized between optimal background therapy plus placebo or a new agent. The primary endpoint is survival. To ensure adequate power, the study is required to remain open until a pre-specified maximum number of deaths have been observed. Since activation of the study, however, new evidence from other trials has been published documenting that the standard of care for patients in this disease area has improved substantially. It is plausible that overall survival on the placebo arm might be 20 per cent higher than was thought to be the case when the study was designed. If this is true the study may be underpowered. It might be beneficial to examine this premise at the next interim look. If the survival rate for the control arm appears to be substantially superior to what was anticipated at the design stage, but no corresponding gain is observed on the treatment arm, one could increase the maximum number of deaths, and thus recover the power of the study.
4. *The integrated phase II/III trial.* For many indications it is common to run a phase II trial that includes several doses of the new compound along with an active comparator. The best phase II arm and the active comparator are then carried forward to a separate phase III trial. It is possible to accelerate this sequence and conduct it more efficiently by designing a two-stage adaptive trial in which the optimal dose of the new compound and the stage-2 sample size are determined from the stage-1 data. Technical details of this two-stage procedure are available in Reference [12]. See also Reference [13].

## 5. DECISION THEORY AND SAMPLE SIZE DETERMINATION

Section 3 presented various ways to change the sample size of an on-going trial after observing the interim results without inflating the type 1 error. We now address the question of how to determine what the new sample size should be. At the design stage, before any data have been collected, the primary determinant of sample size is the power desired to detect a clinically meaningful effect size. At the interim analysis stage, in an adaptive trial, one has the opportunity to revise the sample size based on internal and external information available at that time. The usual way that the revision is accomplished is through the use of conditional power. In an industry trial, however, economic considerations and conditional power considerations might not always coincide. Consider again Plan 4 for the depression study. The study has 90 per cent power to detect  $\delta = 4$  with a maximum sample size of 267 and one interim look after observing data on 200 subjects. Suppose that we observe  $\hat{\delta}_1 = 2$  at the interim look. The conditional power under  $\delta = \hat{\delta}_1$  for various values of the stage-2 sample size are displayed in Table III along with the additional study duration assuming an enrolment rate of 10 patients/week.

If sample size is not increased, the trial will end in six weeks. However, the conditional power is only 22.5 per cent. By extending the trial, the conditional power can be increased, but at

Table III. The trade-off between conditional power and additional study duration assuming an enrollment rate of 10 patients/week.

Stage-2 sample size	Conditional power (%)	Additional study duration (weeks)
67	22.5	6
400	66.5	40
600	81	60
800	89.5	80

Table IV. Declining exclusivity periods by product class. Source, CGEY 2002 report.

Product class	Years exclusive	Product class	Years exclusive
Beta blocker	9	SSRI	3
H2 blocker	3	PPRI	2
Ace inhibitor	6	Red blood cell	1
Anti-histamine	4	Anti-emesis	4
RTI	4	A2 antagonist	2
Cholesterol lowering	5	Cox-2 inhibitor	0.4

the cost of prolonging the duration of the trial. In a registration trial, especially, the sponsor faces a difficult decision problem: on the one hand, prolonging the trial duration will cut into the period of market exclusivity for the compound; on the other hand, if the sample size is not increased, the trial might fail to reach statistical significance. For example, one could strive for 80 per cent conditional power and lose 60 weeks of market exclusivity or strive for 90 per cent conditional power and lose 80 weeks of market exclusivity. Which is better? Moreover, this decision problem is especially acute because the period of market exclusivity for new products is being eroded by the emergence of advanced technologies that make it possible to design drugs with similar or improved medical profiles to existing drugs without infringing their patents. Table IV, taken from the Cap-Gemeni Earnest and Young (CGEY) 2002 annual report, shows the declining period of exclusivity that novel products, costing in excess of \$600M to develop, are experiencing.

The traditional frequentist statistical methods cannot help with this trade-off. They can only ensure that the type 1 error will be preserved, no matter what decision is taken. On the other hand, this decision problem fits very naturally into the framework of Bayesian inference and decision analysis. It is therefore worth considering a hybrid approach wherein the criterion for declaring statistical significance is based entirely on the frequentist methodology described in Section 3 but the magnitude of the sample size increase is determined by maximizing expected net present value (NPV) through a decision tree. Figure 5 displays one possible decision tree that can help determine the optimal sample size at stage 2.

Let us examine the components of this decision tree from left to right. At stage 1 we select a sample size  $n^{(1)}$  and observe a standardized statistic  $z^{(1)}$ . Although the choice of  $n^{(1)}$  could be incorporated into the backward induction step to be described later, we will assume for simplicity that  $n^{(1)}$  is determined by conventional frequentist power considerations. For example, under Plan 4

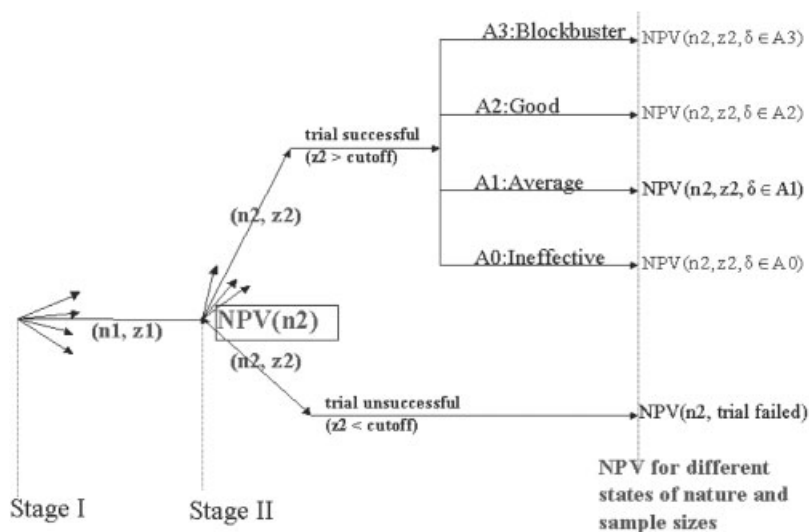


Figure 5. Decision tree for maximizing expected NPV.

of Section 2.2,  $n^{(1)} = 200$ . If  $|z^{(1)}|$  equals or exceeds the stage-1 stopping boundary of 2.34, the trial is terminated and there is no further need to construct a decision tree. Therefore, we will assume that  $|z^{(1)}| < 2.34$ . The decision problem consists in determining the optimal value for  $n^{(2*)}$ , the stage-2 sample size. In the remainder of this section, for notational convenience we will drop the '\*' and denote  $n^{(2*)}$  by  $n^{(2)}$ .

Suppose we have selected a particular value for  $n^{(2)}$  and have observed the standardized statistic  $z^{(2)}$ . The decision tree now unfolds in two directions depending on whether the trial was unsuccessful or successful.

*Trial unsuccessful.* If  $z_2$  is less than the critical cut-off for declaring statistical significance at stage 2, the trial is unsuccessful. The cash flow of this outcome is negative. It is simply the cost of the  $n^{(2)}$  subjects who were enrolled at stage 2 and may be assumed to increase linearly with  $n^{(2)}$ . This cost, discounted to the net present value, is denoted on the decision tree as  $\text{NPV}(n_2, \text{trial failed})$ .

*Trial successful.* If, on the other hand,  $z^{(2)}$  equals or exceeds the critical cut-off, the trial is successful and we may assume that the compound will eventually generate a positive cash flow. It is reasonable to assume that this positive cash flow depends on the true value of  $\delta$ . If  $\delta$  is large, the compound will be very successful in the marketplace whereas if  $\delta$  is small, even though the clinical trial was positive and the compound was brought to market, it may not generate large revenues. As a concrete example we consider the cash flow that would be generated if the depression trial was successful. For convenience we will partition  $\delta$  in four distinct intervals  $A_j$ , for  $j = 0, \dots, 3$ , denoting blockbuster, good, average and ineffective products, respectively. Let the peak annual revenues, while the compound is still in the exclusivity period, be denoted by  $u(A_j)$ .

*Blockbuster:*  $A_3 = \{\delta: \delta > 4\}$ ,  $u(A_3) = \$900\text{M/yr}$

*Good:*  $A_2 = \{\delta: 3 < \delta \leq 4\}$ ,  $u(A_2) = \$600\text{M/yr}$

*Average:*  $A_1 = \{\delta: 1 < \delta \leq 3\}$ ,  $u(A_1) = \$300\text{M/yr}$

*Ineffective:*  $A_0 = \{\delta: \delta \leq 1\}$ ,  $u(A_0) = \$50\text{M/yr}$

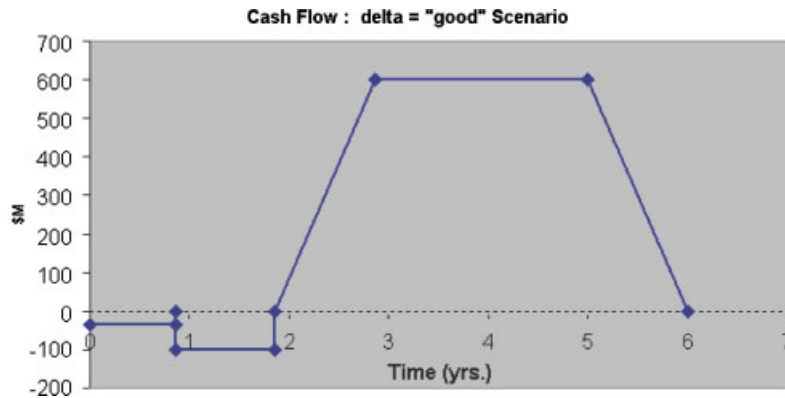


Figure 6. Shape of cash flow function.

The shape of the cash flow function is the same for all four categories, but the peak sales are different as specified above. As an example, the shape of the cash flow function when  $\delta \in A_2$ , is displayed in Figure 6.

There are three distinct parts to this cash flow function. There is an initial constant negative cash flow rate attributed to the cost per subject while the stage 2 portion of the clinical trial is still underway, enrolling  $n^{(2)}$  subjects. After the trial is terminated, if the outcome is a success, there is a second period of negative cash flow, which can be attributed to the cost of preparing for regulatory approval and for launching the product in the marketplace. After that the product generates a positive cash flow that climbs steadily to the peak value  $u(A_j)$  and then declines down to zero at the end of the exclusivity period. The total net profit contribution generated by the compound is the area under the curve. This quantity is discounted to net present value and denoted on the ends of the decision tree as  $NPV(n_2, z_2, \delta \in A_j)$ . It should be evident from the shape of the cash flow function that if the stage 2 sample size is over-extended, the area under the curve will decrease and the NPV will decrease because of the reduction in profit as well as by the increased discount due to these profits accruing later in time. This is shown in Figure 7 where the duration of the stage 2 portion of the clinical trial has been extended by increasing  $n^{(2)}$ .

The analysis of the decision tree proceeds in the usual way with a prior-to-posterior analysis in the forward direction to obtain probabilities on the branches of the tree, and a backward induction to obtain expected net present value (see, for example, Reference [14]).

*Prior-to-posterior analysis.* In the Bayesian paradigm  $\delta$  is a random variable with prior

$$f_0(\delta) \sim N(a_0, b_0)$$

At the end of stage-I, after we observe  $z_1$ , the posterior distribution of  $\delta$  is obtained by Bayes theorem as

$$f_1(\delta|z^{(1)}) \propto f_0(\delta)l(z^{(1)}|\delta)$$

where  $l(\cdot|\delta)$  is the likelihood function, treated as a function of  $\delta$ . Let  $n^{(2)}$  be the sample size and  $z^{(2)}$  be the outcome of stage 2. The predictive distribution of  $Z^{(2)}$  is

$$f_{\text{pred}}(z^{(2)}) = \int_{\delta} f(z^{(2)}|\delta) f_1(\delta) d\delta$$

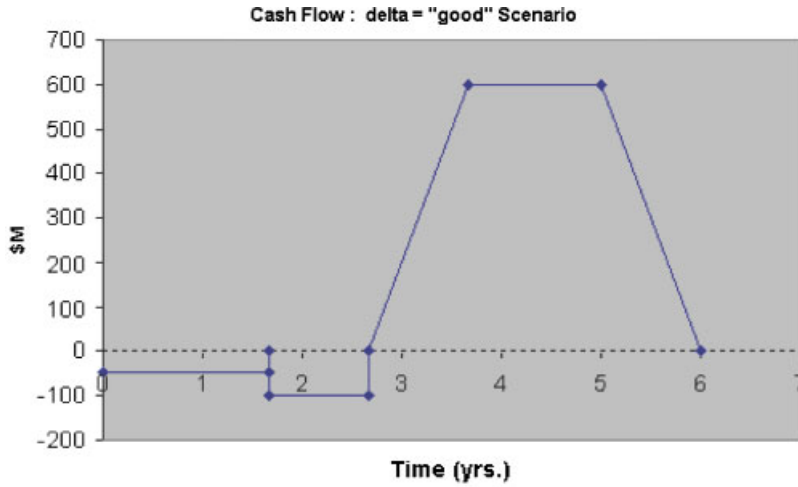


Figure 7. Drop in cash flow if stage 2 is extended.

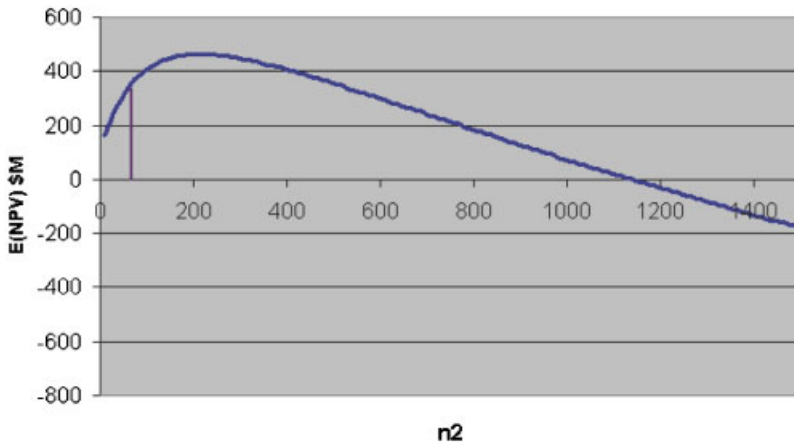


Figure 8. Plot of  $\bar{NPV}(n_2)$  versus  $n_2$  when  $\hat{\delta}_1 = 2$ .

After we have observed  $z^{(2)}$ , the posterior distribution of  $\delta$  is obtained by again applying Bayes theorem:

$$f_2(\delta|z^{(2)}) \propto f_1(\delta|z_1)l(z^{(2)}|\delta)$$

*Backward induction.* First we take the expectation over posterior distribution of  $\delta$  given  $z^{(2)}$

$$\begin{aligned} \bar{NPV}(n^{(2)}, z^{(2)}) &= \sum_{j=0}^3 NPV(n^{(2)}, z^{(2)}, \delta \in A_j) \Pr(\delta \in A_j|n^{(2)}, z^{(2)}) \\ &\quad + NPV0 \Pr(Z^{(2)} < \text{cut-off}) \end{aligned}$$

Then we take the expectation over predictive distribution of  $Z^{(2)}$

$$\text{N}\bar{\text{P}}\text{V}(n^{(2)}) = \int_{-\infty}^{\infty} \text{N}\bar{\text{P}}\text{V}(z^{(2)}, n^{(2)}) f_{\text{pred}}(z^{(2)}) dz^{(2)}$$

By computing  $\text{N}\bar{\text{P}}\text{V}(n^{(2)})$  in this manner for different values of  $n^{(2)}$  the optimal value may be determined. Figure 8 displays a plot of  $\text{N}\bar{\text{P}}\text{V}(n^{(2)})$  versus  $n^{(2)}$  for the depression trial when  $n^{(1)} = 200$  and  $\hat{\delta}_1 = 2$ . We used a very flat prior (normal with mean equal to 2 and standard deviation equal to 10). The discount rate used to compute net present value was 15 per cent per annum.

The optimal value of  $n^{(2)}$  is 270 new subjects at stage 2. It is interesting to contrast this number with the value of  $n^{(2)}$  that would be required to increase the conditional power at stage 2 to 80 per cent. The required sample size is 580.

## 6. CONCLUDING REMARKS

Sample size re-estimation to ensure that a study is adequately powered can be handled naturally within the group sequential framework where the analysis of interim data is already built into the process. It is important, however, to distinguish between re-estimation based on revised estimates of nuisance parameters like the variance of the response variable and re-estimation because of second thoughts about the value of  $\delta$  at which to power the trial. In the first type of re-estimation the traditional group sequential methodology is utilized without any modification, with Fisher information playing the role of sample size. The maximum information required to provide  $1 - \beta$  power for a group sequential level- $\alpha$  test to detect a difference  $\delta_1$  is determined at the design stage by the formula

$$I_{\max} = \left[ \frac{z_{\alpha} + z_{\beta}}{\delta} \right]^2 \times \text{IF} \quad (20)$$

where IF is an inflation factor whose value depends on  $\alpha$ ,  $\beta$ , the number of interim looks, and a pre-specified  $\alpha$ -spending function. Observe that the computation of  $I_{\max}$  by equation (20) does not involve any unknown nuisance parameters whereas the computation of  $N_{\max}$  by equation (4) requires specification of the nuisance parameter  $\sigma^2$ . The study is monitored at administratively convenient times with information at the  $j$ th interim look being estimated by

$$I_j = [\text{se}(\hat{\delta}_j)]^{-2}$$

The information fraction at look  $j$  is thus  $t_j = I_j / I_{\max}$ , the cumulative amount of type 1 error that may be spent by look  $j$  is given by  $\alpha(t_j)$ , and the corresponding stopping boundary is derived by inverting this cumulative error as discussed in Reference [5]. In a maximum information trial the maximum sample size,  $N_{\max}$ , need not be fixed in advance. The study remains open with a floating sample size until either  $I_{\max}$ , the primary determinant of statistical power, is attained or a stopping boundary is crossed. Examples of maximum information group sequential trials for longitudinal and survival studies are provided in References [15, 16], respectively. For a more detailed expository discussion of maximum information group sequential trials see Reference [2].

The key feature of maximum information group sequential trials is that while sample size may float,  $I_{\max}$  is computed once and for all at the design stage, by equation (20), and may not be changed during the course of the trial. Adaptive trials, in contrast, permit  $I_{\max}$  to be altered after the trial is underway. The idea of altering the maximum information itself after the trial is underway on the basis of an interim estimate of  $\delta$  is relatively new. The first papers to introduce this idea appear to be those of Bauer and Kohne [17], Proschan and Hunsberger [7], and Fisher [18]. These authors did not deal with group sequential trials, however, Cui *et al.* [9] were the first to extend the approach to the group sequential setting. Since these pioneering papers were published, the literature on this topic has exploded. By some counts there are now more than 100 papers on adaptive design in print. Despite this evidence of huge interest in the topic, there do not appear to have been many instances of actual clinical trials in which these methods were implemented. It is possible that efficiency considerations, as presented in Reference [4] or Reference [3], are partially responsible for this hesitancy to adopt adaptive designs. We have seen in Section 4, however, that some loss of efficiency is acceptable in exchange for greater flexibility to make adaptive changes to the trial design. Thus, it is more likely that sponsors of industry trials are concerned primarily about the regulatory implications of adopting these new adaptive designs. There are no published examples, standard operating procedures or guidance documents that can be cited should the regulatory authorities raise questions about their validity. Serious logistical issues, such as who should have the authority to examine the interim data and make adaptive changes to the study design, have never been discussed in a public forum. It seems likely that for a while, each adaptive proposal will be considered on a case by case basis. It will be the responsibility of the sponsor to demonstrate prior to implementation that the trial adheres to sound statistical principles and carries no risk of leakage of interim results to trial investigators.

The primary appeal of the adaptive design is the flexibility to make data-dependent changes in mid-stream if conditions warrant the change. The manner in which this flexibility will be exercised must be pre-specified and documented in the study protocol so as to ensure that the type 1 error will be preserved. However, it is important to clarify exactly what needs to be pre-specified in the protocol and what can be decided later, after the study has been activated. The protocol must declare the method that will be used to preserve the type 1 error, should a data-dependent change be made after the trial is activated. We stress, however, that the protocol does *not* have to make an *a priori* commitment to increase the sample size. Nor does the protocol have to pre-specify the manner in which the new sample size will be determined, should the decision to increase the sample size be taken at a later time. That decision might be best made at the time of the interim analysis itself, by a combination of internal data, external information from other trials, market conditions, and clinical judgement.

Finally, we wish to comment on the Bayesian decision theoretic methodology discussed in Section 5. This is a promising approach new approach that needs much further investigation and detailed inputs from the business and economic departments of a pharmaceutical corporation. It represents a first attempt to capture the real business issues of the sponsor through the specification of a realistic cost function. At the same time this approach utilizes the traditional frequentist paradigm for preserving the type 1 error and so is less likely to face resistance at the regulatory level.

#### ACKNOWLEDGEMENTS

The authors wish to thank Professor Butch Tsiatis for many insightful discussions concerning this problem.

## REFERENCES

1. Wittes J, Brittain E. The role of internal pilot studies in increasing efficiency of clinical trials. *Statistics in Medicine* 1990; **9**:65–72.
2. Mehta CR, Tsiatis AA. Flexible sample size considerations using information based interim monitoring. *Drug Information Journal* 2001; **35**:1095–1112.
3. Tsiatis AA, Mehta CR. On the inefficiency of the adaptive design for monitoring clinical trials. *Biometrika* 2003; **90**:367–378.
4. Jennison C, Turnbull BW. Mid-course sample size modification in clinical trials based on the observed treatment effect. *Statistics in Medicine* 2003; **22**:971–993.
5. Lan KKG, DeMets DL. Discrete sequential boundaries for clinical trials. *Biometrika* 1983; **70**:659–663.
6. East 3.1. *Software for Design Simulation and Interim Monitoring of Flexible Clinical Trials*. Cytel Software Corporation: Cambridge, MA, 2004.
7. Proschan MA, Hunsberger SA. Designed extension of studies based on conditional power. *Biometrics* 1995; **51**:1315–1324.
8. Lan KKG, Trost DC. Estimation of parameters and sample size reestimation. *Proceedings of the Biopharmaceutical Section of the American Statistical Association*, 1997; 48–51.
9. Cui L, Hung HMJ, Wang S-J. Modification of sample size in group sequential trials. *Biometrics* 1999; **55**: 853–857.
10. Cox DR, Hinkley DV. *Theoretical Statistics*. Chapman & Hall: New York, 1986.
11. Müller HH, Schäfer H. Adaptive group sequential designs for clinical trials: combining the advantages of adaptive and classical group sequential approaches. *Biometrics* 2001; **57**:886–891.
12. Bauer P, Kieser M. Combining different phases in the development of medical treatments within a single trial. *Statistics in Medicine* 1999; **18**:1833–1848.
13. Stallard N, Todd S. Sequential designs for phase III trials incorporating treatment selection. *Statistics in Medicine* 2003; **22**(5):689–703.
14. Raiffa H, Schlaifer R. *Applied Statistical Decision Theory*. MIT Press: Cambridge, MA, 1968.
15. Scharfstein DO, Tsiatis AA, Robins JM. Semiparametric efficiency and its implication on the design and analysis of group-sequential studies. *Journal of the American Statistical Association* 1997; **92**:1342–1350.
16. Scharfstein DO, Tsiatis AA. The use of simulation and bootstrap in information-based group sequential studies. *Statistics in Medicine* 1998; **17**:75–87.
17. Bauer P, Kohne K. Evaluation of experiments with adaptive interim analysis. *Biometrics* 1994; **50**:1029–1041.
18. Fisher LD. Self-designing clinical trials. *Statistics in Medicine* 1998; **17**:1551–1562.