Mid-course sample size modification in clinical trials based on the observed treatment effect *

Christopher Jennison
Department of Mathematical Sciences,
University of Bath, Bath BA2 7AY, U. K.
email: cj@maths.bath.ac.uk, *tel:* +44 (0)1225 826468, *fax:* +44 (0)1225 323436
and
Bruce W. Turnbull
Department of Statistical Science,
227 Rhodes Hall, Cornell University, Ithaca, New York 14853-3801, U. S. A.
email: turnbull@orie.cornell.edu, *tel:* +1 607 255 9131, *fax:* +1 607 255 9129

SUMMARY

It is not uncommon to set the sample size in a clinical trial to attain specified power at a value for the treatment effect deemed likely by the experimenters, even though a smaller treatment effect would still be clinically important. Recent papers have addressed the situation where such a study produces only weak evidence of a positive treatment effect at an interim stage and the organizers wish to modify the design in order to increase the power to detect a smaller treatment effect than originally expected. Raising the power at a small treatment effect usually leads to considerably higher power than was first specified at the original alternative.

Several authors have proposed methods which are not based on sufficient statistics of the data after the adaptive re-design of the trial. We discuss these proposals and show in an example how the same objectives can be met while maintaining the sufficiency principle, as long as the eventuality that the treatment effect may be small is considered at the design stage. The group sequential designs we suggest are quite standard in many ways but unusual in that they place emphasis on reducing the expected sample size at a parameter value under which extremely high power is to be achieved. Comparisons of power and expected sample size show that our proposed methods can out-perform L. Fisher's "variance spending" procedure. Although the flexibility to re-design an experiment in mid course may be appealing, the cost in terms of the number of observations needed to correct an initial design may be substantial.

Key words: Clinical trials; Group sequential tests; Sample size re-estimation; Two-stage procedure; Flexible design; Variance spending.

*Paper presented at 4th International Meeting on Statistical Methods in Biopharmacy, Paris Sept 24-25, 2001

1 Introduction

We consider a situation which is not unusual in Phase III clinical trials that involve the comparison of a new treatment with a placebo or standard therapy. A statistical design is specified in the protocol based in part on specification of a Type I error rate, α, and power $1-\beta$ at a given effect size, δ. The design may be for a fixed sample size or there may be provision for early stopping via a group sequential monitoring boundary. At some intermediate point during the course of the trial, the principal investigators examine the outcome data collected so far and decide they wish to modify the original design. They may, for example, have been over-optimistic in the choice of the design effect size, δ, whereas it is now apparent that the benefit of the new treatment is liable to be somewhat less than δ and it is unlikely that a significant result will be achieved at the planned end of the trial. Even so, the estimated effect may still be large enough to be deemed clinically significant and worthwhile.

At this stage, the question is often posed to the study statistician (and perhaps to a regulatory body, such as the FDA) whether the trial design can be modified and the sample size enlarged, without violating the trial's credibility and statistical validity. In the past, a strict answer to this question was usually "No". The alternative strategy of abandoning a trial if early results appear unpromising and starting a new trial can also lead to grossly inflated Type I error rates — in extreme, this is akin to "sampling to a foregone conclusion" (Cornfield[1]). However, recently Fisher,[2] and Shen & Fisher[3] have proposed the so-called "variance spending" method whereby the sample size and other features of the current trial can be modified while maintaining the α level, even though these modifications were unplanned at the start of the trial. Other authors have proposed similar and related methods that can adapt a design to interim outcome data. These include Bauer & Köhne,[4] Proschan & Hunsberger,[5] Lan & Trost,[6] Cui, Hung & Wang,[7] Lehmacher & Wassmer,[8] Chi & Liu,[9] Denne,[10,11] Müller & Schäfer[12,13] and Wang, Hung, Tsong & Cui.[14] Wassmer[15] summarizes and reviews many of these articles. In some of these papers, the authors describe methods in which the design is changed in response to interim results according to *pre-specified* rules; other methods offer greater freedom to adapt to interim data in an *unplanned* way. Fisher[2] emphasizes that the variance spending method allows mid-course changes in trial design "undreamed of" before the study started and such considerations lead him to term these "self-designing" trials.

In the next section we describe an example of a trial in which examination of response data at an unplanned interim analysis suggests a larger study should have been conducted. In Section 3 we show how Fisher's variance spending method can overcome this problem, we explain its equivalence to other

methods and present new methodology for deriving a confidence interval on termination; for the most part we confine attention to the two-stage method proposed by Fisher,[2] but we make occasional remarks on multi-stage versions. In order to assess the efficiency of the variance spending method it is necessary to consider its use with a fully specified rule for revising sample size. In Section 3.4 we present a typical version of a sample size rule and in Section 3.5 we discuss possible inefficiencies in the variance spending approach due to use of a non-sufficient statistic and variability in the final sample size. In Section 4 we calculate the overall power function and average sample size function of this procedure and show that in this example flexibility does not come without a price. In comparison, we present standard group sequential procedures which provide the same power for fewer observations on average: these procedures could have been used had the experimenters considered the possibility of a smaller effect size and agreed on a minimal clinically significant effect size at the initial design stage. Our conclusion is that more careful initial planning can lead to significant savings in resources. Although flexible procedures allow a mid-course correction to be made in a statistically valid manner, it is better still to determine the correct objective at the start.

Note that the sample size re-estimation procedures we consider here should not be confused with those used when there is an unknown nuisance parameter such as a response variance. There is already an extensive literature on this topic; see, for example, the recent paper by Whitehead et al[16] or, for a survey, Jennison & Turnbull,[17] Chapter 14. The procedures are also not to be confused with adaptive designs where treatment allocation proportions can be varied depending on accumulating results. The even more numerous papers on this topic are surveyed in Jennison & Turnbull,[17] Chapter 17.

2 Problems posed by unplanned interim analyses

Consider a balanced two-sample comparison in which observations X_{Ai} on treatment A and X_{Bi} on treatment B, $i = 1, 2, \ldots$, are independent, normally distributed with common variance σ^2 and means μ_A and μ_B, respectively. We assume the variance is known and, without loss of generality, take σ^2 to be 0.5. The parameter of interest is the unknown difference in treatment means, $\theta = \mu_A - \mu_B$, and it is desired to test the null hypothesis H_0: $\theta = 0$ against the one-sided alternative $\theta > 0$ with Type I error probability α. Although this problem may seem unrealistically simple, it does in fact serve as a prototype for a wide variety of responses, and methods developed for this situation have wide applicability; see, for example, Jennison & Turnbull,[17] Chapter 3.

We suppose the experimenters initially plan a fixed sample test attaining power $1 - \beta$ at the alternative

$\theta = \delta$. This requires a sample size

$$n = \frac{(z_\alpha + z_\beta)^2}{\delta^2} \tag{1}$$

per treatment arm (recall $2\sigma^2 = 1$) where $z_p = \Phi^{-1}(1-p)$ denotes the upper p tail point of the standard normal distribution.

Now suppose the data are examined at an intermediate stage of the trial when a fraction r of the planned observations have been collected. Denote the estimate of θ computed from the rn observations per treatment accumulated so far by

$$\widehat{\theta}_1 = \frac{1}{rn} \sum_{i=1}^{rn} (X_{Ai} - X_{Bi}).$$

Consider the not uncommon scenario where $\widehat{\theta}_1$ is positive but somewhat smaller than the effect size δ at which power $1-\beta$ was specified. If the true value of θ is close to $\widehat{\theta}_1$, it is unlikely that H_0 will be rejected — the conditional power at $\theta = \widehat{\theta}_1$ is low. However, the experimenters now realize that the magnitude of $\widehat{\theta}_1$ *is* clinically meaningful and the original target effect size, δ, was over-optimistic. In retrospect, they regret not designing the test to have power $1-\beta$ at $\theta = \widehat{\theta}_1$ rather than at $\theta = \delta$. This would have required the larger sample size $\xi^2 n$, where $\xi = \delta/\widehat{\theta}_1$.

A naive approach to "rescue" this trial would be simply to increase the number of remaining observations on each arm from $(1-r)n$ to $\gamma(1-r)n$, where γ is chosen so that $rn + \gamma(1-r)n = \xi^2 n$, i.e.,

$$\gamma = \frac{\xi^2 - r}{1-r},$$

and proceed to use the naive final test statistic

$$Z = \frac{1}{\sqrt{(\xi^2 n)}} \sum_{i=1}^{\xi^2 n} (X_{Ai} - X_{Bi}).$$

However, since the random variable ξ is a function of the first stage data, this Z statistic does not follow a $N(0,1)$ distribution under H_0 and the test that rejects H_0 when $Z > z_\alpha$ does *not* have Type I error rate α. Cui et al[7] show that, typically, the Type I error rate of such a test is inflated by 30% to 40%; using other rules to determine the second stage sample size, it can more than double — see Proschan, Follmann & Waclawiw,[18] Table 4, Proschan & Hunsberger[5] and Shun et al.[19]

The experimenters may consider the alternative option of ignoring the data collected so far and starting a completely new trial with a larger sample size. Not only is this wasteful of data but, as noted in Section 1, persistent use of this strategy is liable to produce an excess of false positive results in a manner akin to the process of sampling to a foregone conclusion discussed by Cornfield.[1]

3 L. Fisher's variance spending approach

3.1 Definition

Fisher[2] has proposed a method which allows changes to the sample size at an unplanned interim analysis while still preserving the Type I error rate. At an intermediate stage when rn $(0 < r < 1)$ observations have been observed on each treatment arm, define

$$S_1 = \sum_{i=1}^{rn} (X_{Ai} - X_{Bi}),$$

then

$$S_1 \sim N(rn\theta,\, rn) \quad \text{and} \quad W_1 = \frac{S_1}{\sqrt{n}} \sim N(r\sqrt{n}\,\theta,\, r). \tag{2}$$

In the variance spending framework, W_1 is said to have spent a fraction r of the total variance 1 in the final z-statistic.

Suppose the second stage sample size is now changed from $(1-r)n$ to $\gamma(1-r)n$, where $\gamma > 0$. One might, for example, choose γ to give a total sample size that would attain a certain power at $\theta = \widehat{\theta}_1$ or to meet a conditional power requirement at $\theta = \widehat{\theta}_1$. We shall consider a specific rule for choosing γ later but it should also be remembered that Fisher's method allows a free choice of γ without reference to any pre-specified rule.

Let $n^* = rn + \gamma(1-r)n$ denote the total sample size per treatment arm, after revision, and define the second stage variables

$$S_2 = \sum_{i=rn+1}^{n^*} (X_{Ai} - X_{Bi})$$

and

$$W_2 = \frac{\gamma^{-1/2} S_2}{\sqrt{n}}\,.$$

Then, *conditional on the first stage data*,

$$S_2 \sim N(\gamma(1-r)n\theta,\, \gamma(1-r)n)$$

and

$$W_2 \sim N(\sqrt{\gamma}(1-r)\sqrt{n}\,\theta,\, 1-r). \tag{3}$$

The key point to note is that under H_0: $\theta = 0$, we have $W_2 \sim N(0, 1-r)$ whatever the data-dependent choice of γ, so W_2 is independent of the first stage data and of W_1. (For independence results in a more general adaptive design framework, see Liu, Proschan and Pledger.[20])

The variance of W_2 , $1 - r$, is the remaining part of the total variance 1 not already used by W_1. The variance spending test statistic is

$$Z = W_1 + W_2 = \frac{S_1 + \gamma^{-1/2} S_2}{\sqrt{n}} \tag{4}$$

and this has a $N(0, 1)$ distribution under H_0. Rejecting H_0 when $Z > z_\alpha$ maintains the Type I error probability α, even though γ depends on the first stage data.

To see that this test has power greater than α for $\theta > 0$, write $W_1 = r\sqrt{n}\,\theta + Y_1$ where $Y_1 \sim N(0,\, r)$ and $W_2 = \sqrt{\gamma}(1-r)\sqrt{n}\,\theta + Y_2$ where $Y_2 \sim N(0,\, 1-r)$. Here, γ is a positive random variable dependent on Y_1, but Y_2 is independent of Y_1. Thus, for $\theta > 0$,

$$Z \;=\; W_1 + W_2 \;>\; Y_1 + Y_2 \;\sim\; N(0,\, 1)$$

and so the probability that Z exceeds z_α and H_0 is rejected is greater than α.

The variance spending approach can be extended to allow more than one re-design point. Fisher[2] and Shen & Fisher[3] describe a multi-stage procedure in which a final z-statistic is created from a sequence of statistics $W_1, \ldots, W_K$ (where K is not necessarily fixed in advance) based on adaptively weighted sample sums. Under H_0, the conditional distribution of each W_k given its predecessors is normal with zero mean and their conditional variances sum to one. It follows by, for example, embedding the partial sums $W_1 + \ldots + W_k$ in a standard Brownian motion, that $Z = W_1 + \ldots + W_K \sim N(0, 1)$.

3.2 Equivalent methods

Cui et al[7] propose a method for modifying group sizes in the later stages of a group sequential test while maintaining the original Type I error rate. The following description generalizes their procedure to allow more than one re-design point — although in practice at most one change to a trial's protocol may well be desirable. In the original design, the K group sizes are $r_1 n, \ldots, r_K n$, where n is the maximum sample size chosen for the group sequential test and $r_1 + \ldots + r_K = 1$. In each group $k = 1, \ldots, K$, observations are summarized by the statistic

$$W_k = \frac{S_k}{\sqrt{n}} \;\sim\; N(r_k\sqrt{n}\,\theta,\, r_k)$$

where S_k is the difference between sums of responses on treatments A and B. When the group k sample size is increased by a factor γ_k in response to data in groups 1 to $k-1$, the new statistic

$$W_k = \frac{\gamma_k^{-1/2} S_k}{\sqrt{n}} \tag{5}$$

is formed which has a $N(\gamma_k^{-1/2} r_k \sqrt{n}\,\theta,\ r_k)$ conditional distribution. Thus, under H_0: $\theta = 0$, the distribution of each W_k is unaffected by the change in group size and the original group sequential boundary can still be used to give a test with Type I error rate α. The factor $\gamma_k^{-1/2}$ in (5) applies the same weighting as in Fisher's variance spending approach. With a single re-design point, the factors γ_k are equal to one for a certain number of groups, l say, and then change to a new, common value for groups $l+1$ to K. Using Cui et al's method in a group sequential test with two stages and no early stopping at the first stage produces exactly Fisher's two-stage variance spending test.

If a formal interim analysis is included in a trial protocol, one would expect the experimenters to consider the option of stopping to reject or to accept H_0 at the interim analysis. When a variance spending test is adopted because of an unplanned interim analysis, such early stopping is not strictly allowable. However, if interim results are very negative, one may decide for ethical or economic reasons to stop for "futility" with acceptance of H_0; indeed, Shen & Fisher[3] advocate the use of such a futility boundary. This can only reduce the Type I error, producing a conservative test.

Another method for adapting a group sequential test to deal with data-dependent modifications to group sizes has been proposed by Denne[11] and Müller & Schäfer.[12,13] The key to this method is preserving the conditional Type I error probability, given current data and the original experimental design, when the future course of the trial is changed. The following calculation shows that applying this principle at stage one of a two-stage group sequential test with re-calculation of the remaining sample size but no stopping at the first stage yields Fisher's variance spending test. With the notation of Section 3.1, if $S_1 = s_1$ is observed after rn pairs of observations, the conditional Type I error probability in the original design is

$$P_{\theta=0}\left\{\frac{S_1 + S_2}{\sqrt{n}} > z_\alpha \mid S_1 = s_1\right\},$$

where $S_2 \sim N(0,\ (1-r)n)$ under $\theta = 0$, and this probability is equal to

$$A(s_1) = 1 - \Phi\left(\frac{z_\alpha}{\sqrt{(1-r)}} - \frac{s_1}{\sqrt{\{(1-r)n\}}}\right). \tag{6}$$

With a new second stage sample size of $\gamma(1-r)n$ per treatment arm, $S_2 \sim N(0,\ \gamma(1-r)n)$ under $\theta = 0$ and probability $A(s_1)$ is maintained by rejecting H_0 for

$$\frac{S_2}{\sqrt{\{\gamma(1-r)n\}}} \ > \ \Phi^{-1}\{1 - A(s_1)\},$$

a condition which simplifies to

$$\frac{s_1 + \gamma^{-1/2} S_2}{\sqrt{n}} \ > \ z_\alpha,$$

just as in Fisher's test.

There is good reason why all these methods should be equivalent. Integrating over the distribution of S_1, we can write the Type I error probability of the original test as

$$\int_{-\infty}^{\infty} A(s_1) f(s_1)\, ds_1 = \alpha$$

where $A(s_1)$ is the conditional Type I error probability given $S_1 = s_1$ defined in (6) and f is the $N(0,\, rn)$ density of S_1. In an adaptive approach, the second stage sample size may be revised after learning the value of S_1, then, after observing S_2, H_0 is accepted or rejected according to a rule stated in terms of S_1 and S_2. Our concern is the nature of this rule. Suppose that under a certain rule the conditional Type I error probability given $S_1 = s_1$ and $\gamma(1-r)n$ pairs of observations in stage two is $B(s_1, \gamma)$. In order to retain Type I error probability α whatever system is used to choose the second stage sample size, we require

$$\int_{-\infty}^{\infty} B(s_1, \tilde{\gamma}(s_1)) f(s_1)\, ds_1 = \alpha \tag{7}$$

for any positive function $\tilde{\gamma}$. If no change is made to the initial design, the original test will be used so we know $B(s_1, 1) = A(s_1)$. Suppose there is a set of values s_1 of positive Lebesgue measure for which $B(s_1, \gamma^*(s_1)) > A(s_1)$ for some positive function γ^*, then defining $\tilde{\gamma}(s_1) = \gamma^*(s_1)$ on this set and $\tilde{\gamma}(s_1) = 1$ otherwise would make the left hand side of (7) greater than α, so we can deduce that such a set of s_1 values does not exist. Likewise, there can be no similar set on which $B(s_1, \gamma^*(s_1)) < A(s_1)$. It follows that for each s_1 (with the possible exception of a set of measure zero), $B(s_1, \gamma)$ is independent of γ and equal to $A(s_1)$ — and the preceding discussion shows the final decision rule is therefore that of Fisher's variance spending test.

Several authors, including Wassmer,[21] Chi & Liu[9] and Posch & Bauer,[22] have described the two-stage tests of Bauer & Köhne,[4] Proschan & Hunsberger[5] and others in a common framework. In these procedures, the second stage design is chosen in the light of first stage outcomes and data from the two stages are combined according to a pre-specified rule. Responses from each stage can be summarized by a P-value or z-statistic for testing the null hypothesis. Working in terms of the z-statistics Z_1 and Z_2 calculated from stage one and stage two data respectively, a conditional Type I error function $C(z_1)$ is defined with the property

$$\int_{-\infty}^{\infty} C(z_1) \phi(z_1)\, dz_1 = \alpha \tag{8}$$

where $\phi(x)$ denotes the standard normal density at x. Having observed $Z_1 = z_1$, H_0 is rejected in favor of

$\theta > 0$ after stage two if

$$Z_2 > \Phi^{-1}\{1 - C(z_1)\}$$

or, equivalently, if the stage two P-value is less than $C(z_1)$. The condition (8) ensures the overall Type I error rate is equal to α. It may seem surprising that Fisher's variance spending test can be described in the same manner since it is applicable in the absence of any initial intent to use a two-stage procedure. In this case, the fixed sample analysis originally planned determines the conditional Type I error function $A(s_1)$ defined in (6) and this plays the same role as $C(z_1)$ above.

3.3 P-values and confidence intervals

It is useful to augment the result of a hypothesis test by stating a P-value for testing the null hypothesis and a confidence interval for the parameter of interest. In a two-stage variance spending test with no early stopping at the first stage, defining a P-value of the observed data for testing H_0: $\theta = 0$ is straightforward. Since the standardized statistic Z defined by (4) has a standard normal distribution under H_0, the one-sided P-value for testing H_0 against $\theta > 0$ is simply

$$p = 1 - \Phi(Z).$$

Shen & Fisher[3] (p. 197) note that their method does not provide an estimate of θ. In the following, we overcome the complication that γ, and hence the mean of S_2, depends on S_1 to produce a confidence interval for θ. Our derivation generalizes to Shen & Fisher's multi-stage setting. In a two-stage variance spending test, as defined in Section 3.1,

$$S_1 - rn\theta \sim N(0,\, rn)$$

and

$$\gamma^{-1/2} S_2 - \sqrt{\gamma}\,(1-r)n\theta \sim N(0,\, (1-r)n)$$

independently of S_1. Thus,

$$S_1 + \gamma^{-1/2} S_2 - \{r + \sqrt{\gamma}\,(1-r)\}n\theta \sim N(0,\, n)$$

so

$$P_\theta \left\{ -z_\alpha \sqrt{n} \;\le\; S_1 + \gamma^{-1/2} S_2 - \{r + \sqrt{\gamma}\,(1-r)\}n\theta \;\le\; z_\alpha \sqrt{n} \right\} = 1 - \alpha$$

and, by the usual pivoting argument,

$$\frac{S_1 + \gamma^{-1/2} S_2}{\{r + \sqrt{\gamma}\,(1-r)\}n} \;\pm\; \frac{z_\alpha}{\{r + \sqrt{\gamma}\,(1-r)\}\sqrt{n}}$$

is a $1 - \alpha$ confidence interval for θ.

This confidence interval can also be derived by inverting a family of tests of hypotheses $H\colon\ \theta = \tilde{\theta}$ where the critical region of each test is defined using the conditional error probability argument applied in testing H_0 in Section 3.2. This method has the advantage of extending to the adaptively re-designed group sequential tests of Cui et al.[7] To do this, we start with a test of a single value $\theta = \tilde{\theta}$ with data collected under the original group sequential design. A $1 - \alpha$ critical region is constructed based on, say, the stage-wise ordering of the sample space (see Jennison & Turnbull,[17] Section 8.4 and references therein). At the time of an adaptive re-design, one computes the conditional probability of rejecting $\theta = \tilde{\theta}$ if the study were to continue as originally planned and then maintains this conditional probability in setting up the critical region for a test of $\theta = \tilde{\theta}$ in the modified design, using the same rule to order the remaining sample space. On termination, the confidence interval comprises all values $\theta = \tilde{\theta}$ which have not been rejected in their individual tests.

Brannath, Posch and Bauer[23] present more general methods for obtaining P-values and confidence intervals when repeated design adaptations are allowed. See also Liu and Chi,[24] Section 6.

A more troublesome problem arises if unplanned early stopping is introduced at an interim stage, such as stopping for futility with acceptance of H_0 when a positive result looks very unlikely. It is then unclear what the space of outcomes that could have arisen really is (to specify this, one needs to say what the experimenters would have decided to do in every possible eventuality) and this appears to preclude construction of a confidence interval with the required frequentist coverage probability.

3.4 A rule for choosing γ

The prime motivation for variance spending tests and related methods is the desire for flexible modification of a trial design in response to intermediate results when no such adaptation was originally planned. Nevertheless, it is helpful to consider formal rules for how such adaptation might be carried out. Examining the overall properties of response adaptive designs conducted according to specific rules will aid in understanding these methods and help assess the cost, in terms of efficiency, of this flexibility.

As before, we denote by $\widehat{\theta}_1 = S_1/(rn)$ the estimate of θ computed from rn observations per treatment arm observed at an unplanned, intermediate analysis and define $\xi = \delta/\widehat{\theta}_1$. Fisher[2] discusses a strategy for

choosing γ to obtain conditional power $1-\beta$, given the data observed so far, if in fact $\theta = \delta/\xi = \widehat{\theta}_1$. In the variance spending test, the conditional power given $S_1 = s_1$ under $\theta = \widehat{\theta}_1$ is

$$\Phi\left\{\frac{\{r+\sqrt{\gamma}(1-r)\}\sqrt{n}\,\widehat{\theta}_1 - z_\alpha}{\sqrt{(1-r)}}\right\}$$

and equating this probability to $1-\beta$ gives

$$\gamma = \frac{(\sqrt{1-r}\,z_\beta + z_\alpha - r\sqrt{n}\,\widehat{\theta}_1)^2}{(1-r)^2 n\,\widehat{\theta}_1^2}. \tag{9}$$

Some truncation may be necessary to ensure that γ is positive but does not exceed a reasonable upper bound.

We shall pursue the alternative proposal by Cui et al[7] of equating the total sample size to that which would achieve *unconditional* power $1-\beta$ for a true value of θ equal to $\widehat{\theta}_1$, but we adapt this rule to allow for the special weighting of the second stage data. Recall from (2) and (3) that in the statistic $Z = W_1 + W_2$,

$$W_1 \sim N(r\sqrt{n}\,\theta,\, r)$$

and, conditionally on the first stage data,

$$W_2 \sim N(\sqrt{\gamma}(1-r)\sqrt{n}\,\theta,\, 1-r).$$

A fixed sample test designed from the outset to achieve power $1-\beta$ at $\theta = \delta/\xi$ would have $\xi^2 n$ observations per treatment arm and use a standardized test statistic Z' with distribution

$$Z' \sim N(\xi\sqrt{n}\,\theta,\, 1).$$

Equating the mean of Z' with the sum of $E(W_1)$ and the conditional expectation of W_2 gives

$$\xi = r + \sqrt{\gamma}(1-r)$$

or, equivalently,

$$\gamma = \left(\frac{\xi - r}{1-r}\right)^2. \tag{10}$$

If $\widehat{\theta}_1$ is small and positive, ξ and γ will be very large. Thus, it is advisable to bound the value of ξ used in (10). Since a small positive value of θ may give rise to negative values of $\widehat{\theta}_1$, the maximal value of γ is also appropriate for negative $\widehat{\theta}_1$. (However, if $\widehat{\theta}_1$ is sufficiently low, one may choose to abandon the trial for futility and stop at this point with acceptance of H_0.) If $\widehat{\theta}_1 > \delta$, then $\xi < 1$ and the above rule leads to

$\gamma < 1$, i.e., a decrease in the second stage sample size. This is quite acceptable but at least some truncation is necessary to keep γ positive. With these modifications, we obtain the general rule

$$\gamma(\widehat{\theta}_1) = \left(\frac{\tilde{\xi}(\widehat{\theta}_1) - r}{1 - r} \right)^2 \tag{11}$$

where

$$\tilde{\xi}(\widehat{\theta}_1) = \begin{cases} M & \text{for} \quad \widehat{\theta}_1/\delta \leq M^{-1} \\ \delta/\widehat{\theta}_1 & \text{for} \quad M^{-1} < \widehat{\theta}_1/\delta \leq h^{-1} \\ h & \text{for} \quad \widehat{\theta}_1/\delta > h^{-1}. \end{cases} \tag{12}$$

The values of γ generated by this rule lie in the range $(h-r)^2/(1-r)^2$ to $(M-r)^2/(1-r)^2$. When $h = 1$, no decrease is allowed from the originally planned sample size, n.

3.5 Properties of variance spending tests

By design, a variance spending test has Type I error probability α. Further properties depend on how the sample size inflation factor γ is chosen in the light of first stage data. The fact that the final test statistic Z defined by (4) is not a function of a sufficient statistic for θ gives some cause for concern. Of course, the unequal weighting of first and second stage observations is necessary to ensure independence of W_1 and W_2 and, indeed, the argument of Section 3.2 shows the final test *must* have this form if Type I error rate α is to be maintained when unplanned design changes take place. Nevertheless, it is instructive to make comparisons with trial designs the experimenters could have chosen had they anticipated the possibility of a smaller effect size before commencing the study.

An initial measure of possible inefficiency can be obtained from the derivation of the rule for choosing γ in Section 3.4. There, we noted that the sample size needed for a fixed sample test designed to achieve power $1 - \beta$ at $\theta = \delta/\xi$ is $\xi^2 n$ per treatment arm, where n is given by (1). In contrast, a variance spending test adapting an initial design with power $1-\beta$ at $\theta = \delta$ when an estimate $\widehat{\theta}_1 = \delta/\xi$ is observed at an interim analysis requires $n^* = \{r + \gamma(1-r)\}n$ observations per arm where $\gamma = (\xi - r)^2/(1-r)^2$. A measure of inefficiency of the variance spending test is thus

$$\frac{\{r + \gamma(1-r)\}n}{\xi^2 n} \;=\; \left\{ r + \frac{(\xi - r)^2}{1 - r} \right\} \frac{1}{\xi^2}\,. \tag{13}$$

Table 1 shows numerical values of this measure for the case $r = 0.5$, i.e., when the trial is re-designed after half the originally planned sample size. In the limit as $\xi \to \infty$, the second stage term W_2 contributes

Table 1: A measure of inefficiency of a variance spending test with $r = 0.5$, as given by equation (13), and the relative cost of re-starting the trial afresh with increased power.

ξ	0.5	0.6	0.8	1	2	3	4	10	∞
Inefficiency of Z	2	1.44	1.06	1	1.25	1.44	1.56	1.81	2
Relative cost to re-start	3	2.39	1.78	1.50	1.12	1.06	1.03	1.01	1

essentially all the information about θ and this is diluted by adding W_1 which has the same variance but, by comparison, negligible information about θ; the situation is reversed as ξ decreases to $r = 0.5$ where, in the limit, all the information about θ comes from W_1.

Particularly when ξ is much greater than 1, experimenters may be tempted to abandon the original experiment, discard the observations, and start a new fixed sample trial with power $1 - \beta$ at $\theta = \delta/\xi$. This new trial would require $\xi^2 n$ observations per treatment arm in addition to the rn in the abandoned study. The "relative cost" in the bottom line of Table 1 is the ratio of the total sample size, $rn + \xi^2 n$, involved in this strategy to the sample size of $\xi^2 n$ needed by a fixed sample test designed from the outset with power $1 - \beta$ at $\theta = \delta/\xi$. When $(\xi - 1)^2 > 1 - r$, i.e., when $\xi > 1.71$ for the case $r = 0.5$, starting a fresh trial would be more efficient than using the variance spending test. However, as mentioned previously, this is *not* really a valid option since it inflates the overall Type I error rate.

The "inefficiencies" in Table 1 are suggestive of the cost of using a non-sufficient statistic in the variance spending method. They do not, however, take account of the variability in $\widehat{\theta}_1$ as an estimate of θ and the resulting random distribution of the factor γ. A proper assessment of the overall performance of a variance spending test requires integration over the distribution of $\widehat{\theta}_1$, which is normal with mean θ and variance $1/(rn)$. We present such integrals for the overall power and ASN functions below and we use these criteria in assessing the example in Section 4.

If anything, the variation in second stage sample size could have a detrimental effect. Consider a study with a random sample size of N observations per treatment arm, where N is not itself influenced by the observations' values. A hypothesis test of H_0: $\theta = 0$ conducted with Type I error rate α conditional on the value of N has power

$$E\{\Phi(\sqrt{N}\,\theta - z_\alpha)\}.$$

Since $\Phi(x)$ is an increasing, concave function of x for $x > 0$ and $\sqrt{N}\,\theta - z_\alpha$ is concave in N, the conditional

power $\Phi(\sqrt{N}\,\theta - z_\alpha)$ is concave in N when $\sqrt{N}\,\theta - z_\alpha > 0$, i.e., for values of N which give power at least 0.5. It follows by Jensen's inequality that, when N varies in this range, overall power is maximized if N is equal to its expectation with probability one, i.e., when sample size does not in fact vary. Under the initial design, there is a good chance of distinguishing between the cases $\theta = 0$ and $\theta = \delta$ using a sample of n observations per treatment arm. At an intermediate stage with only a fraction of these observations, $\widehat{\theta}_1$ is liable to vary over the range zero to δ, leading to considerable variation in the sample size implied by (11) and (12). We should not, therefore, be surprised if a variance spending test has rather low power for its expected sample size.

The power of the variance spending test can be calculated as

$$P_\theta\{\text{Reject } H_0\} = P_\theta\{Z > z_\alpha\} = \int P_\theta\{Z > z_\alpha \mid \widehat{\theta}_1\} f_\theta(\widehat{\theta}_1)\, d\widehat{\theta}_1. \tag{14}$$

It follows from the definition of Z and the distribution of W_2 stated in (3) that

$$P_\theta\{Z > z_\alpha \mid \widehat{\theta}_1\} = \Phi\left\{ \frac{r\sqrt{n}}{\sqrt{(1-r)}}\,\widehat{\theta}_1 + \sqrt{\{\gamma(\widehat{\theta}_1)(1-r)n\}}\,\theta - \frac{z_\alpha}{\sqrt{(1-r)}} \right\}.$$

The density of $\widehat{\theta}_1$ is

$$f_\theta(\widehat{\theta}_1) = \sqrt{\frac{rn}{2\pi}}\ \exp\left\{ -\frac{rn}{2}(\widehat{\theta}_1 - \theta)^2 \right\}$$

and hence (14) can be evaluated numerically. The expected sample size per treatment arm or Average Sample Number (ASN) is

$$\mathit{ASN} = E(n^*) = rn + (1-r)n \int \gamma(\widehat{\theta}_1) f_\theta(\widehat{\theta}_1)\, d\widehat{\theta}_1$$

which, again, is readily evaluated by numerical integration.

In the next section we shall apply the above formulae to evaluate the power function and ASN curve of a representative example of a variance spending test. We then use these results to assess the price one may have to pay for the flexibility of the variance spending approach as compared to other less flexible procedures.

4 An example

4.1 Sampling and stopping rules

We shall use the following example to evaluate a typical variance spending test by the standard criteria of power and expected sample size functions. The original plan is for a fixed sample test and sample

size is modified at an intermediate analysis using the adaptation of Cui et al's[7] sampling rule described in Section 3.4; early stopping for futility introduced at the interim analysis allows H_0 to be accepted for sufficiently poor responses. Although the results presented here are for this single example, we have found very similar results in a variety of other examples using different values of r, M and h, or calculating second stage sample size by the conditional power formula (9) proposed by Fisher.[2]

As before, observations on treatments A and B are distributed as $X_{Ai} \sim N(\mu_A, 0.5)$ and $X_{Bi} \sim N(\mu_B, 0.5)$, interest is in the parameter $\theta = \mu_A - \mu_B$, and the null hypothesis H_0: $\theta = 0$ is to be tested against the one-sided alternative $\theta > 0$ with Type I error rate $\alpha = 0.025$. In the non-sequential test originally planned, power $1 - \beta = 0.9$ is set at $\theta = \delta$, requiring a sample size

$$n = \frac{(z_\alpha + z_\beta)^2}{\delta^2} = \frac{10.51}{\delta^2}$$

per treatment arm. Intermediate data are examined halfway through the trial, i.e., $r = 0.5$, and the second stage sample size is inflated by the factor $\gamma(\widehat{\theta}_1)$ defined by (11) and (12) using $M = 4$ and $h = 0.5$.

Specifically, $\widehat{\theta}_1$ is calculated from the first $n/2$ observations per treatment, we define

$$\xi = \begin{cases} 4 & \text{for} \quad \widehat{\theta}_1/\delta \leq 0.25 \\ \delta/\widehat{\theta}_1 & \text{for} \quad 0.25 < \widehat{\theta}_1/\delta \leq 2 \\ 0.5 & \text{for} \quad \widehat{\theta}_1/\delta > 2, \end{cases}$$

and a further $\gamma\, n/2$ observations are taken on each arm where

$$\gamma = 4\,(\xi - 0.5)^2.$$

The second stage sample increases if $\widehat{\theta}_1 < \delta$, remains unchanged if $\widehat{\theta}_1 = \delta$, and decreases if $\widehat{\theta}_1 > \delta$. The inflation factor γ lies in the interval $(0, 49)$ and the total sample size, $n^* = (1 + \gamma)n/2$, is bounded by $(0.5 + 49 \times 0.5)n = 25n$.

The null hypothesis is rejected in favor of $\theta > 0$ if

$$Z \;=\; \frac{S_1 + \gamma^{-1/2} S_2}{\sqrt{n}} \;>\; z_{0.025}$$

where

$$S_1 = \sum_{i=1}^{n/2} (X_{Ai} - X_{Bi}) \quad \text{and} \quad S_2 = \sum_{i=n/2+1}^{n^*} (X_{Ai} - X_{Bi}).$$

As it stands, whenever $\widehat{\theta}_1 < 0.25\,\delta$ this rule gives $\xi = 4$ and $\gamma = 49$, the value associated with a test achieving power 0.9 at $\theta = 0.25\,\delta$. In order to save sample size when there is little prospect of a

positive outcome, we add a futility boundary at the first stage which stipulates stopping to accept H_0 if the conditional probability of rejecting H_0 under $\theta = 0.25\,\delta$ and with $\gamma = 49$ is less than 0.8, a condition which is met when $\widehat{\theta}_1/\delta < -0.1735$.

At the other extreme, substituting $\widehat{\theta}_1 = 2\,\delta$ into the formula $\xi = \delta/\widehat{\theta}_1$ gives $\xi = 0.5$ and $\gamma = 0$. The $N(\sqrt{\gamma}\,n/2, n/2)$ distribution of $\gamma^{-1/2}S_2$ tends to a $N(0, n/2)$ distribution as γ approaches zero so we simply take $\gamma^{-1/2}S_2 \sim N(0, n/2)$ for the case $\xi = 0.5$ arising when $\widehat{\theta}_1 > 2\,\delta$. In practice one might prefer to take a single observation on each treatment — but as γ will be small, the expectation of $\gamma^{-1/2}S_2$ will be close to zero and the main role of this term is still to contribute the required amount to the variance of Z.

4.2 Power and ASN functions

Figure 1 shows the power function of the variance spending test along with that of the original fixed sample size test. It is evident that the variance spending test has been successful in increasing power over the range of θ values. After a sharp initial rise, its power function increases slowly as θ moves from around $0.3\,\delta$ to δ and the overall shape of the power curve is quite different from that of any fixed sample test.

The argument of Section 3.1 that power is greater than α for all positive θ does not readily extend to prove that the power function increases monotonely with θ. Indeed, a general result is not possible since examples exist where power is *not* monotone. The power function in Figure 2 is for a sampling rule similar to our example but with ξ replaced by the maximum of ξ^2 and 0.5: after rising to 0.914 at $\theta = 0.8\,\delta$, power falls back to 0.884 at $\theta = 1.4\,\delta$ before starting to increase again.

It is possible that

$$S_1 + \gamma^{-1/2}S_2 \;>\; z_\alpha\sqrt{n} \tag{15}$$

and the variance spending test rejects H_0, while

$$S_1 + S_2 \;<\; z_\alpha\sqrt{n^*} \tag{16}$$

and a standard z-test calculated from all n^* observations per treatment would not reject H_0. Denne[11] notes an analogous problem in adaptive group sequential tests and we may choose to follow his suggestion for such a situation, rejecting H_0 only if *both* conditions (15) and (16) are satisfied. Although this lowers both Type I error rate and power, the effect is surprisingly small and the maximum reduction in the variance spending test's power at any point is less than 0.02.

Figure 1: Power of the variance spending (VS) test and the originally planned fixed sample size test with power 0.9 at $\theta = \delta$.

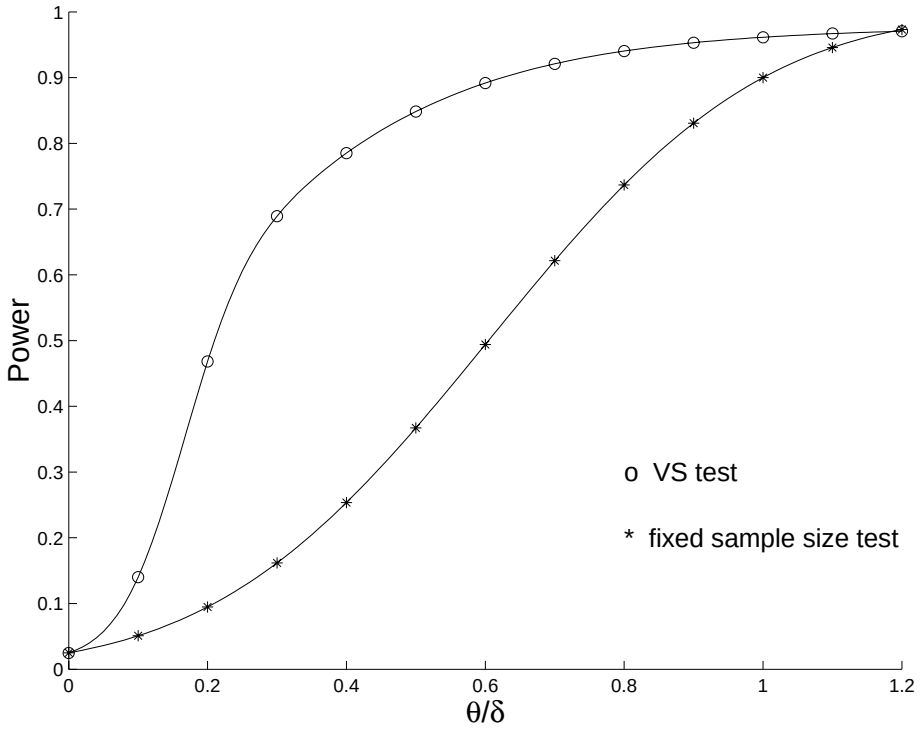
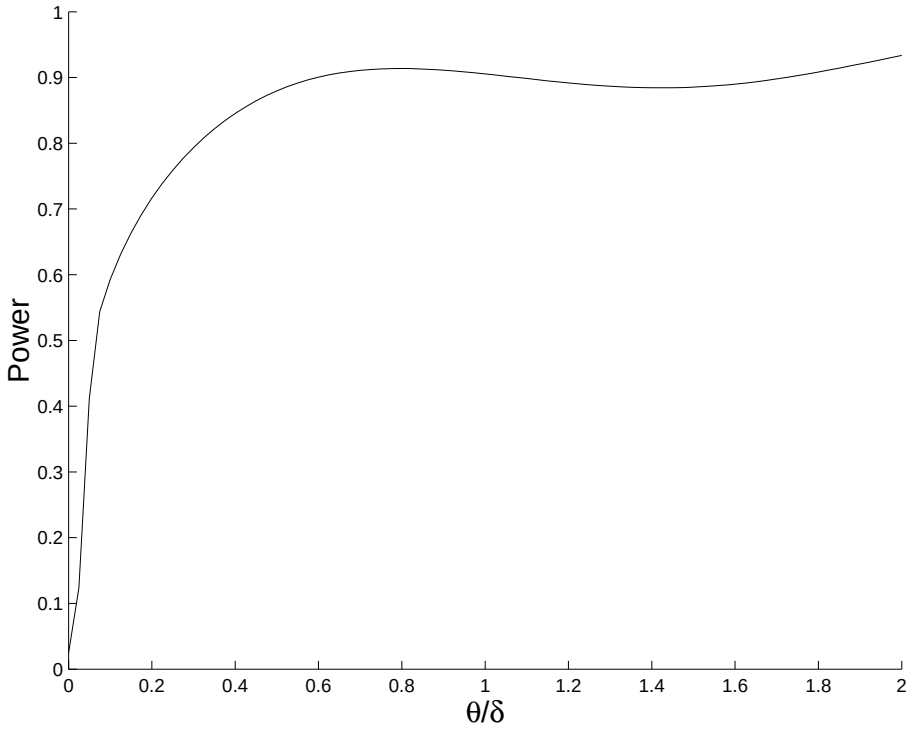


Figure 3 compares properties of the variance spending test with the fixed sample test that has power 0.9 at $\theta = 0.6\,\delta$. The left hand panel shows the difference in the shapes of the two power curves, the fixed sample test having the greater power for $\theta > 0.6\,\delta$ but the lower power, by some margin, at smaller θ values. The ASN per treatment arm of the variance spending test is plotted in the right hand panel of Figure 3, expressed in units of $n = 10.51/\delta^2$, the number of observations per arm originally planned. The steep rise in ASN as θ decreases from δ towards zero is in keeping with the goal of a sample size inversely proportional to θ^2 for θ between $0.25\,\delta$ and $2\,\delta$ that motivated this sampling rule. The variation in ASN is substantial with values around $12n$ for θ near zero compared to $3n$ or less for $\theta > \delta$. In contrast, the fixed sample test has constant sample size of $n/0.6^2 = 2.78n$. If it had been realized at the outset that greater power was desirable, this fixed sample test would have been an attractive candidate, offering broadly similar gains in power to the variance spending test for a generally lower sample size.

Figure 2: Non-monotone power function of an adaptively defined test.



4.3 Lack of efficiency of the variance spending test

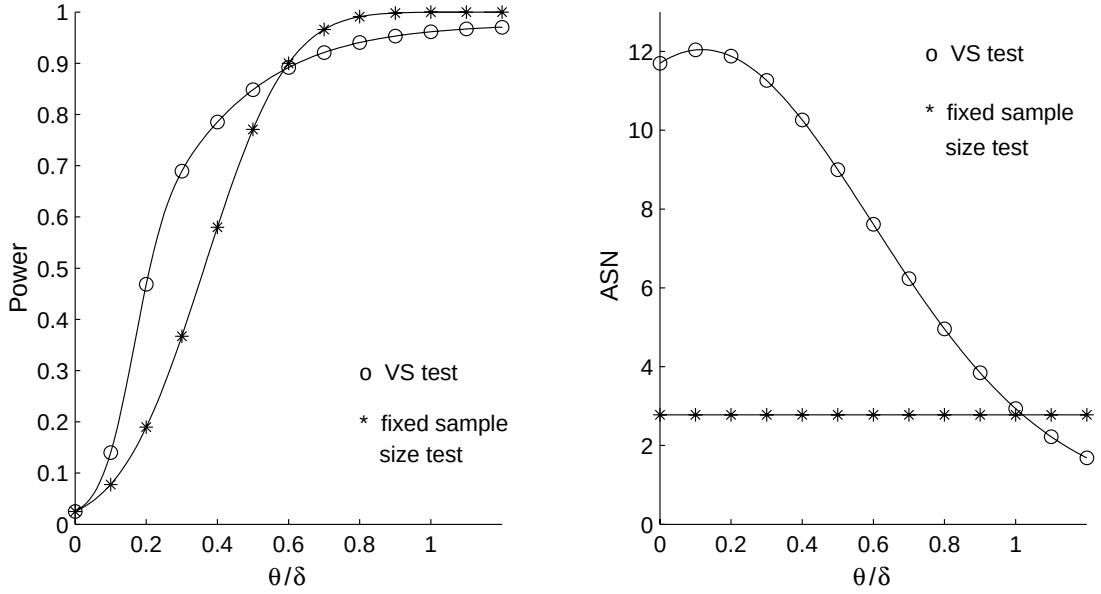
In Section 3.5 we presented a measure of possible inefficiency in the variance spending test due to its use of a non-sufficient statistic for θ. The high ASN seen in Figure 3 relative to that of a fixed sample test with broadly similar power curve is further evidence of such inefficiency. Figure 4 compares the variance spending test's ASN function with the fixed sample size needed to obtain the same power at each individual value of θ. Since this fixed sample size varies with θ, the values plotted on the line labeled FSS do not represent the ASN curve of any particular test, but this is still a reasonable point of comparison: many sequential and group sequential tests would have a *lower* ASN at each value of θ in such a comparison.

For general r, if the value of γ were independent of S_1, the expectation of

$$Z \;=\; \frac{S_1 + \gamma^{-1/2} S_2}{\sqrt{n}}$$

would be $\{r + \sqrt{\gamma}(1-r)\}\sqrt{n}\,\theta$, which is the same as the expectation of a standardized statistic based on $\{r + \sqrt{\gamma}(1-r)\}^2 n$ equally weighted observations per treatment arm. We therefore define the *effective*

Figure 3: Power and ASN of the variance spending (VS) test and a fixed sample size test with power 0.9 at $\theta = 0.6\,\delta$.



ASN scale is in multiples of n, the sample size originally chosen to give power 0.9 at $\theta = \delta$.

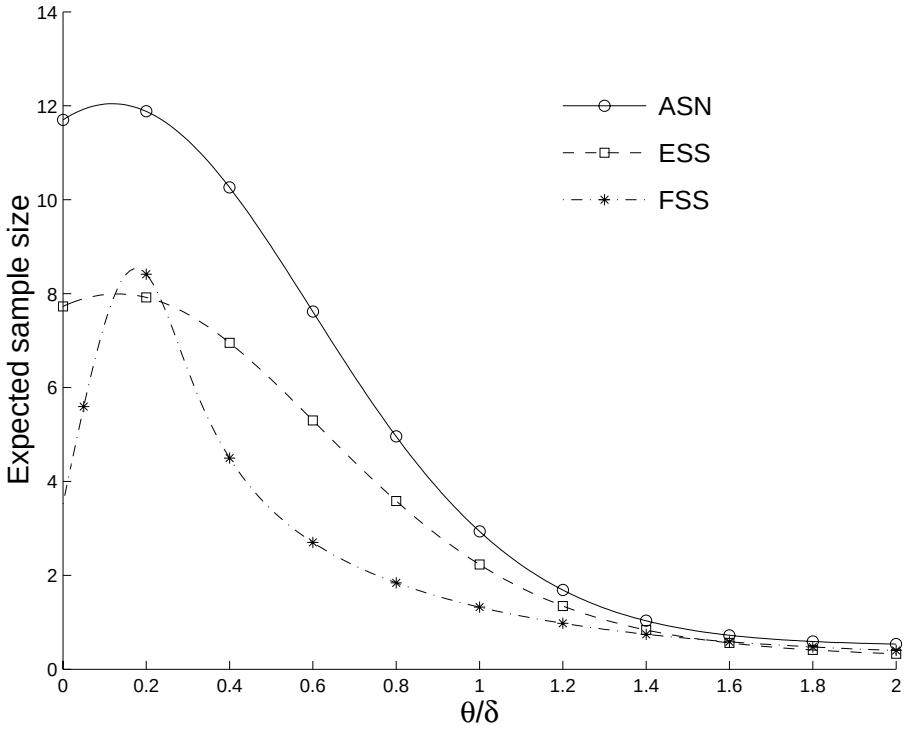
sample size in the variance spending test as

$$N_{\mathit{eff}} = \{r + \sqrt{\gamma}(1-r)\}^2 n.$$

A little algebra shows that N_{eff} is always less than or equal to the actual sample size $n^* = \{r + \gamma(1-r)\}n$ with equality only when $\gamma = 1$. The average effective sample size for our example test (with $r = 0.5$), labeled ESS in Figure 4, lies below the ASN but, for the most part, above the equivalent fixed sample size, FSS. Thus, at most θ values, power is still less than one might expect given the average effective sample size.

The remaining lack of power can be attributed to the variability in N_{eff}, along the lines of the discussion of variable sample size in Section 3.5. As an example, consider the case $\theta = 0.5\,\delta$. The density of $\widehat{\theta}_1$ when $\theta = 0.5\,\delta$ is shown in the left hand panel of Figure 5 and the resulting distribution of N_{eff} in the right hand panel. This distribution comprises a density plus two point probability masses arising from $\xi = 0.5$ and 4, for which $\gamma = 0$ and 49 and $N_{\mathit{eff}} = n/4$ and $16n$ respectively. The average effective sample size, $6.17n$, is noticeably less than the ASN of $9.00n$. A size $\alpha = 0.025$ fixed sample test with $6.17n$ observations per

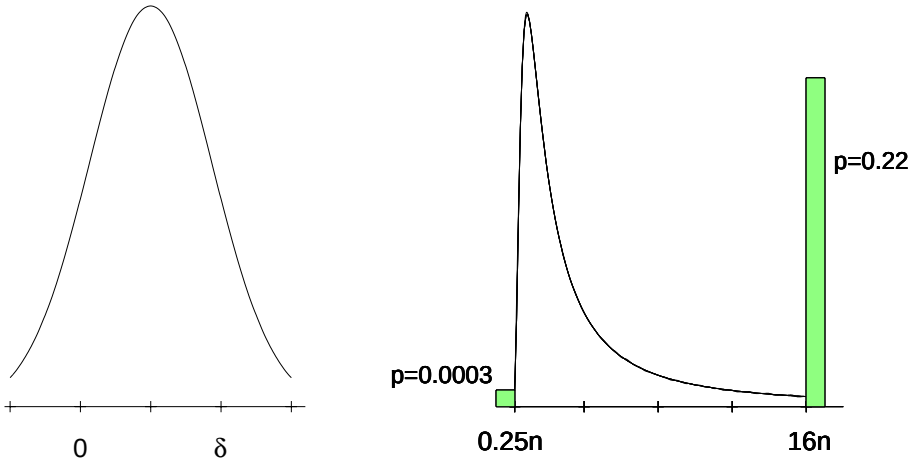
Figure 4: ASN and average effective sample size (ESS) of the variance spending test compared to the fixed sample size (FSS) needed to obtain the same power at each individual value of θ.



Sample size scale is in multiples of the original fixed sample size, n.

treatment arm has power 0.981 at $\theta = 0.5\,\delta$. If, however, a test is carried out with a random sample size taken from the distribution of N_{eff} using a conditional significance level α given the observed sample size, its overall power is only 0.703 at $\theta = 0.5\,\delta$. The variance spending test's power there of 0.848 lies between these two values, indicating that it suffers from the effects of the variable sample size but these are partly ameliorated by the way in which γ depends on S_1: low values of γ are chosen when S_1 is high and good conditional power can be achieved with a small number of additional observations, while high values of γ occur when S_1 is low and a higher stage two sample size provides a substantial rise in conditional power. We note, however, that this beneficial effect is of limited value since, as the line FSS in Figure 4 shows, a fixed sample size of just $3.40n$ per arm is all that is needed to achieve the variance spending test's power of 0.848 in a simple, fixed sample test.

Figure 5: Density of $\widehat{\theta}_1$ when $\theta = 0.5\delta$ (left panel) and consequent distribution of N_{eff} (right panel).



Density of $\widehat{\theta}_1$ is $N(0.5\delta, 2\delta^2/10.51)$

N_{eff} has a density on $(0.25n, 16n)$ plus point masses at $N_{\mathit{eff}} = 0.25n$ and $16n$.

One can ask whether better results might have been obtained if the first and second stage data had been combined through some other test statistic. As explained in Section 3.2, use of such a test is only allowable if designated in the initial protocol, thus, this is not a legitimate option in the scenario of an unplanned interim analysis in what was intended to be a simple fixed sample size trial. Bauer & Köhne[4] use R.A. Fisher's method for combining P-values for a one-sided test of H_0: $\theta = 0$ against $\theta > 0$. The first and second stage P-values are

$$p_1 = 1 - \Phi(S_1/\sqrt{\{rn\}}) \quad \text{and} \quad p_2 = 1 - \Phi(S_2/\sqrt{\{\gamma(1-r)n\}}),$$

respectively. Under H_0, $-\log(p_1\, p_2)$ has 0.5 times a χ^2 distribution on 4 degrees of freedom, so a hypothesis test with Type I error rate α can be obtained by rejecting H_0 if

$$p_1\, p_2 \;<\; \exp\{-0.5\, \chi^2_{4,\alpha}\},$$

where $\chi^2_{\nu,p}$ denotes the upper p tail point of a χ^2_ν distribution. For $\alpha = 0.025$, the critical value for $p_1\, p_2$ is 0.0038. Combining this rule with the sampling rule of our example, produces a fairly similar power curve to that of the variance spending test: the curves cross twice between $\theta = 0$ and $\theta = \delta$ and are within 0.03 of each other everywhere. However, the power of Fisher's combination test does approach one more

rapidly and the difference between, for example, power 0.989 for Fisher's combination test and 0.961 for the variance spending test at $\theta = \delta$ may be regarded as significant. More might have been expected of Fisher's combination test in view of the very good power properties of this test in a simpler application reported by Bauer & Köhne[4] (Table 3); however, it should be noted that the design in our example, in particular the rule for stopping for futility at the first analysis, is not tailored to Fisher's combination test.

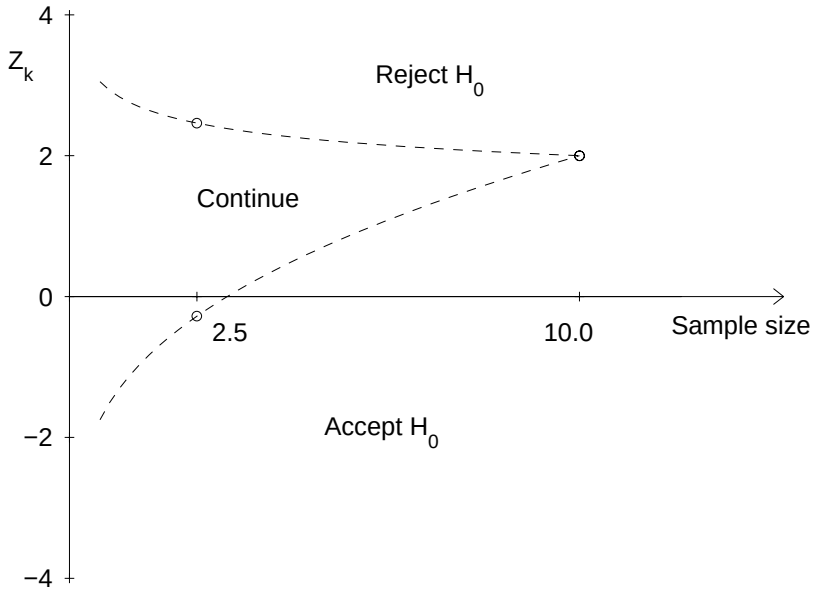
4.4 Competing group sequential tests

In Section 4.2 we compared the variance spending test with a fixed sample test achieving similar overall power. This fixed sample design could have been chosen if the experimenters had anticipated the need for greater power. In this case, there are other options too: group sequential tests can satisfy error probability requirements with lower average sample size than fixed sample tests. Error spending group sequential tests are a currently popular choice and have the ability to deal with variation in observed group sizes about their intended values. We shall present results for one-sided error spending tests in the "ρ-family" described by Jennison & Turnbull,[17] Section 7.3; for simplicity, we present results when group sizes are actually equal to their planned values. The tests described below are chosen to dominate the variance spending test in terms of both power and ASN for θ values in the region of primary interest between zero and δ.

The two-stage, one-sided group sequential test shown in Figure 6 has Type I error rate 0.025 and power 0.9 at $\theta = 0.33\,\delta$. The form of stopping boundary is quite standard, namely a ρ-family error spending test with $\rho = 1$. An unusual feature of the design is the timing of the first analysis after just $2.5n$ observations per treatment, one fourth of the maximum sample size: this allows sufficiently early stopping to make good reductions in ASN at parameter values near $\theta = \delta$, where power is very close to one. Setting power 0.9 at $\theta = 0.33\,\delta$ ensures that the group sequential test's power curve lies completely above that of the variance spending test. The left hand panel of Figure 7 shows that, in addition, the group sequential test provides *much* greater power for values of θ around $0.3\,\delta$ and above. At the same time, the ASN curves in the right hand panel demonstrate that this is achieved with considerably lower average sample size. Furthermore, the group sequential test's maximum sample size of $10.0n$ per treatment arm is far below that of $25n$ for the variance spending test.

A two-stage group sequential test is comparable with the variance spending test in that both have a total of two analyses. However, the variance spending test has the freedom to vary the second stage group size in the light of first stage data while that of the group sequential test is pre-determined. Careful timing of a

Figure 6: A two-stage, one-sided group sequential test of H_0: $\theta = 0$ with Type I error rate 0.025 and power 0.9 at $\theta = 0.33\,\delta$.

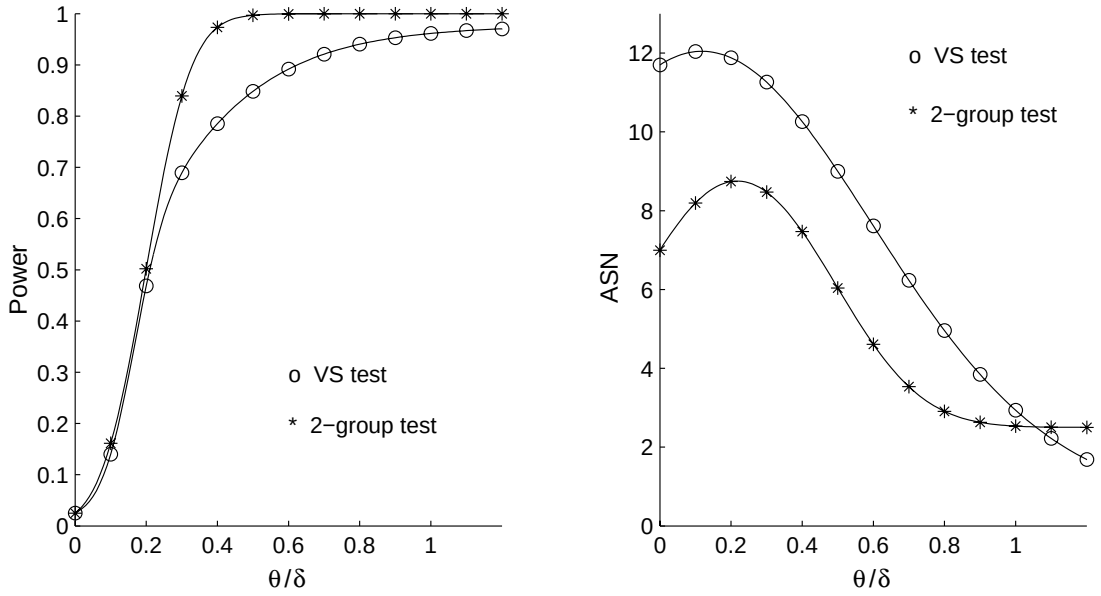


Sample size scale is in multiples of the original fixed sample size, n.

group sequential test's early analyses helps attain low average sample sizes at the higher values of θ, where power is close to one. This is evident in the above two group design where the first analysis is set at one quarter of the total sample size. The five and ten-stage, one-sided group sequential boundaries shown in Figures 8 and 9 are also for ρ-family error spending tests with $\rho = 1$. Both tests have Type I error rate 0.025 and power 0.9 at $\theta = 0.33\,\delta$ and their power curves are indistinguishable from that of the two-stage test in Figure 6. The five-stage test has its first analysis at one tenth of the total sample size, with equal group sizes thereafter, while the ten-stage test has ten equally sized groups. The ASN curves in Figure 10 show the usual improvements in ASN arising from more frequent analyses and particular improvement at higher values of θ, helped by the additional, well placed, early analyses. Again, the maximum sample sizes per treatment arm of $11.4n$ for the five group test and $11.9n$ for the ten group test are well below the variance spending test's $25n$.

These comparisons with standard group sequential designs illustrate the possible cost of the flexibility available in the variance spending approach. The increased power and reduced sample size of the group

Figure 7: Power and ASN curves of the variance spending (VS) test and two-stage group sequential test.



ASN scale is in multiples of the original fixed sample size, n.

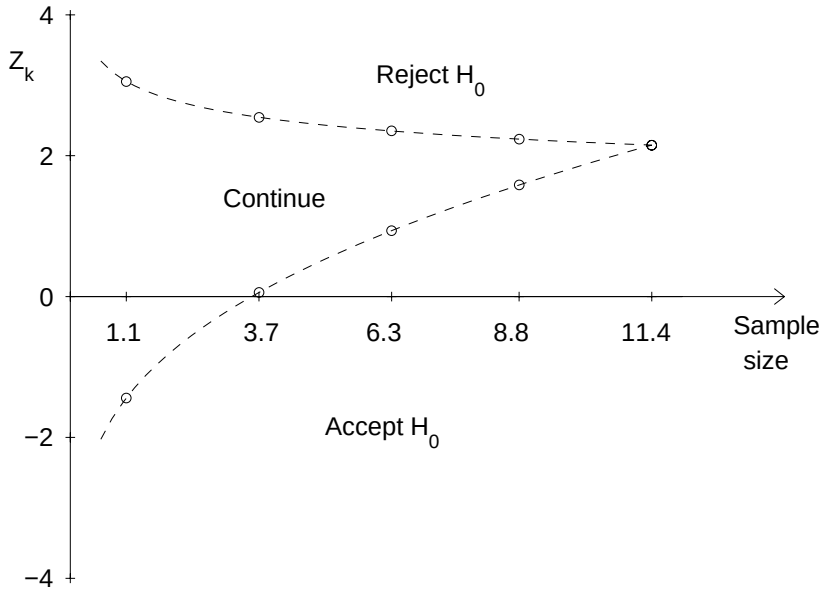
sequential tests argue eloquently for more careful consideration of the appropriate power requirement, and choice of a suitable design, well before a trial gets under way.

5 Discussion

There is no dispute that a variance spending test can rescue a trial from a poor initial design. The flexibility of these tests can also be used to adapt a trial to a change in treatment definition (such as a new dosage or selection of one dose from the range of doses used initially), or to the substitution of an alternate endpoint; see, for example, Bauer & Köhne,[4] Bauer & Röhmel,[25] Fisher[2] and Lehmacher & Wassmer.[8] In another form of adaptation, Wang, et al[14] use Cui et al's[7] method to create a group sequential test which can switch adaptively between hypothesis tests of superiority and non-inferiority.

It may not be so obvious that this flexibility can come at a substantial price. Our evaluations have been in the context of changing a trial's sample size in mid-course in order to attain power at a smaller effect size than originally anticipated. The message from our example is clear: a variance spending test can require

Figure 8: A five-stage, one-sided group sequential test of H_0: $\theta = 0$ with Type I error rate 0.025 and power 0.9 at $\theta = 0.33\,\delta$.

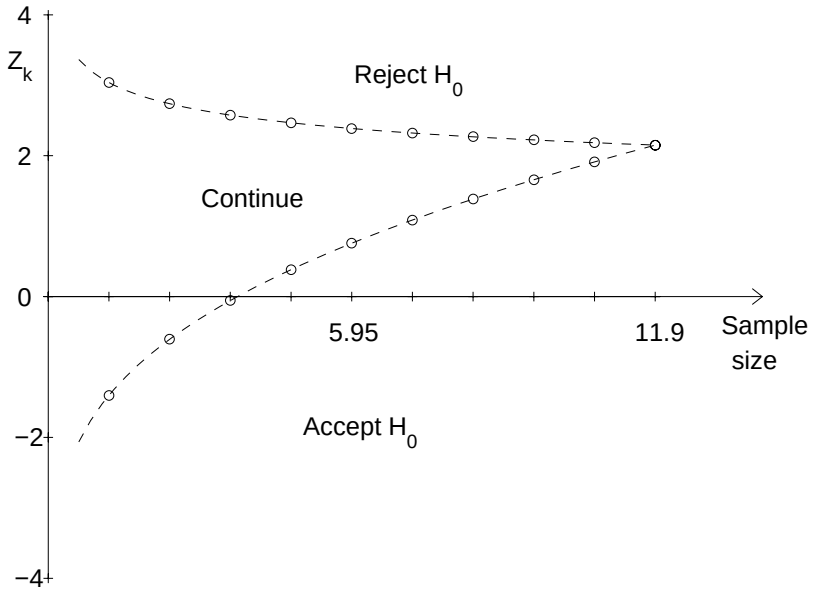


Sample size scale is in multiples of the original fixed sample size, n.

many more observations than a group sequential test with superior power. Thus, consideration should be given at the planning stage to what is desirable for the full range of possible effect sizes, including those deemed unlikely at that point. If observing a particular value of $\widehat{\theta}$ at an interim analysis will be enough to persuade the investigators that high power is appropriate at that value of θ, then it makes sense to design the study with such power from the outset.

We have concentrated in this paper on an in depth analysis of one example. In addition to this, we have studied similar examples with different values for r, the fraction of data available when the design is adapted, and for the parameters M and h which govern truncation of the modified sample size through the definition (12) of $\tilde{\xi}(\widehat{\theta}_1)$. We have implemented sample size rules based on conditional power, as described at the start of Section 3.4. In Section 4.3, we reported on methods in which data from the two stages are combined through R.A. Fisher's χ^2 method rather than the variance spending rule. We have also recently investigated methods proposed by Fisher,[2] Shen & Fisher[3] and Cui et al[7] which allow re-design within an initially planned group sequential test. Our findings in all these cases follow the same pattern as for the

Figure 9: A ten-stage, one-sided group sequential test of H_0: $\theta = 0$ with Type I error rate 0.025 and power 0.9 at $\theta = 0.33\,\delta$.

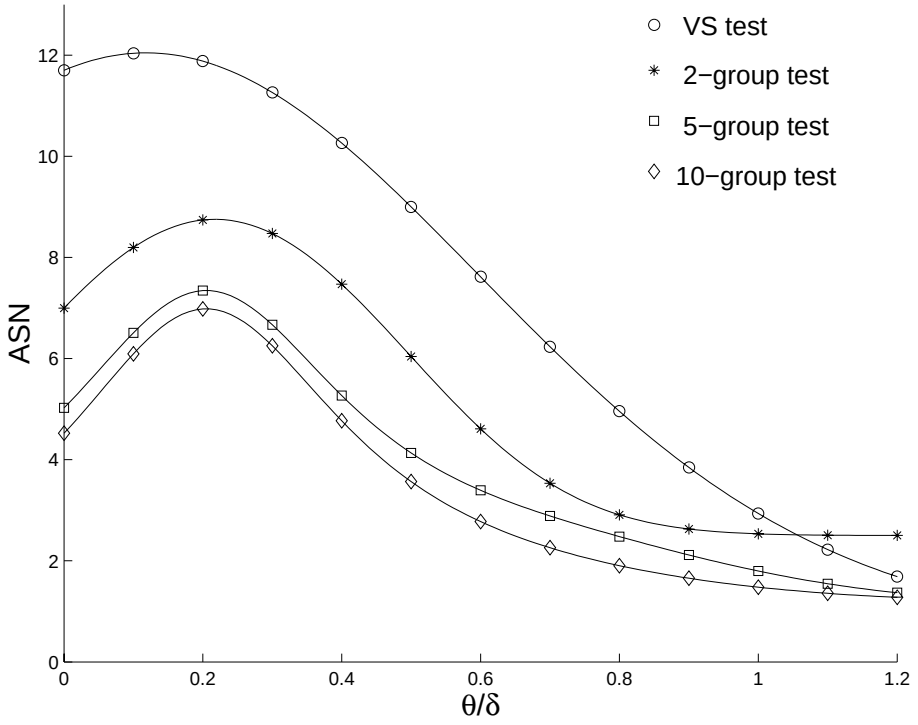


Sample size scale is in multiples of the original fixed sample size, n.

example presented in Section 4: on calculating the overall power curve of the adaptive procedure, with specified rules for the re-calculation of sampling size and the terminal decision, it is evident that non-adaptive group sequential tests could have achieved better power with smaller expected sample size across the range of θ values of interest. We do not claim that all adaptive designs must be seriously inefficient, but the examples we have investigated indicate the need to evaluate procedures and compare other options before possibly applying them.

The increase in sample size by a factor of 25 allowed in our example is clearly very high and many adaptive modifications will be on a smaller scale. It is noteworthy, however, that this increase is not particularly effective since Figure 3 shows that a fixed sample test with just 2.8 times the initial sample size broadly matches the attained power curve. If one seeks to attain power $1 - \beta$ for θ as low as $\delta/2$ (rather than $\delta/4$), the value $M = 2$ should be used in (12) and this gives the more plausible maximum sample size of $5n$. Calculations show the variance spending procedure using this sampling rule attains power 0.9 at $\theta = 0.7\,\delta$. The power curve is dominated at all points by fixed sample and group sequential tests with

Figure 10: ASN curves of the variance spending (VS) test and group sequential tests with 2, 5 and 10 stages.



ASN scale is in multiples of the original fixed sample size, n.

power 0.9 at $0.57\,\delta$ and, as for our earlier example, group sequential tests offer lower ASN than the variance spending test over a wide range of θ values. The two group test has a maximum sample size of $3.35\,n$ and lower ASN for all θ values between zero and $1.2\,\delta$; five and ten group tests offer greater reductions in ASN.

Remarks by some authors suggest a desire to set a specific power, $1-\beta$, at whatever is the true value of the effect size parameter; for example, Shen and Fisher[3] (Section 3) refer to the value δ at which power is set as being an underestimate, a proper estimate, or an overestimate of the underlying treatment difference θ. This is a curious motivation for adaptive designs with the apparent objective of a power curve which rises sharply near $\theta = 0$ and then remains perfectly flat. What is surprising is that adaptive designs with the sampling rules we have presented do actually come close to having such power curves! However, maintaining a significant risk of a negative outcome when the effect size is high seems quite perverse. This whole philosophy seems to be generated from a misconception about the role of power calculations: power should be guaranteed at values of θ that would be of clinical or commercial interest, bearing in mind the

sampling cost needed to detect a particularly small effect size. Then, the design will be suitable whatever the true value of θ.

As long as the experimenters' objectives can be properly established at the outset, there are good reasons to expect standard group sequential designs to dominate variance spending tests. Knowing the correct goal helps design the trial efficiently, use of a sufficient statistic is in keeping with fundamental principles, and one can choose from tests optimized to a selection of criteria (see Barber & Jennison[26]). Variance spending tests have the special feature that future group sizes can be adapted to previously observed responses. The extension of group sequential tests to "sequentially planned" designs proposed by Schmitz[27] provides this property, which may be of value when only a small number of groups are permitted. However, we should not let consideration of these more complex designs obscure the excellent performance, seen in our example, of standard group sequential tests with pre-specified group sizes.

Acknowledgement

This research was supported in part by NIH grant R01 CA66218.

References

1. Cornfield, J. A Bayesian test of some classical hypotheses — with applications to sequential clinical trials. *J. American Statistical Association*, **61**, 577–594 (1966).

2. Fisher, L.D. Self-designing clinical trials. *Statistics in Medicine*, **17**, 1551–1562 (1998).

3. Shen, Y. and Fisher, L. Statistical inference for self-designing designing clinical trials with a one-sided hypothesis. *Biometrics*, **55**, 190–197 (1999).

4. Bauer, P. and Köhne, K. Evaluation of experiments with adaptive interim analyses. *Biometrics*, **50**, 1029–1041 (1994). Correction *Biometrics*, **52**, 380 (1996).

5. Proschan, M.A. and Hunsberger, S.A. Designed extension of studies based on conditional power. *Biometrics*, **51**, 1315–1324 (1995).

6. Lan, K.K.G. and Trost, D.C. Estimation of parameters and sample size reestimation. *Proceedings of Biopharmaceutical Section*, American Statistical Association, Alexandria, Virginia, pp. 48–51 (1997).

7. Cui, L., Hung, H.M.J. and Wang, S-J. Modification of sample size in group sequential clinical trials. *Biometrics*, **55**, 853–857 (1999).

8. Lehmacher, W. and Wassmer, G. Adaptive sample size calculation in group sequential trials. *Biometrics*, **55**, 1286–1290 (1999).

9. Chi, G.Y.H. and Liu, Q. The attractiveness of the concept of a prospectively designed two-stage clinical trial. *J. Biopharmaceutical Statistics*, **9**, 537–547 (1999).

10. Denne, J.S. Estimation following extension of a study on the basis of conditional power. *J. Biopharmaceutical Statistics*, **10**, 131-144 (2000).

11. Denne, J.S. Sample size recalculation using conditional power. *Statistics in Medicine*, **20**, 2645–2660 (2001).

12. Müller, H-H. and Schäfer, H. Changing a design during the course of an experiment. Unpublished manuscript (2000).

13. Müller, H-H. and Schäfer, H. Adaptive group sequential designs for clinical trials: Combining the advantages of adaptive and of classical group sequential procedures. *Biometrics*, **57**, 886–891 (2001).

14. Wang, S-J, Hung, H.M.J., Tsong, Y. and Cui, L. Group sequential test strategies for superiority and non-inferiority hypotheses in active controlled clinical trials. *Statistics in Medicine*, **20**, 1903–1912 (2001).

15. Wassmer, G. Basic concepts of group sequential and adaptive group sequential test procedures. *Statistical Papers*, **41**, 253–279 (2000).

16. Whitehead, J., Whitehead, A., Todd, S., Bolland, K. and Sooriyarachchi, M.R. Mid-trial design reviews for sequential clinical trials. *Statistics in Medicine*, **20**, 165–176 (2001).

17. Jennison, C. and Turnbull, B.W. *Group Sequential Methods with Applications to Clinical Trials*, Chapman & Hall/CRC, Boca Raton (2000).

18. Proschan, M.A., Follmann, D.A. and Waclawiw, M.A. Effects of assumption violations on Type I error rate in group sequential monitoring. *Biometrics*, **49**, 1131–1143 (1992).

19. Shun, Z., Yuan, W., Brady, W.E. and Hsu, H. Type I error in sample size reestimations based on observed treatment difference (with commentary). *Statistics in Medicine*, **20**, 497–520 (2001).

20. Liu, Q., Proschan, M.A. and Pledger, G.W. A unified theory of two-stage adaptive designs. *J. American Statistical Association*. In press.

21. Wassmer, G. A comparison of two methods for adaptive interim analyses in clinical trials. *Biometrics*, **54**, 696–705 (1998).

22. Posch, M. and Bauer, P. Adaptive two stage designs and the conditional error function. *Biometrical Journal*, **41**, 689–696 (1999).

23. Brannath, W., Posch, M. and Bauer, P. Recursive combination tests. *J. American Statistical Association*, **97**, 236–244 (2002).

24. Liu, Q. and Chi, G.Y.H. On sample size and inference for two-stage adaptive designs. *Biometrics*, **57**, 172–177 (2001).

25. Bauer, P. and Röhmel, J. An adaptive method for establishing a dose-response relationship. *Statistics in Medicine*, **14**, 1595–1607 (1995).

26. Barber, S. and Jennison, C. Optimal asymmetric one-sided group sequential tests. *Biometrika*, **89**, 49–60 (2002).

27. Schmitz, N. *Optimal Sequentially Planned Decision Procedures.* Lecture Notes in Statistics, 79, Springer-Verlag: New York (1993).