Adaptive and non-adaptive group sequential tests

Christopher Jennison
Department of Mathematical Sciences,
University of Bath, Bath, UK

email: cj@maths.bath.ac.uk

and Bruce W. Turnbull
Department of Statistical Science,
Cornell University, Ithaca, New York, USA

17 August 2004

SUMMARY

Methods have been proposed to re-design a clinical trial at an interim stage in order to increase power. This may be in response to external factors which indicate power should be sought at a smaller effect size, or it could be a reaction to data observed in the study itself. In order to preserve the type I error rate, methods for unplanned design change have to be defined in terms of non-sufficient statistics and this calls into question their efficiency and the credibility of conclusions reached. We evaluate methods for adaptive re-design, extending the theoretical arguments for use of sufficient statistics of Tsiatis & Mehta (2003) and assessing the possible benefits of pre-planned adaptive designs by numerical computation of optimal tests; these optimal adaptive designs are concrete examples of optimal sequentially planned sequential tests proposed by Schmitz (1993). We conclude that the flexibility of unplanned adaptive designs comes at a price and we recommend the appropriate power for a study should be determined as thoroughly as possible at the outset. Then, standard error spending tests, possibly with unevenly spaced analyses, provide efficient designs but it is still possible to fall back on flexible methods for re-design should study objectives change unexpectedly once the trial is under way.

Key words: Adaptive re-design; Admissibility; Clinical trials; Conditional power; Efficiency; Group sequential tests; Sufficiency.

1 Introduction

There has been much recent interest in adaptive methods for modifying the power, or conditional power, of a clinical trial at an interim stage. Such adaptation may be in response to external developments or to information arising in the study itself. We consider the situation where there is a change in the alternative at which a specified power is to be attained. This should not be confused with the problem of "re-estimating" the sample size needed to meet a fixed power requirement as more is learnt about a nuisance parameter that controls the necessary sample size; see, for example, Wittes & Brittain (1990) or, for updating sample size in a group sequential test, Denne & Jennison (2000). Adaptive strategies have also been proposed to deal with changes in treatment definition, changes in the primary response or the way it is measured, switching between tests for superiority and non-inferiority, or adaptive randomisation rules for reducing the number of subjects on an inferior treatment. Many of these adaptations can be accommodated in non-adaptive group sequential tests and are essentially orthogonal to the issues we consider here.

Suppose θ represents the improvement in efficacy offered by a new treatment and a study has been designed to test H_0: $\theta \leq 0$ against the alternative $\theta > 0$ with type I error probability α and power $1 - \beta$ at $\theta = \delta$. Motivation for re-design may be from external factors, for example, the withdrawal of a rival treatment may mean that smaller effect sizes for the new treatment are now of interest and power $1 - \beta$ should be sought at an alternative $\theta = \delta'$ for some $0 < \delta' < \delta$. A similar conclusion might be reached in the light of information internal to the study but on a secondary endpoint, for example, good safety results combined with a positive efficacy effect at a level below δ could justify use of the new treatment.

There may, instead, be completely internal reasons for re-design, arising from interim data on the primary endpoint. It may be deemed appropriate to increase the remaining sample size of a study if continuing as planned would give low conditional power under $\theta = \delta$. Alternatively, when an interim estimate $\widehat{\theta}$ below δ is reported, investigators may realise that, although $\widehat{\theta}$ is lower than the effect size they had expected or hoped for, it still represents a worthwhile improvement and they would like to extend the study to ensure high power can be achieved under such an effect size. Monitoring a study by repeated confidence intervals, as described by Jennison & Turnbull (1989), gives flexibility to modify criteria for early stopping but this approach still assumes adherence to a specified sampling plan: attaining power $1-\beta$ at an alternative closer to the null hypothesis necessitates an increase in sample size.

Special methods are needed to preserve the type I error probability at level α if sample size is changed in order to modify power on the basis of observed data. Bauer & Köhne (1994) propose two-stage designs in which P-values calculated separately from the two stages are combined through R. A. Fisher's (1932) method; this allows great flexibility in adapting the second stage to interim data but, to be valid, the method must be adopted at the outset. More recently, Cui et al. (1999), L. D. Fisher (1998), Shen & Fisher (1999) and Müller & Schäfer (2001), among others, have proposed a variety of methods that preserve the type I error rate despite completely unplanned design changes. Although differing in appearance and derivation, these methods are closely related in that each preserves the conditional type I error probability whenever the design is modified; Jennison & Turnbull (2003) prove this must be the case for any unplanned re-design that preserves the overall type I error rate.

The publication of well over a hundred papers on adaptive designs in recent years indicates great enthusiasm for these methods, with potential uses well beyond the rescue of under-powered studies described by Cui et al. (1999). In their illustrative examples, Lehmacher & Wassmer (1999) and Brannath, Posch & Bauer (2002) note the freedom given to investigators to re-design the remainder of a study at an interim stage. Shen & Fisher (1999) promote "variance spending" tests as a means to gain the benefits of low sample size for given power achieved by group sequential tests. Thach & Fisher (2002) search for optimal designs within a class of two-stage variance spending tests. In Shen & Fisher's (1999) examples, a power curve is not decided on at the outset, instead, sample sizes are modified to aim for power $1-\beta$ under the actual effect size, using an estimate from interim data. We shall return in our discussion to the contentious issue of whether it is reasonable to postpone full consideration of the power requirement until interim data become available.

Several authors explain adaptive re-design in terms of a weighting factor for later observations: thus, the responses of different subjects are weighted unequally and decisions are not functions of the sufficient statistic for θ. Failure to observe the principal of sufficiency (Cox & Hinkley, 1974, Sec. 2.3) raises questions both about the statistical efficiency of the experimental designs and the credibility results will have when reported to a wider audience. In an analysis of selected examples, Jennison & Turnbull (2003) show that adaptive sampling rules can be highly inefficient in comparison with standard group sequential tests. Tsiatis & Mehta (2003) give a formal proof that any adaptive test using a non-sufficient statistic can be out-performed by a sequential test using the sufficient statistic; however, the sequential test they construct to do this is allowed a greater number of analyses than the adaptive test. Proponents of adaptive designs have responded to these criticisms: in a comparison of certain classes of adaptive and non-adaptive designs, Posch, Bauer & Brannath (2003) found optimal adaptive designs to have a small advantage over their optimal non-adaptive counterparts. These adaptive designs are examples of the "sequentially planned sequential designs" proposed by Schmitz (1993) and are implemented according to a precisely defined set of rules, a quite different prospect from the flexible schemes discussed above.

Our objectives in this paper are to illustrate and critically appraise methods of adaptive re-design for power criteria, in particular, to answer the questions:

> Does the use of non-sufficient statistics in adaptive designs automatically imply a loss of efficiency?
>
> How great an improvement over non-adaptive tests can the most efficient adaptive sequential tests offer and is this large enough to justify their use in practice?

Examples in Section 2 illustrate how adaptive re-design can be used to meet new objectives arising from external or internal information. We measure the cost of delay in learning about a study's real objective by comparing performance of the adaptive scheme with a non-adaptive test designed knowing the ultimate objective at the outset. In Section 3 we present a complete class theorem that characterises optimal adaptive and non-adaptive tests when the experimental design is pre-planned. To be admissible, a sequential test must be the solution of a Bayes decision problem, and one implication of this is that tests which do not adhere to the sufficiency principle are inadmissible. The theoretical results of Section 3 answer our first question in the affirmative, re-inforcing the evidence of specific examples in Section 2.

In Section 4 we quantify the benefits adaptivity can yield in pre-planned designs. Our results show that small gains are indeed possible but these are unlikely to be regarded as sufficiently great to justify the extra complexity of an adaptive design. Nor do the positive benefits of optimal adaptive tests provide a useful margin to offset the inefficiency arising from use of non-sufficient statistics or sub-optimal sampling rules in unplanned adaptive tests.

Our conclusion is that the strength of adaptive re-design lies in coping with the unexpected, in particular responding to external information that could not have been anticipated at the start of a study. The efficiency cost when adaptive methods are used to rescue an under-powered study is inescapable and we would recommend investigators avoid such problems by thinking through the power requirement carefully at the planning stage.

2 Sample size adaptation to alter power

2.1 Adaptation preserving the type I error rate

Cui et al. (1999) cite instances in their experience at the U. S. Food and Drug Administration of researchers proposing an increase in sample size during the course of a group sequential trial based on the observed sample path. In one example, a Phase III study of a drug to prevent myocardial infarction in patients undergoing coronary artery bypass graft surgery was designed to have power 0.95 to detect a 50% reduction in incidence. At an interim point, the incidence rate in the placebo group was in line with expectations but the observed rate for patients receiving the drug was only 25% lower. The investigators recognised that a 25% reduction in incidence was still clinically significant but, as designed, the study had little power to detect such an effect: consequently a proposal was submitted to expand the study's sample size. However, no valid testing procedure was available to account for such an outcome-dependent adjustment of sample size.

Such events motivated Cui et al. (1999) to propose a method of adapting sample size during the course of a group sequential study which preserves type I error. We describe their proposal in the context of a general group sequential test of a treatment effect θ. Suppose efficient score statistics S_k for θ are available at analyses $k = 1, \ldots, K$ with

$$\begin{aligned} &S_1 \sim N(\,\theta\mathcal{I}_1,\, \mathcal{I}_1), \\ &S_k - S_{k-1} \sim N(\,\theta(\mathcal{I}_k - \mathcal{I}_{k-1}),\, \mathcal{I}_k - \mathcal{I}_{k-1}), \quad k = 2, \ldots, K, \end{aligned} \tag{1}$$

and increments S_1, $S_2 - S_1, \ldots, S_K - S_{K-1}$ are independent. This joint distribution for a sequence of score statistics arises very generally, holding exactly in normal linear models and for large samples in other cases; see, for example, Jennison & Turnbull (1997). A one-sided group sequential test of the null hypothesis H_0: $\theta \leq 0$ against $\theta > 0$ takes the form

$$\begin{array}{ll} \text{after group } k = 1, \ldots, K-1 & \\ \qquad \text{if } S_k \geq b_k & \text{stop, reject } H_0 \\ \qquad \text{if } S_k \leq a_k & \text{stop, accept } H_0 \\ \qquad \text{otherwise} & \text{continue to group } k+1, \\ \text{after group } K & \\ \qquad \text{if } S_K \geq b_K & \text{stop, reject } H_0 \\ \qquad \text{if } S_K < a_K & \text{stop, accept } H_0, \end{array} \tag{2}$$

where $a_K = b_K$ to ensure termination at analysis K. Typically, tests are designed with analyses at equally spaced information levels $\mathcal{I}_1, \ldots, \mathcal{I}_K$. Then, for given K, the maximum information $\mathcal{I}_K$ and boundary values (a_k, b_k), $k = 1, \ldots, K$, can be chosen to attain type I error probability α under $\theta = 0$ and power $1 - \beta$ at a specified alternative $\theta = \delta$.

Suppose a test of the above form is under way and, based on data observed at analysis j, it is desired to increase the size of the remaining groups of observations. Let S'_k, $k > j$, denote the new score statistics and, for notational convenience, define $S'_j = S_j$. Assume each information increment is increased by a factor γ so that for each $k = j+1, \ldots, K$,

$$S'_k - S'_{k-1} \sim N(\,\theta\,\gamma\,(\mathcal{I}_k - \mathcal{I}_{k-1}),\, \gamma\,(\mathcal{I}_k - \mathcal{I}_{k-1}))$$

independently of other increments. For $k > j$, define

$$S_k = S_j + \sum_{i=j+1}^{k} \gamma^{-1/2}(S'_i - S'_{i-1}). \tag{3}$$

Then, under $\theta = 0$, increments remain independent, $N(0, \mathcal{I}_k - \mathcal{I}_{k-1})$ and applying the boundary (2) to the newly defined S_k preserves the type I error probability α exactly. The means of the new increments are multiplied by $\gamma^{1/2}$, so if $\gamma > 1$ this increases the test's power for $\theta > 0$. Cui et al. (1999) suggest a single re-design point will usually suffice but the method easily extends to more.

A key feature of this proposal is that it gives investigators freedom to decide how to modify a study at an interim point. However, in order to assess the method, it is necessary to consider specific strategies for adaptive re-design.

2.2 Example 1: Re-design in response to external information

We consider the example of a group sequential test with 5 analyses testing H_0: $\theta \leq 0$ against $\theta > 0$ with type I error probability $\alpha = 0.025$ and power $1 - \beta = 0.9$ at $\theta = \delta$. A fixed sample size test for this problem requires information for θ

$$\mathcal{I}_f = \{z_\alpha + z_\beta\}^2/\delta^2, \tag{4}$$

where z_p denotes the $1 - p$ quantile of the standard normal distribution. We suppose the study is designed as a one-sided test from the ρ-family of error spending tests described by Jennison & Turnbull (2000, Sec. 7.3) With the choice $\rho = 3$, the boundary values $a_1, \ldots, a_5$ and $b_1, \ldots, b_5$ are chosen to satisfy

$$P_\theta\{S_1 > b_1 \text{ or} \ldots \text{or } S_1 \in (a_1, b_1), \ldots, S_{k-1} \in (a_{k-1}, b_{k-1}), S_k > b_k\} = (\mathcal{I}_k/\mathcal{I}_5)^3\, \alpha$$

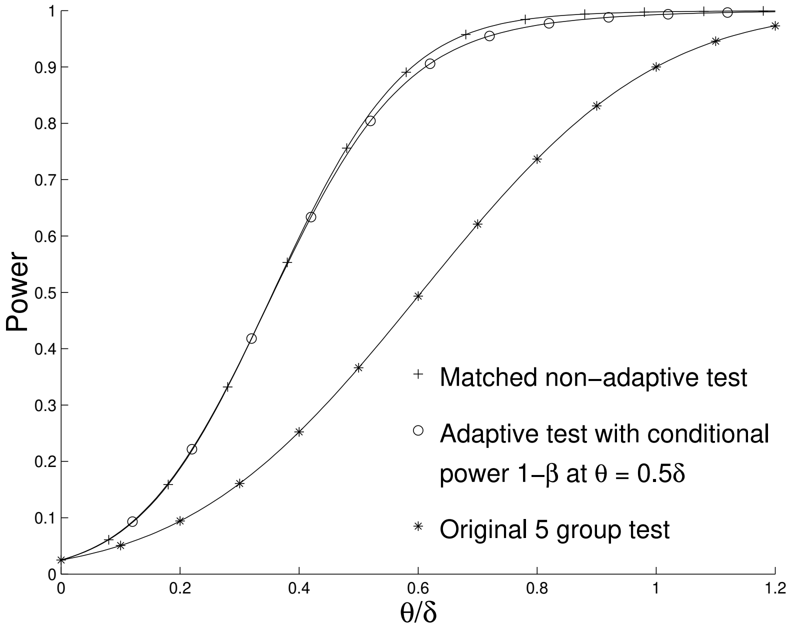
and

$$P_\theta\{S_1 < a_1 \text{ or} \ldots \text{or } S_1 \in (a_1, b_1), \ldots, S_{k-1} \in (a_{k-1}, b_{k-1}), S_k < a_k\} = (\mathcal{I}_k/\mathcal{I}_5)^3\, \beta$$

for $k = 1, \ldots, 5$. At the design stage, equally spaced information levels are assumed and calculations show that a maximum information $\mathcal{I}_5 = 1.049\, \mathcal{I}_f$ is needed for the boundaries to meet up with $a_5 = b_5$.

Suppose now that at the second analysis, information becomes available that leads the investigators to seek power 0.9 at $\theta = \delta/2$ rather than $\theta = \delta$. Since this decision is independent of data observed in the study, one might argue a design modification could be made without prejudicing the type I error rate. However, it would be difficult to prove the data revealed at interim analyses had played no part in the decision to re-design. We consider design modification according to Cui et al's (1999) general method. We choose γ so that the conditional power under $\theta = \delta/2$ given the observed value of S_2 is equal to $1 - \beta = 0.9$, with

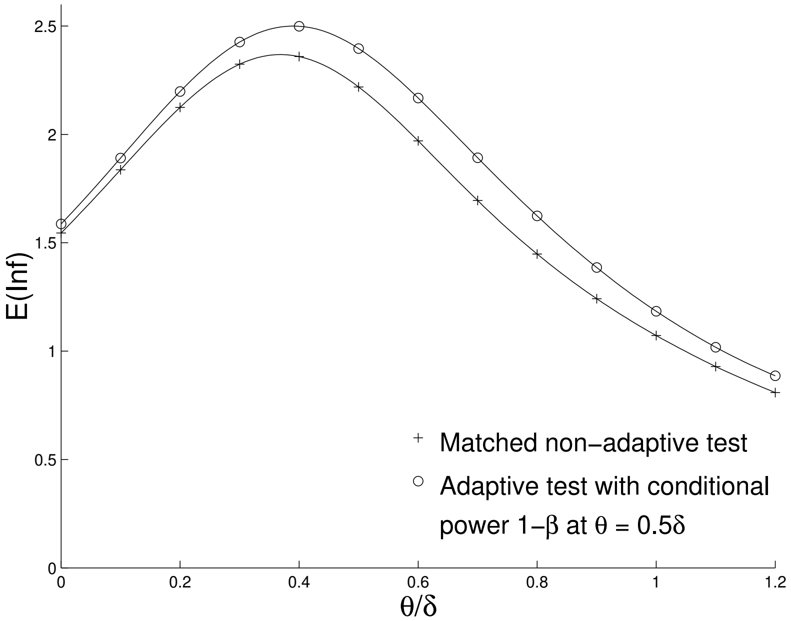
Figure 1: Power of the original test ($\rho = 3$), Cui et al's adaptive design with sample size revised to attain power at $\theta = 0.5\,\delta$, and matched non-adaptive test ($\rho = 0.75$) with power 0.9 at $\theta = 0.59\,\delta$.



the exception that γ is truncated to lie in the range 1 to 6, so sample size is never reduced and the maximum total information is increased by at most a factor of 4. Figure 1 shows the power curve of the adaptive test lies well above that of the original group sequential design. The power 0.78 attained at $\theta = 0.5\,\delta$ falls short of the target of 0.9 due to the impossibility of increasing conditional power when the test has already terminated to accept H_0 and the truncation of γ for values of S_2 just above a_2

It is of interest to assess the cost of the delay in learning the ultimate objective of the study. Our comparison is with a ρ-family error spending test with $\rho = 0.75$, power 0.9 at $0.59\,\delta$ and the first four analyses at fractions 0.1, 0.2, 0.45 and 0.7 of the final information level $\mathcal{I}_5 = 3.78\,\mathcal{I}_f$. This choice ensures the power of the non-adaptive test is everywhere as high as that of the adaptive test, as seen in Figure 1, and the expected information curves of the two tests are of a similar shape. Figure 2 shows the expected information on termination as a function of θ/δ for these two tests; the vertical axis is in units of $\mathcal{I}_f$, the information required in a fixed sample size with power 0.9 at $\theta = \delta$. Together, Figures 1 and 2 show the

Figure 2: $E_\theta(\mathcal{I})$ of Cui et al's adaptive design with sample size revised to attain power at $\theta = 0.5\,\delta$ and of the matched non-adaptive test with power 0.9 at $\theta = 0.59\,\delta$, expressed in units of $\mathcal{I}_f$.



non-adaptive test dominates the adaptive test in both power and expected information over the range of θ values. Also, the non-adaptive test's maximum information level of $3.78\,\mathcal{I}_f$ is about 10% lower than the adaptive test's $4.20\mathcal{I}_f$.

It is useful to have a single summary of relative efficiency when two tests differ in both power and expected information. If test A with type I error rate α at $\theta = 0$ has power $1 - b_A(\theta)$ and expected information $E_{A,\theta}(\mathcal{I})$ under a particular $\theta > 0$, we define its efficiency index at θ to be

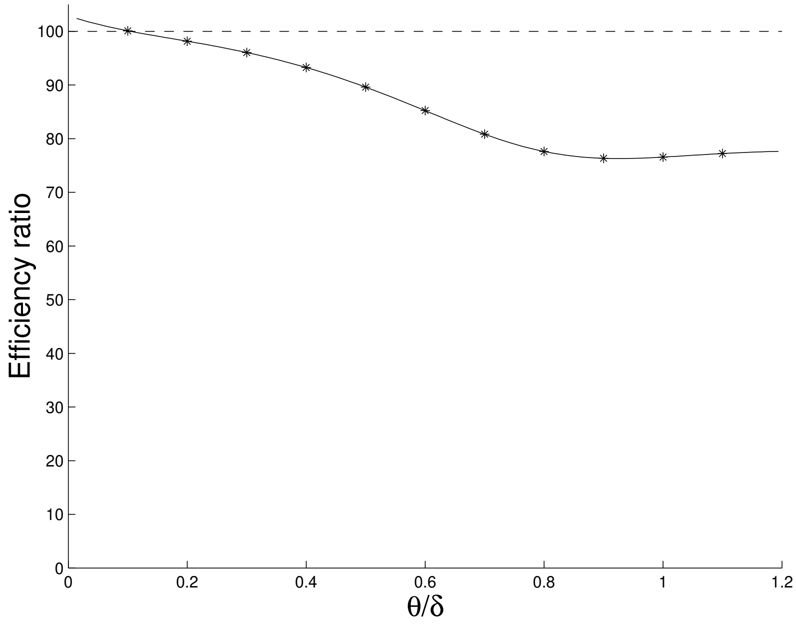
$$EI_A(\theta) \;=\; \frac{\{z_\alpha + z_{b_A(\theta)}\}^2}{\theta^2}\,\frac{1}{E_{A,\theta}(\mathcal{I})},$$

the ratio of the information needed to achieve power $1 - b_A(\theta)$ in a fixed sample test to $E_{A,\theta}(\mathcal{I})$. In comparing tests A and B, we take the ratio of their efficiency indices to obtain the efficiency ratio

$$ER_{A,B}(\theta) \;=\; \frac{EI_A(\theta)}{EI_B(\theta)} \times 100. \;=\; \frac{E_{B,\theta}(\mathcal{I})}{E_{A,\theta}(\mathcal{I})}\,\frac{\{z_\alpha + z_{b_A(\theta)}\}^2}{\{z_\alpha + z_{b_B(\theta)}\}^2} \times 100.$$

This can be regarded as a ratio of expected information for the two tests adjusted for the difference in attained power.

Figure 3: Efficiency ratio between Cui et al's adaptive design with sample size revised to attain power at $\theta = 0.5\,\delta$ and the matched non-adaptive test with power 0.9 at $\theta = 0.59\,\delta$.

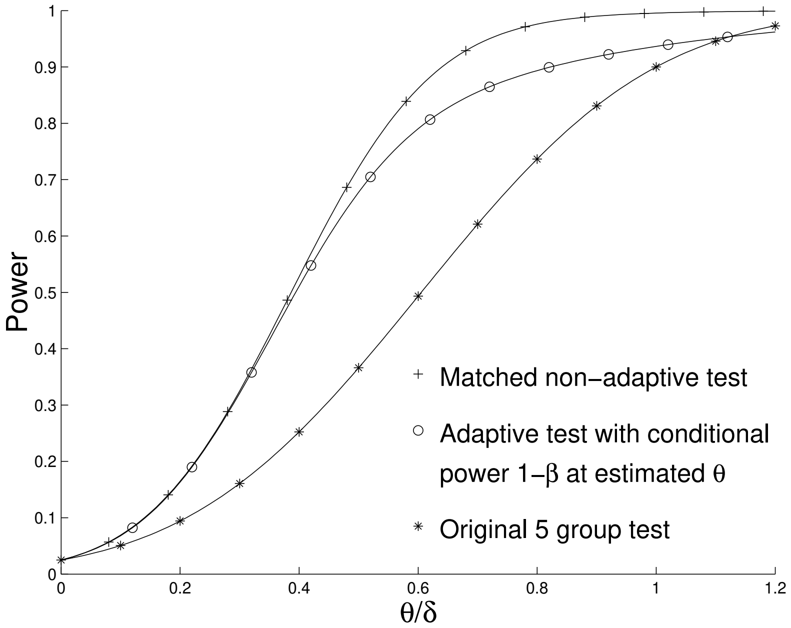


The plot in Figure 3 of the efficiency ratio between the adaptive and non-adaptive tests for our example quantifies the cost of delay in learning the study's objective as a decrease in efficiency of over 20% at higher values of θ, falling to around zero near $\theta = 0$. Values of the efficiency ratio in excess of 100 just above $\theta = 0$ reflect slightly higher power of the adaptive test, not visible to the naked eye in Figure 1.

2.3 Example 2: Re-design in response to internal information

We start with the same initial test as in Example 1, but now suppose the decision to modify the design at the second analysis is prompted by the estimate $\widehat{\theta}_2 = S_2/\mathcal{I}_2$ and the realisation that high power is desirable at lower values of θ which were overlooked originally but now appear plausible in the light of interim data. This time we choose γ so that conditional power given the observed S_2, if θ is in fact equal to the current estimate $\widehat{\theta}_2$, is equal to $1 - \beta = 0.9$. A decrease in sample size is allowed if $\widehat{\theta}_2$ is sufficiently high to imply $\gamma < 1$. As in Example 1, γ is truncated to 6 to restrict the maximum information level to at most 4 times that of the original design; this has the effect that conditional power is equal to 0.9 for $\widehat{\theta} \geq 0.5\,\delta$ but

Figure 4: Power of the original test ($\rho = 3$), Cui et al's adaptive design with sample size revised to attain power at $\theta = \widehat{\theta}_2$, and matched non-adaptive test ($\rho = 0.75$) with power 0.9 at $\theta = 0.64\,\delta$.

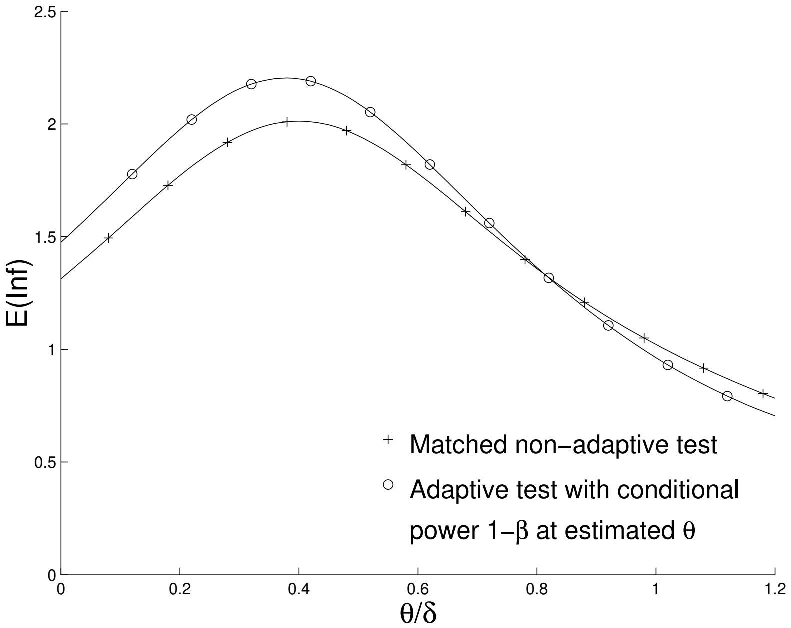


lower for smaller values of $\widehat{\theta}$.

The power curves in Figure 4 show this adaptation has been effective in increasing power above that of the original test, with power at $\theta = \delta/2$ rising from 0.37 to 0.68. In this example, the reason for re-design arose purely from observing $\widehat{\theta}_2$ and did not depend on information from external sources. It should, therefore, have been possible for investigators to consider at the design stage how they would respond to data seen at the second analysis. Let us suppose the above adaptive rule is in accord with such considerations and the power curve in Figure 4 is deemed to be satisfactory. We shall compare this adaptive design with a non-adaptive group sequential test achieving similar power that could have been chosen for the original study design. Our choice is the error spending test from the ρ-family with $\rho = 0.75$, power 0.9 at $0.64\,\delta$ and the first four analyses at fractions 0.1, 0.2, 0.45 and 0.7 of the final information level $\mathcal{I}_5 = 3.21\,\mathcal{I}_f$. Figure 4 shows the power of this non-adaptive test exceeds that of the adaptive test at all θ values and by a substantial margin at the highest θs.

Figure 5 shows that the non-adaptive test has considerably lower expected information

Figure 5: $E(\mathcal{I})$ of Cui et al's adaptive design with sample size revised to attain power at $\theta = \widehat{\theta}_2$ and of the matched non-adaptive test with power 0.9 at $\theta = 0.64\,\delta$, expressed in units of $\mathcal{I}_f$.

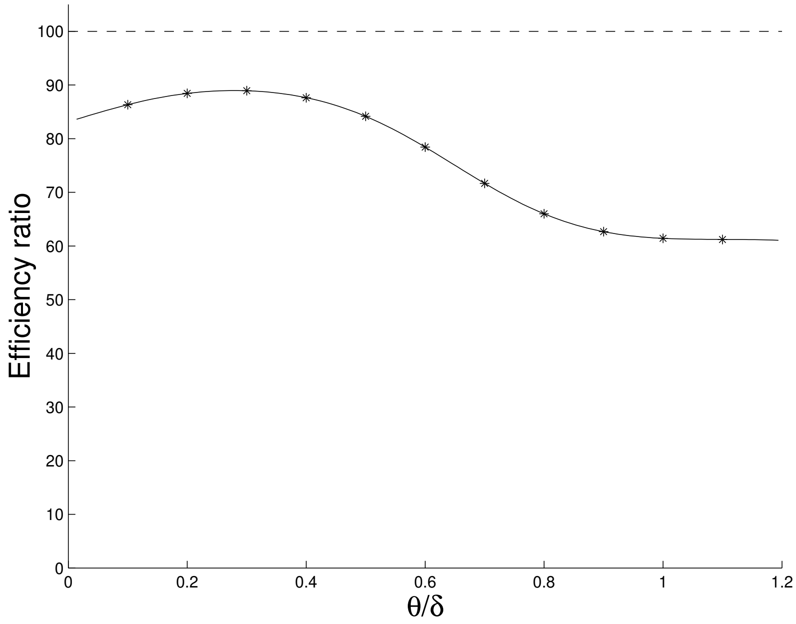


over a wide range of θ values but slightly higher expected information for θ above $0.8\,\delta$ where the non-adaptive test's power advantage is greatest. The efficiency ratio is particularly helpful here. The plot in Figure 6 shows that, with adjustment for attained power, the adaptive test is up to 39% less efficient than the non-adaptive alternative. The maximum information of $4.20\,\mathcal{I}_f$ for the adaptive test is also substantially higher than the non-adaptive test's $3.21\,\mathcal{I}_f$.

2.4 Discussion of examples

The positive conclusion from the preceding examples is that adaptive methods do exist for making mid-course design modifications to meet changes in study objectives due to external or internal factors while preserving the type I error rate. Although a more cost effective design could have been chosen had the ultimate objective been known at the outset, this is not an option in the first example; moreover, it would appear that instances of under-powered studies in need of mid-course rescue continue to arise.

Figure 6: Efficiency ratio between Cui et al's adaptive design with sample size revised to attain power at $\theta = \widehat{\theta}_2$ and the matched non-adaptive test with power 0.9 at $\theta = 0.64\,\delta$.



The negative aspect of flexible adaptive designs is their inefficiency relative to designs set up to achieve the correct power requirement at the outset. Use of non-sufficient statistics as a result of the weighting by $\gamma^{-1/2}$ in (3) is a source of inefficiency in both examples. Part of the additional efficiency loss in Example 2 can be attributed to over-reliance on the interim estimate $\widehat{\theta}_2$ which is, in fact, highly variable. This results in random variation in sample size that is in itself inefficient: see Jennison & Turnbull (2003) for further discussion of this point in the context of a two-stage design. The lack of precision of early estimates of θ argues against the "wait and see" approach in which a firm decision on the desired power curve is delayed until interim data are available and an adaptive design modification is then used to attain this.

We have carried out many more comparisons of adaptive designs and matched non-adaptive error spending tests with similar qualitative conclusions to the two examples described here. In general, allowing a greater increase in the maximum sample size of an adaptive test leads to higher inefficiency. The examples of Sections 2.2 and 2.3 follow the recommendation of many authors to base sample size revision on conditional power. The adaptive tests have the benefit of early stopping to accept H_0 in the original design: this stopping rule was carefully chosen to reduce the risk of stopping to accept H_0 under values of θ in the range $\delta/2$ to δ under which it may be decided later that higher power is desirable. In our experience, adaptations which make a noticeable change to a test's power curve are liable to introduce inefficiency at least as great as that seen in our two examples and often much larger; see Jennison & Turnbull (2003) for an example of a two-stage adaptive design with much higher efficiency loss. In the following sections we complement this empirical evidence with theory and numerical evaluation of optimal tests within well-defined adaptive and non-adaptive classes.

3 Theory of optimal adaptive group sequential designs

Consider the problem of testing H_0: $\theta \leq 0$ against $\theta > 0$. Suppose there are M possible analysis times to choose from with associated information levels $\mathcal{I}_1, \ldots, \mathcal{I}_M$. We assume the statistic S_m is sufficient for θ at the analysis with information $\mathcal{I}_m$ and the sequence $S_1, \ldots, S_M$ has the joint distribution specified in (1). We shall consider group sequential tests with a maximum of K analyses, where $K \leq M$. When the study continues at an interim analysis, the timing of the next analysis is chosen as a function of currently observed data. The set of available information levels $\{\mathcal{I}_1, \ldots, \mathcal{I}_M\}$ is to be regarded as fixed. For adaptive tests, we are interested in $M \gg K$; the case $M = K$ applies to non-adaptive group sequential tests.

Denote the indices of the information levels arising in a particular realisation of the experiment by m_1, m_2, $\ldots$, so the kth analysis has information level $\mathcal{I}_{m_k}$. An adaptive group sequential design is defined by a decision rule specifying the action at each stage. A deterministic rule specifies $m_1 \in \{1, \ldots, M - K + 1\}$, then for each k and observed X_k it chooses an action from the set of possibilities: stop and accept H_0; stop and reject H_0; continue to analysis $k+1$ at information level $\mathcal{I}_{m_{k+1}}$ where $m_{k+1} \in \{m_k + 1, \ldots, M - K + k + 1\}$. The option of continuing is not available at analysis K. It is helpful in deriving theoretical results to allow randomised rules which correspond to probability distributions on the set of deterministic rules. We denote the set of all randomised and non-randomised rules by $\mathcal{D}$.

Let $\mathcal{A}$ denote the final decision taken, either to accept or to reject H_0, and $\mathcal{I}$ the information on termination. The risk or expected loss of decision rule d comprises the type I error function

$$R_1(\theta, d) = P_\theta(\mathcal{A} = \text{Reject } H_0), \quad \theta \leq 0,$$

the type II error function

$$R_2(\theta, d) = P_\theta(\mathcal{A} = \text{Accept } H_0), \quad \theta > 0,$$

and the expected information function

$$R_3(\theta, d) = E_\theta(\mathcal{I}).$$

We assume the preferred decision is to reject H_0 whenever $\theta > 0$ but it is straightforward to modify R_1 and R_2 to change the threshold for this preference. We avoid technical difficulties in our proofs by considering risk at a finite set of θ values, $\Theta = \{\theta_1, \ldots, \theta_Q\}$, where $\theta_1 < \ldots < \theta_P \leq 0 < \theta_{P+1} < \ldots < \theta_Q$. This restriction has little impact on the practical implications of theoretical results as it is perfectly acceptable to take, say, ten million points spaced at very small intervals over the range of θ values of interest.

We combine R_1, R_2 and R_3 into a single risk vector

$$\begin{aligned} R(d) &= (R(1, d), \ldots, R(2Q, d)) \\ &= (R_1(\theta_1, d), \ldots, R_1(\theta_P, d), R_2(\theta_{P+1}, d), \ldots, R_2(\theta_Q, d), \\ &\qquad R_3(\theta_1, d), \ldots, R_3(\theta_Q, d)). \end{aligned}$$

A decision rule $d \in \mathcal{D}$ is said to be inadmissible if there is a rule d' with

$$R(i, d') \leq R(i, d) \quad \text{for all } i = 1, \ldots, 2Q$$

and

$$R(i, d') < R(i, d) \quad \text{for at least one } i \in \{1, \ldots, 2Q\}.$$

A decision rule which is not inadmissible is admissible.

A Bayes decision problem is defined by a prior distribution $\pi = (\pi_1, \ldots, \pi_Q)$ on Θ and costs for each element of the risk vector R. The Bayes risk is

$$\sum_{q=1}^{Q} \pi_q \sum_{j=1}^{3} c_{qj} R_j(\theta_q, d) \tag{5}$$

where c_{q1} is the cost of rejecting H_0, c_{q2} the cost of accepting H_0 and c_{q3} the cost per unit of observed information under $\theta = \theta_q$. Here $c_{q1} = 0$ for $q > P$ and $c_{q2} = 0$ for $q \leq P$. We shall write the Bayes risk as

$$w^T R(d) \;=\; \sum_{i=1}^{2Q} w(i) R(i,d), \tag{6}$$

where each $w(i) \geq 0$, $i = 1, \ldots, 2Q$. A Bayes rule is a decision rule d which minimises the Bayes risk for some w. In characterising the admissible rules as Bayes rules, the risk set

$$\mathcal{S} = \{R(d); d \in \mathcal{D}\}$$

plays a central role.

Theorem 1. For the problem defined above, the risk set $\mathcal{S}$ is closed and convex.

Corollary 1. Each admissible rule $d \in \mathcal{D}$ is a Bayes rule for a problem in which $w(i) \geq 0$, $i = 1, \ldots, 2Q$, and at least two of the following hold:

1. $w(i) > 0$ for some $i \leq P$
2. $w(i) > 0$ for some $P + 1 \leq i \leq Q$
3. $w(i) > 0$ for some $i \geq Q + 1$.

Let $\mathcal{D}_{NS}$ denote the set of "non-sequential" decision rules which terminate at the minimum information level $\mathcal{I}_1$ with probability 1 or terminate at the maximum information level $\mathcal{I}_M$ with probability 1. Then, each admissible rule in $\mathcal{D} \backslash \mathcal{D}_{NS}$ is a Bayes rule for a problem in which all three of the above conditions hold. $\square$

We refer the reader to Chapter 2 of Ferguson (1967) for background to complete class theorems which show, broadly speaking, that admissible rules are Bayes and vice versa, and for proofs of the supporting hyperplane and separating hyperplane theorems. The first step in proving a complete class theorem is to show that the risk set is closed and convex. Proving the risk set is closed is often difficult and our problem is no exception. The proofs of Theorem 1 and Corollary 1 are given in Appendix 1.

Ferguson (1967, Sec. 7.1 and 7.2) and Brown, Cohen & Strawderman (1980) characterise admissible rules in the non-adaptive case, $M = K$, but combine error rates and expected sample size into a single risk for each value of θ. Keeping error rates and expected information as separate elements of the risk vector in our treatment means that when a decision rule is shown to be inadmissible, the dominating rule has both a superior power function and a lower expected information function — as was very nearly the case in Example 1 of Section 2. In the non-adaptive setting, Chang (1996) considers a risk vector of length three, comprising the type I error rate at a single θ_0, power at an alternative θ_1 and expected sample size at $\theta = (\theta_0 + \theta_1)/2$. He appeals to standard decision theory arguments to conclude that admissible designs are Bayes but does not provide a proof that the risk set is closed.

It follows from Corollary 1 that any decision rule which is properly sequential in that it produces a non-degenerate distribution of sample sizes and which is not a Bayes rule for a problem satisfying the three conditions of the corollary is inadmissible. Since a Bayes problem always has a solution based on sufficient statistics, this establishes the general principle that a sequential test should be defined as a function of the sequence of sufficient statistics for θ. The fact that an adaptive rule is defined through the non-sufficient statistics (3) does not necessarily mean the rule is not Bayes: if the factor γ by which group sizes after analysis j are increased is a one-to-one function of S_j, the stopping rule at information levels $\mathcal{I}_j + \gamma(\mathcal{I}_{j+1} - \mathcal{I}_j), \ldots, \mathcal{I}_j + \gamma(\mathcal{I}_K - \mathcal{I}_j)$ can be re-expressed in terms of the sufficient statistics

$$S_k = S_j + \sum_{i=j+1}^{k} (S'_i - S'_{i-1}), \quad k = j+1, \ldots, K.$$

In the examples of Section 2, truncation of γ to a maximum value means the same sequence of future information levels arises for an interval of S_j values and consideration of the decision rule to accept or reject H_0 at analysis K is sufficient to show these adaptive plans fail to agree with any given Bayes rule on a set of sample paths with positive probability. The variance spending tests of Shen & Fisher (1999) are easily dealt with since the sequence of information levels is fixed and it is the weights for each group of observations that are chosen adaptively: any non-trivial departure from equal weights implies a positive probability of disagreement with a given Bayes rule so, by the corollary, the variance spending test is inadmissible.

Corollary 1 with $K < M$ characterises admissible designs which are truly adaptive in that data-dependent choices are made for each successive group size. Adaptive designs using non-sufficient statistics, as required in the flexible adaptive approach, are dominated by admissible designs based on sufficient statistics. The Bayes optimal dominating designs are examples of the optimal "sequentially planned sequential designs" described by Schmitz (1993), all aspects of which are pre-planned.

The general class of "designed extension" procedures proposed by Proschan & Hunsberger (1995) comprises pre-planned adaptive designs with $K = 2$. Some of these procedures can be expressed in terms of sufficient statistics and Li, Shih, Xie & Lu (2002) advocate one such procedure. While it is necessary for admissibility that a design can be expressed as a function of the sufficient statistic, this is not a sufficient condition: Corollary 1 shows what is required of the conditional error function and sampling rule for such a design to be admissible.

In the case $K = M$, adaptive tests become non-adaptive and the corollary tells us that within the class of non-adaptive group sequential tests, those with stopping rules or decision rules based on non-sufficient statistics are dominated by Bayes optimal designs defined in terms of sufficient statistics. With the sequence of M possible information levels held fixed, increasing the maximum number of analyses from K to M only adds to the available options. Thus for any Bayes problem, the optimal adaptive test with $K < M$ analyses can do no better than the optimal non-adaptive design with M analyses. It follows that any K-analysis adaptive design using non-sufficient statistics is dominated by a non-adaptive M-analysis design based on sufficient statistics. This conclusion is similar in nature to the result proved by Tsiatis & Mehta (2003) who start with a K-analysis adaptive design using non-sufficient statistics and construct an M-analysis non-adaptive test which increases power and reduces expected information at parameter values θ in the alternative hypothesis. Our result goes further in showing the type I error probability and expected information function can also be maintained or reduced at all values of θ in a composite null hypothesis. Tsiatis & Mehta (2003) consider only a simple null hypothesis and expected information at this value of θ may increase in their construction; also, while this construction improves on the test based on non-sufficient statistics, the result is not necessarily an admissible test. Our Corollary 1 provides a characterisation of admissible tests and we shall use this in Section 4 to derive and study optimal tests.

Calculations for optimal group sequential tests in Eales & Jennison (1992) show that most of the reductions in expected sample size to be gained by sequential analysis are obtained in tests with as few as 5 or 10 analyses, supporting Tsiatis & Mehta's (2003, p. 375) argument that non-adaptive group sequential tests with 5 or 10 groups should be able to match the performance of adaptive tests fairly closely. This leaves open the question of how great an advantage carefully designed adaptive tests may have when the maximum number of analyses is restricted to $K = 2$ or 3. Allowing adaptive choice of group sizes extends the class of sequential designs and there are intuitive arguments why, for example, one might wish to take a smaller group size when current data lie close to the testing boundary. If the efficiency gains for optimal adaptive tests are substantial, there could be a case for using pre-planned adaptive designs. Also, advantages of adaptivity might mean that the performance of sub-optimal tests using non-sufficient statistics is still comparable with that of the best non-adaptive tests. We shall explore the extent of these possible gains from adaptivity in Section 4.

As well as pointing to the need for efficient tests to be defined in terms of sufficient statistics, Corollary 1 also imposes strict requirements on the stopping rule and sampling rule of an admissible adaptive test. It will be of interest to see how similar the optimal rules derived in Section 4 are to the proposals involving conditional power at a pre-specified or estimated effect size that have been proposed by other authors.

4 Computing optimal adaptive designs

The theory of Section 3 indicates the importance of Bayes optimal adaptive designs as the set of such designs coincides with the class of admissible adaptive tests. Eales & Jennison (1992) and Barber & Jennison (2002) have exploited the analogous correspondence in the non-adaptive setting to compute optimal frequentist tests, using backwards induction to solve an unconstrained Bayes decision problem and searching over costs in this Bayes problem to find the optimal test with a specific type I error rate and power.

We have extended this computational technique to find optimal adaptive tests. The results reported here are for tests of H_0: $\theta \leq 0$ against $\theta > 0$ with type I error rate $\alpha = 0.025$ under $\theta = 0$ and power $1 - \beta = 0.9$ at $\theta = \delta$. Tests are designed to minimise the integral of $E_\theta(\mathcal{I})$

over a normal distribution for θ with mean δ and standard deviation $\delta/2$, reflecting optimism that the effect size may be higher than the value δ at which power $1-\beta$ is set and a desire to stop particularly early if this is the case. Optimal adaptive tests are calculated with $M = 50$ and $\mathcal{I}_1, \ldots, \mathcal{I}_{50}$ equally spaced between 0 and $R\mathcal{I}_f$, for several values of R, where $\mathcal{I}_f$ is the information needed for a fixed sample test given in (4). Comparison of results for $M = 10$, 25 and 50 indicates no significant improvement is to be obtained by increasing M further.

In order to find the optimal tests we adopt the device of Eales & Jennison (1992) and formulate Bayes decision problems with a prior comprising point masses at $\theta = 0$ and δ mixed with a $N(\delta,\ \delta^2/4)$ kernel, costs c_1 for rejecting H_0 when $\theta = 0$ and c_2 for accepting H_0 when $\theta = \delta$, and a cost of one per unit of observed information under the continuous component of the prior. The backwards induction algorithm for finding Bayes optimal adaptive rules is similar to that employed by Eales & Jennison (1992) and Barber & Jennison (2002) but now the set of interim states is indexed both by the analysis number k and the index m_k of the information level at which this analysis occurs. Further details of the algorithm are given in Appendix 2. These calculations provide the optimal adaptive tests described in abstract form, but without numerical examples, by Schmitz (1993).

Let $f(\theta)$ denote the density of a $N(\delta,\ \delta^2/4)$ distribution. Table 1 shows the integrated expected information

$$\int E_\theta(\mathcal{I}) f(\theta)\, d\theta \tag{7}$$

for optimal adaptive tests expressed as a percentage of $\mathcal{I}_f$. The numbers of analyses are $K = 2$, 3, 4, 5, 6, 8 and 10 and maximum sample size is $R = 1.05$, 1.1, 1.2 and 1.3 times $\mathcal{I}_f$. The tables also show the minimum possible value of (7) for (a) a non-adaptive test with K analyses at information levels $\mathcal{I}_1, \ldots, \mathcal{I}_K$ placed optimally between 0 and $R\mathcal{I}_f$ and (b) a non-adaptive test with K analyses at information levels equally spaced between 0 and $R\mathcal{I}_f$. The search for optimal information levels in (a) was by the simplex algorithm of Nelder & Mead (1965).

The results show that adaptive tests can reduce expected information well below that of a fixed sample test. This reduction increases with the number of analyses K and, at least initially, with the factor R specifying the maximum allowable information. However, well chosen non-adaptive tests are almost as efficient. For a given number of analyses K and maximum information $R\mathcal{I}_f$, the average $E_\theta(\mathcal{I})$ of the best non-adaptive test with equally

Table 1: Minimum possible values of $\int E_\theta(\mathcal{I}) f(\theta)\, d\theta$ for adaptive and non-adaptive tests with type I error rate $\alpha = 0.025$, power $1 - \beta = 0.9$ at $\theta = \delta$, K analyses and maximum information level $R\,\mathcal{I}_f$. Values are expressed as a percentage of $\mathcal{I}_f$.

	Number of analyses, K	Adaptive tests	Non-adaptive tests with optimised $\mathcal{I}_1, \ldots, \mathcal{I}_K$	Non-adaptive tests with $\mathcal{I}_k = (k/K)R\,\mathcal{I}_f$
$R = 1.05$				
	2	74.7	74.7	74.7
	3	68.0	68.8	69.0
	4	64.9	66.2	66.5
	5	63.3	64.7	65.1
	6	62.3	63.7	64.1
	8	61.1	62.5	62.8
	10	60.5	61.8	62.1
$R = 1.1$				
	2	73.2	73.3	73.8
	3	66.0	66.8	67.0
	4	62.8	63.9	64.2
	5	61.0	62.3	62.7
	6	59.9	61.3	61.6
	8	58.6	60.0	60.3
	10	58.0	59.3	59.5
$R = 1.2$				
	2	72.5	73.2	74.8
	3	64.8	65.6	66.1
	4	61.2	62.4	62.7
	5	59.2	60.5	60.9
	6	58.0	59.4	59.8
	8	56.6	58.0	58.3
	10	55.9	57.2	57.5
$R = 1.3$				
	2	72.4	73.0	77.1
	3	64.5	65.5	66.6
	4	60.8	61.9	62.5
	5	58.6	60.0	60.5
	6	57.3	58.7	59.2
	8	55.8	57.2	57.6
	10	55.0	56.3	56.7

spaced information levels is within 2% of $\mathcal{I}_f$ of the optimal adaptive test in most cases: exceptions when $K = 2$ and $R \geq 1.2$ or $K = 3$ and $R = 1.3$ are due to these values of R being unnecessarily high for the number of analyses. Eales & Jennison (1992) and Brittain & Bailey (1993) have proposed optimising the information levels at which analyses of a group sequential test are performed. Optimising $\mathcal{I}_1, \ldots, \mathcal{I}_K$ subject to an upper bound of $R\mathcal{I}_f$ for $\mathcal{I}_K$ gives the middle column of results in Table 1, none of which is more than 1.5% of $\mathcal{I}_f$ higher than the average $E_\theta(\mathcal{I})$ of the optimal adaptive test. We have found similar differences with other choices of α and β and when tests are optimised for a variety of criteria, for example, $E_{\delta/2}(\mathcal{I})$ or $\{E_0(\mathcal{I}) + E_\delta(\mathcal{I})\}/2$. Optimising the information levels is more crucial for criteria such as $\{E_0(\mathcal{I}) + E_\delta(\mathcal{I}) + E_{H\delta}(\mathcal{I})\}/3$ with $H = 3$ or 4, say, where it is important to stop early under a high θ value: in these cases non-adaptive tests with optimised information levels still do almost as well as the optimal adaptive tests.

The small advantages of adaptive designs over non-adaptive tests are in keeping with results reported by Posch, Bauer & Brannath (2003) for the case $K = 2$. Our results are more far-reaching in that we optimise over completely general sampling rules and stopping boundaries and consider higher values of K. Controlling comparisons at a fixed maximum information level is appropriate since the maximum possible sample size is often constrained in practice and its value has a substantial effect on the performance of a sequential design. Even if the limited benefits of adaptive designs are deemed worthwhile, from an administrative perspective it may well be preferable to achieve these in a non-adaptive design with additional analyses. That these gains can be obtained in many instances with just one or two extra analyses is a very tight result when compared with the argument that one can expect to dominate an adaptive test with 2 or 3 analyses by a non-adaptive test with 10 analyses since this is known to deliver almost all the benefits of continuous monitoring.

The gap between the best adaptive and best non-adaptive tests is large enough that an adaptive test based on non-sufficient statistics may not be dominated by a non-adaptive test. However the margin for error here is small. We have found sampling rules for optimal adaptive tests to follow a consistent pattern: at analysis k, smaller increments in information are chosen when S_{m_k} is close to either stopping boundary and larger increments are taken when S_{m_k} is near the middle of the continuation region. This is in contrast to the monotone increase in information increments as S_{m_k} decreases seen in sampling rules based on constant conditional power at $\theta = \delta/2$ or $\theta = \widehat{\theta}$, as in the examples of Section 2, or based on constant conditional power at $\theta = \delta$ as suggested by Denne (2001). Thus, although conditional power criteria have an intuitive appeal, they should not be expected to lead to efficient sequential designs.

5 Discussion

Non-adaptive group sequential tests are well studied and optimal tests have been derived for a variety of design criteria. Barber & Jennison (2002) show that members of the ρ-family of error spending tests with equally spaced information levels are highly efficient for a range of criteria involving $E_\theta(\mathcal{I})$ at values of θ between $-\delta/2$ and $3\delta/2$. These error spending tests are easily implemented and provide flexibility to deal with unpredictable information sequences.

Adding an element of adaptivity in pre-planned group sequential designs, as proposed by Schmitz (1993), produces a small benefit over non-adaptive tests with the same number of analyses. However, similar or better performance is often achieved by a non-adaptive design with one extra analysis, avoiding the administrative complications of a pre-planned adaptive design.

Using adaptive methods in an unplanned manner offers flexibility to study organisers but, since the sufficiency principle is contravened, there is an automatic efficiency cost. One argument for flexible adaptive designs is that they allow investigators to choose a study's power curve in the light of early estimates of the effect size, θ. Such an approach may be suggested when there is uncertainty about the likely effect and, in particular, optimistic estimates of the effect size are considerably larger than the minimum clinically or commercially significant effect. Schäfer & Müller (2004) consider tests for a range of detectable treatment effects and propose a novel group sequential design in which attention shifts to smaller effect sizes at successive analyses. An alternative solution is simply to specify high power at the small but clinically significant effect size and choose a stopping boundary that gives low expected sample size under the larger effects investigators hope to see. Reducing expected information under values of θ well above that at which power is set may require specialised versions of standard group sequential tests. Examples of these are the ρ-family error spending tests seen in Section 2 with $\rho = 0.75$ and a special sequence of information levels including a couple of very early analyses. Jennison & Turnbull (2004) investigate related tests, assessing performance by the average expected information $\{E_0(\mathcal{I}) + E_\delta(\mathcal{I}) + E_{H\delta}(\mathcal{I})\}/3$ with $H = 2$, 3 and 4; they show that ρ-family error spending tests with an optimised information level for the first analysis perform almost as well as the best pre-planned adaptive designs.

In some studies a nuisance parameter, such as the variance of a normal response, determines the sample size needed to achieve a given power at a specified effect size. There is a substantial literature on methods for modifying sample size in response to estimates of such a nuisance parameter and these methods can be incorporated in group sequential tests. In the "information-based monitoring" approach described by Mehta & Tsiatis (2001), the maximum information level needed in an error spending test is known at the outset but, since the relationship between sample size and information depends on parameters which are initially unknown, the target sample size is adjusted repeatedly during the study as new estimates of these parameter are obtained. Since this modification of sample size follows pre-specified rules, independent of the estimated treatment effect, there is no need for any adjustment to preserve the type I error rate. The importance for our discussion is that this process operates independently of any design changes that might be made to alter the original power requirement and this form of sample size adjustment should not be confused with the issues addressed in this paper.

A key role that remains for flexible adaptive methods is to help investigators respond to unexpected external events. As Müller & Schäfer (2001) and Posch, Bauer & Brannath (2003) point out, it is good practice to design a study as efficiently as possible given initial assumptions, so the benefits of this design are obtained in the usual circumstances where no mid-course change is required. But, if the unexpected occurs, adaptive methods are available to deal with this. The approach based on maintaining conditional type I error probability put forward by Denne (2001) and by Müller & Schäfer (2001) is particularly promising as it has the potential to be used with error spending designs that already adapt to unpredictable information sequences and, possibly, update sample size in response to estimates of a nuisance parameter.

Finally, the use of flexible adaptive methods to rescue an under-powered study should not be overlooked. While it is easy to be critical of a poor initial choice of sample size, it would be naive to think such problems will cease to occur.

ACKNOWLEDGEMENT

This research was supported in part by an NIH grant.

APPENDIX 1

Proofs

Proof of Theorem 1. In the problem formulated in Section 3, we restrict attention to cases with $\pi_q = 1/Q$, $q = 1, \ldots, Q$. Then, the Bayes risk $w^T R(d)$ in (6) implies costs under $\theta = \theta_q$ of $Qw(q)$ for a wrong decision and $Qw(Q+q)\mathcal{I}$ for observed information $\mathcal{I}$. Since any expression (5) can be written in the form (6), this does not reduce the class of problems considered. It will be helpful to treat the Bayes risk as the expectation, under prior $\pi_q = 1/Q$, $q = 1, \ldots, Q$, of the "loss function" $L(w, \mathcal{A}, \mathcal{I}, \theta)$ where

$$L(w, \mathcal{A}, \mathcal{I}, \theta_q) \;=\; \begin{cases} Q\, w(q)\, I(\mathcal{A} = \text{Reject } H_0) \;+\; Q\, w(Q+q)\, \mathcal{I}, & q \leq P, \\ Q\, w(q)\, I(\mathcal{A} = \text{Accept } H_0) \;+\; Q\, w(Q+q)\, \mathcal{I}, & q > P. \end{cases} \tag{8}$$

Suppose risk vectors r_1 and r_2 belong to the risk set $\mathcal{S}$ and $0 < \lambda < 1$. Then, there are decision rules d_1 and d_2 for which r_1 and r_2 are the risk vectors. Define d_3 as the randomised rule which mixes d_1 with probability λ and d_2 with probability $1 - \lambda$. The risk vector of d_3 is $R(d_3) = \lambda r_1 + (1 - \lambda) r_2 \in \mathcal{S}$. Showing that a general linear combination of points in $\mathcal{S}$ is also in $\mathcal{S}$ proves that $\mathcal{S}$ is convex.

To prove that $\mathcal{S}$ is closed, we take a general point r_1 on the boundary of $\mathcal{S}$ and show that it is in $\mathcal{S}$. We use the supporting hyperplane at r_1 to define a Bayes decision problem. A decision rule solving this Bayes problem can be found by backwards induction; see the proof of Lemma 1 for details. The risk vector of this rule is in $\mathcal{S}$ and lies in the supporting hyperplane. If the hyperplane intersects the closure of $\mathcal{S}$ in a single point, this must be the original point r_1. However, the supporting hyperplane may intersect the closure of $\mathcal{S}$ at a set of points and then we need to prove this set is closed and contains r_1. The full proof of the theorem is by induction. We start by outlining the first two stages of the general scheme to motivate the definition of the inductive hypothesis.

Let $\bar{\mathcal{S}}$ denote the closure of $\mathcal{S}$ and take an arbitrary point r_1 on the boundary of $\bar{\mathcal{S}}$. We wish to show $r_1 \in \mathcal{S}$. By the supporting hyperplane theorem, there is a hyperplane

$$P_1 = \{r \colon w_1^T r = k_1\}$$

which passes through r_1 with $\mathcal{S}$ on one side, i.e., $w_1^T r_1 = k_1$ and $w_1^T r \geq k_1$ for all $r \in \mathcal{S}$. Let

$$\mathcal{S}_1 = P_1 \cap \mathcal{S} \quad \text{and} \quad \mathcal{Q}_1 = P_1 \cap \bar{\mathcal{S}}.$$

Consider choosing a decision rule to minimise $w_1^T R(d)$. If some elements of w_1 are negative this is an unusual Bayes decision problem, but that is unimportant. Direct construction of a decision rule, d_1 say, minimising $w_1^T R(d)$ is possible by backwards induction. Since $w_1^T r \geq k_1$ for all $r \in \mathcal{S}$, we know $w_1^T R(d_1) \geq k_1$. But, there are decision rules with risk vectors approaching r_1, so we also have $w_1^T R(d_1) \leq w_1^T r_1 = k_1$. Hence $w_1^T R(d_1) = k_1$ and $R(d_1) \in \mathcal{S}_1$, demonstrating that $\mathcal{S}_1$ is non-empty. Taking linear combinations of decision rules in the usual way, it is easy to see that $\mathcal{S}_1$ is convex. If $r_1 \in \mathcal{S}_1$, we have the desired result that $r_1 \in \mathcal{S}$; the situation we must consider further is where $\mathcal{S}_1$ is a strict subset of $\mathcal{Q}_1$ and $r_1 \in \mathcal{Q}_1 \backslash \mathcal{S}_1$. Our plan is to show that

(i) $\mathcal{S}_1$ is closed, and then

(ii) $\mathcal{S}_1 = \mathcal{Q}_1$,

from which it follows that $r_1 \in \mathcal{S}_1$. Lemma 1 proves that (ii) holds, given (i). Proving (i) is similar to proving the original theorem but we have made some progress since $\mathcal{S}_1$ is a subset of the $(2Q-1)$-dimensional P_1 whereas $\mathcal{S}$ was a subset of $\Re^{2Q}$.

Taking an arbitrary point r_2 on the boundary of $\mathcal{S}_1$, we can find a supporting hyperplane within P_1,

$$P_2 = \{r \colon w_2^T r = k_2\} \cap P_1$$

for which $w_2^T r_2 = k_2$ and $w_2^T r \geq k_2$ for all $r \in \mathcal{S}_1$. We then define

$$\mathcal{S}_2 = P_2 \cap \mathcal{S} \quad \text{and} \quad \mathcal{Q}_2 = P_2 \cap \bar{\mathcal{S}}.$$

Points in $\mathcal{S}_2$ arise as risk vectors of decision rules solving the following problem: first, minimise $w_1^T R(d)$ then, as a secondary criterion, minimise $w_2^T R(d)$ among rules minimising

$w_1^T R(d)$. Such a rule, d_2 say, can be constructed by backwards induction and, following earlier reasoning, it must satisfy $w_1^T R(d_2) = k_1$ and $w_2^T R(d_2) = k_2$. Thus, $\mathcal{S}_2$ is non-empty and, by the usual argument, convex. We now wish to show (i) $\mathcal{S}_2$ is closed, and then (ii) $\mathcal{S}_2 = \mathcal{Q}_2$, in order to deduce $r_2 \in \mathcal{S}_2$.

Further iterations of this process lead eventually to a non-empty, convex $\mathcal{S}_u$ of dimension zero. As this is a singleton set, it is closed so (i) holds. We still need to show that (ii) holds at this level, i.e., $\mathcal{S}_u = \mathcal{Q}_u$, and work back to deduce $\mathcal{S}_1$ is closed and $\mathcal{S}_1 = \mathcal{Q}_1$. The sequence of hyperplanes and subsets of $\mathcal{S}$ arising in this process is defined below.

For notational consistency, let $\mathcal{S}_0 = \mathcal{S}$ and $P_0 = \Re^{2Q}$. We shall consider sequences $\{(r_v, w_v); v = 1, \ldots, 2Q\}$ such that for $v = 1, \ldots, 2Q$:

> r_v is a point on the boundary of $\mathcal{S}_{v-1}$,
>
> $P_v = \{r : w_v^T r = k_v\} \cap P_{v-1}$ is a supporting hyperplane to $\mathcal{S}_{v-1}$ within P_{v-1} at the point r_v, for which $w_v^T r_v = k_v$ and $w_v^T r \geq k_v$ for all $r \in \mathcal{S}_{v-1}$,
>
> $\mathcal{S}_v = P_v \cap \mathcal{S}$ is non-empty and $\mathcal{Q}_v = P_v \cap \bar{\mathcal{S}}$. (9)

Note that arbitrary choice of boundary points r_v is allowed in (9). A supporting hyperplane P_v exists since $\mathcal{S}_0$ is convex and, hence, so is $\mathcal{S}_{v-1}$; if there is more than one supporting hyperplane, any defining vector w_v may be chosen. To see that each $\mathcal{S}_v$ is non-empty, note that backwards induction can be used to construct a decision rule d_v minimising $w_1^T R(d)$ first, then minimising $w_2^T R(d)$ among rules that minimise $w_1^T R(d)$, and so forth. Arguments outlined above and given more fully in the proof of Lemma 1 show that $w_1^T R(d_v) = w_1^T r_1 = k_1$, etc., so $R(d_v)$ lies in each hyperplane $P_1, \ldots, P_v$ as well as in $\mathcal{S}$, and therefore $R(d_v) \in \mathcal{S}_v$.

Lemma 1, proved below, states that in this setting if, for any $1 \leq v \leq 2Q$, $\mathcal{S}_v$ is closed then $\mathcal{S}_v = \mathcal{Q}_v$. In other words, property (i) implies property (ii) at each level v. We use this lemma in an inductive argument combining results over levels v to prove the theorem. The inductive hypothesis to be proved for $0 \leq h \leq 2Q$ is:

$$\text{If the dimension of } P_v \leq h, \text{ then } \mathcal{S}_v \text{ is closed.} \tag{10}$$

This hypothesis is satisfied for $h = 0$ since then $\mathcal{S}_v$ is a singleton set. To prove the inductive step, suppose (10) is true for $h \leq \tilde{h}$, where $0 \leq \tilde{h} \leq 2Q - 1$. Consider a general $\mathcal{S}_{v-1}$ in a hyperplane P_{v-1} of dimension $\tilde{h}+1$. For any r_v on the boundary of $\mathcal{S}_{v-1}$, take a supporting hyperplane P_v and define $\mathcal{S}_v = P_v \cap \mathcal{S}$ and $\mathcal{Q}_v = P_v \cap \bar{\mathcal{S}}$. The dimension of P_v is $\tilde{h}$ so, by the inductive hypothesis, $\mathcal{S}_v$ is closed. Therefore, by Lemma 1, $\mathcal{S}_v = \mathcal{Q}_v$. Now, r_v is in $\bar{\mathcal{S}}$ and P_v, hence

$$r_v \in \mathcal{Q}_v \;\Rightarrow\; r_v \in \mathcal{S}_v \;\Rightarrow\; r_v \in \mathcal{S} \;\Rightarrow\; r_v \in \mathcal{S}_{v-1}.$$

As r_v was a general boundary point of $\mathcal{S}_{v-1}$, we see that $\mathcal{S}_{v-1}$ is closed and this establishes (10) for $h \leq \tilde{h}+1$. Putting $v=0$ and $h=2Q$ in (10) gives the result that $\mathcal{S}_0 = \mathcal{S}$ is closed, completing the proof of the theorem. □

The proof of Theorem 1 is complicated by the possibility that a Bayes decision problem may have multiple solutions. This does not seem very plausible and an alternative strategy would be to prove directly that the Bayes problem defined by w_1 has a unique solution, up to sets of measure zero. Exceptional cases where whole sections of w_1 are zero do have multiple Bayes solutions and need special treatment. For other cases, a possible route is offered by the properties of analytic functions used by Brown, Cohen & Strawderman (1980) in proving their Theorem 3.3. However, the argument for the non-adaptive problem is not simple and its extension to our setting would be non-trivial. Furthermore, our new method of proof generalises to discrete distributions where some Bayes problems do not have unique solutions.

Lemma 1. In the setting defined at (9), for any $1 \leq v \leq 2Q$, if $\mathcal{S}_v$ is closed, then $\mathcal{S}_v = \mathcal{Q}_v$.

Proof of Lemma 1. As noted in the proof of the theorem, $\mathcal{S}_v$ is non-empty and convex. Suppose $\mathcal{S}_v$ is closed but $\mathcal{S}_v \neq \mathcal{Q}_v$. Then, there is a point $y \in \mathcal{Q}_v \backslash \mathcal{S}_v$ and, by the separating hyperplane theorem, a vector b and constant $\epsilon > 0$ such that

$$b^T y \;\leq\; b^T r - \epsilon \quad \text{for all } r \in \mathcal{S}_v.$$

Since $y \in \bar{\mathcal{S}}$, there are decision rules $\{d_i\}$ with $\lim_{i\to\infty} R(d_i) = y$. We shall prove the lemma by constructing a decision rule $\tilde{d}$ for which $b^T y \geq b^T R(\tilde{d})$, contradicting the assumptions about the point y. The rule $\tilde{d}$ is defined by the following criteria:

1. Minimise $w_1^T R(d)$

2. Subject to satisfying condition 1, minimise $w_2^T R(d)$

$\vdots$

v. Subject to satisfying conditions 1 to $v-1$, minimise $w_v^T R(d)$

$v+1$. Subject to satisfying conditions 1 to v, minimise $b^T R(d)$

$v+2$. Subject to satisfying conditions 1 to $v+1$, take the first action in list $\mathcal{L}$.

Here, $\mathcal{L}$ is the ordered list: (1) Stop, accept H_0; (2) Stop, reject H_0; (3) Continue to an analysis at information level $\mathcal{I}_1$; $\ldots$; $(M+2)$ Continue to an analysis at $\mathcal{I}_M$. Condition $v+2$ ensures that rule $\tilde{d}$ is precisely specified, up to variations on a set of measure zero. The particular ordering of actions is not significant but the labelling of possible actions will be of use later.

A rule satisfying the above criteria can be constructed by backwards induction, finding the optimal actions to take at analyses $K, K-1, \ldots, 0$ in succession. The action at analysis zero refers to the choice of m_1. Writing x_k for $(s_{m_1}, \ldots, s_{m_k}; m_1, \ldots, m_k)$, let $f_{\theta_q}(x_k)$ be the probability density of the path $(s_{m_1}, \ldots, s_{m_k})$ under fixed information levels $\mathcal{I}_{m_1}, \ldots, \mathcal{I}_{m_k}$, and denote by $\alpha(d, x_k)$ the conditional probability under rule d of taking the sequence of actions to continue sampling at stage l with next information level $\mathcal{I}_{m_{l+1}}$, $l = 0, \ldots, k-1$, as the sample path $(s_{m_1}, \ldots, s_{m_k})$ unfolds. For the loss function defined by (8) with a given value of w, we write the conditional expected loss under rule d, when $\theta = \theta_q$ and outcomes $S_{m_1} = s_{m_1}, \ldots, S_{m_k} = s_{m_k}$ have been observed, as

$$E_{\theta_q}\{L(w, \mathcal{A}, \mathcal{I}, \theta_q) \mid x_k; d\}.$$

Thus, the contribution to $w^T R(d)$ from sample paths followed up to at least analysis k can be written as

$$\sum_{(m_1,\ldots,m_k)} \int_{(s_{m_1},\ldots,s_{m_k})} \sum_{q=1}^{Q} \frac{1}{Q} f_{\theta_q}(x_k)\, \alpha(d, x_k)\, E_{\theta_q}\{L(w, \mathcal{A}, \mathcal{I}, \theta_q) \mid x_k; d\}\, ds_{m_1} \ldots ds_{m_k}. \tag{11}$$

Denote the density of the path $(s_{m_1}, \ldots, s_{m_k})$ for fixed information levels $\mathcal{I}_{m_1}, \ldots, \mathcal{I}_{m_k}$ under the assumed uniform prior distribution on θ by

$$f_\pi(x_k) \;=\; \sum_{q=1}^{Q} \frac{1}{Q} f_{\theta_q}(x_k).$$

The posterior distribution of θ given x_k is $\pi(\theta_q \,|\, x_k) = Q^{-1} f_{\theta_q}(x_k)/f_\pi(x_k)$. For conciseness, we write $\sum\!\!\!\!\!\!\int dx_k$ to denote the sum over $(m_1, \ldots, m_k)$ followed by integration over $(s_{m_1}, \ldots, s_{m_k})$. We can re-write (11) as

$$\sum\!\!\!\!\!\!\!\int f_\pi(x_k)\, \alpha(d, x_k) \sum_{q=1}^{Q} \pi(\theta_q \,|\, x_k)\, E_{\theta_q}\{L(w, \mathcal{A}, \mathcal{I}, \theta_q) \,|\, x_k; d\}\, dx_k. \tag{12}$$

In the backwards induction process, the optimal decisions at analyses $k+1, \ldots, K$ are known when analysis k is considered. Standard reasoning shows that a Bayes optimal procedure must minimise the expected conditional loss under the posterior distribution of θ. Let $E_{\theta_q}\{L(w, \mathcal{A}, \mathcal{I}, \theta_q) \,|\, x_k; j, \tilde{d}\}$ denote the conditional expectation of loss $L(w, \mathcal{A}, \mathcal{I}, \theta_q)$ when $\theta = \theta_q$, path x_k has been observed, action j is taken at analysis k and the optimal rule $\tilde{d}$ is followed at analysis $k+1$ and beyond. Following the list of $v+2$ criteria, the optimal choice when in state x_k at analysis k is the action j minimising

$$\sum_{q=1}^{Q} \pi(\theta_q \,|\, x_k)\, E_{\theta_q}\{L(w_1, \mathcal{A}, \mathcal{I}, \theta_q) \,|\, x_k; j, \tilde{d}\}.$$

If two or more actions attain this minimum, the second criterion is applied to break the tie, so we minimise

$$\sum_{q=1}^{Q} \pi(\theta_q \,|\, x_k)\, E_{\theta_q}\{L(w_2, \mathcal{A}, \mathcal{I}, \theta_q) \,|\, x_k; j, \tilde{d}\}$$

among the contending actions, and so forth. The final criterion ensures a uniquely defined decision rule. Continuing this process back to $k = 0$, where m_1 is chosen, determines $\tilde{d}$.

Since $R(\tilde{d}) \in \mathcal{S}$, the definition of w_1 and k_1 implies $w_1^T R(\tilde{d}) \geq k_1$. But $w_1^T r_1 = k_1$ for r_1 on the boundary of $\mathcal{S}$, so there are risk vectors r in $\mathcal{S}$ with $w_1^T r$ arbitrarily close to k_1. As $\tilde{d}$ minimises $w_1^T R(d)$ over $R(d) \in \mathcal{S}$, we conclude $w_1^T R(\tilde{d}) = k_1$, hence $R(\tilde{d}) \in P_1$ and $R(\tilde{d}) \in \mathcal{S}_1$. Similarly, $R(\tilde{d}) \in \mathcal{S}_1$ implies $w_2^T R(\tilde{d}) \geq k_2$ but there are risk vectors r in $\mathcal{S}_1$ with $w_2^T r$ arbitrarily close to $w_2^T r_2 = k_2$ and, as $\tilde{d}$ minimises $w_2^T R(d)$ over $R(d) \in \mathcal{S}_1$, we have $w_2^T R(\tilde{d}) = k_2$, $R(\tilde{d}) \in P_2$ and $R(\tilde{d}) \in \mathcal{S}_2$. Repeating this argument shows, ultimately, that $R(\tilde{d}) \in \mathcal{S}_v$.

If $w_1^T R(d_i) = w_1^T R(\tilde{d}), \ldots, w_v^T R(d_i) = w_v^T R(\tilde{d})$, criterion $v+1$ in the definition of $\tilde{d}$ implies $b^T R(\tilde{d}) \leq b^T R(d_i)$ and it cannot be the case that $b^T R(d_i) = b^T R(\tilde{d}) - \epsilon$. However, we need to show this situation cannot be reached in the limit as $i \to \infty$. To compare rule d_i with $\tilde{d}$, define rules d_i^k, $k = 0, \ldots, K$, where d_i^k behaves as d_i at analyses 0 to k and as

$\tilde{d}$ at analyses $k+1$ to K. By this definition, $d_i^K = d_i$ and for notational consistency we set $d_i^{-1} = \tilde{d}$. Then

$$R(d_i) - R(\tilde{d}) \;=\; R(d_i^K) - R(d_i^{-1}) \;=\; \sum_{k=0}^{K} R(d_i^k) - R(d_i^{k-1}). \tag{13}$$

The term k in this sum involves rules d_i^k and d_i^{k-1} which differ only at analysis k and both proceed optimally, as rule $\tilde{d}$, at analysis $k+1$ and beyond.

Suppose sample path $x_k = (s_{m_1}, \ldots, s_{m_k}; m_1, \ldots, m_k)$ arises and stopping does not occur before analysis k. Then, at analysis k, the conditional expectation of loss $L(w, \mathcal{A}, \mathcal{I}, \theta_q)$ under rule $\tilde{d}$, and therefore under rule d_i^{k-1}, is

$$\sum_{q=1}^{Q} \pi(\theta_q \,|\, x_k)\, E_{\theta_q}\{L(w, \mathcal{A}, \mathcal{I}, \theta_q) \,|\, x_k; \tilde{d}\}.$$

Rule d_i^k may take a different action, j, at analysis k, then proceed as $\tilde{d}$, in which case we write the conditional expected loss as

$$\sum_{q=1}^{Q} \pi(\theta_q \,|\, x_k)\, E_{\theta_q}\{L(w, \mathcal{A}, \mathcal{I}, \theta_q) \,|\, x_k; j, \tilde{d}\}.$$

Let

$$G(w, x_k, j) = \sum_{q=1}^{Q} \pi(\theta_q \,|\, x_k)\, [E_{\theta_q}\{L(w, \mathcal{A}, \mathcal{I}, \theta_q) \,|\, x_k; j, \tilde{d}\} - E_{\theta_q}\{L(w, \mathcal{A}, \mathcal{I}, \theta_q) \,|\, x_k; \tilde{d}\}]. \tag{14}$$

Define $\beta(d_i, x_k, j)$ to be the probability that rule d_i takes action j when in state x_k, indexing actions by $j \in \{1, \ldots, M+2\}$ according to the ordering $\mathcal{L}$. Then, combining (12), (13) and (14),

$$w^T R(d_i) - w^T R(\tilde{d}) \;=\; \sum_{k=0}^{K} w^T R(d_i^k) - w^T R(d_i^{k-1}) \;=$$
$$\sum_{k=0}^{K} \sum_{j=1}^{M+2} \oint f_\pi(x_k)\, \alpha(d_i, x_k)\, \beta(d_i, x_k, j)\, G(w, x_k, j)\, dx_k.$$

Define

$$A_1 = \{(x_k, j) \colon G(w_1, x_k, j) > 0\},$$

and, letting A^c denote the complement of A, define

$$A_t = \{(x_k, j) \colon G(w_t, x_k, j) > 0\} \cap A_{t-1}^c \quad t = 2, \ldots, v.$$

For pairs (x_k, j) in $(A_1 \cup \ldots \cup A_t)^c$, action j is optimal for minimising each of $w_1^T R(d), \ldots, w_t^T R(d)$, in order. For pairs (x_k, j) in A_t, action j is optimal for minimising each of $w_1^T R(d), \ldots, w_{t-1}^T R(d)$ in order but not then optimal for minimising $w_t^T R(d)$. The functions $G(w, x_k, j)$ satisfy: $G(w_1, x_k, j) > 0$ for (x_k, j) in A_1 and $G(w_1, x_k, j) = 0$ for (x_k, j) in A_1^c then, for each $t = 2, \ldots, v$, $G(w_t, x_k, j)$ can be positive or negative on $(A_1 \cup \ldots \cup A_{t-1})$, $G(w_t, x_k, j) > 0$ for (x_k, j) in A_t and $G(w_t, x_k, j) = 0$ for remaining pairs (x_k, j).

Recall $\{d_i\}$ is a sequence of decision rules with $R(d_i) \to y \in \mathcal{Q}_v \backslash \mathcal{S}_v$ where $w_t^T y = w_t^T R(\tilde{d})$ for $t = 1, \ldots, v$ and $b^T y \leq b^T r - \epsilon$ for all $r \in \mathcal{S}$. Since $w_1^T R(d_i) - w_1^T R(\tilde{d}) \to 0$,

$$\sum_{k=0}^{K} \sum_{j=1}^{M+2} \int\!\!\!\!\!\!\sum f_\pi(x_k)\, \alpha(d_i, x_k)\, \beta(d_i, x_k, j)\, G(w_1, x_k, j)\, dx_k =$$

$$\sum_{k=0}^{K} \sum_{j=1}^{M+2} \int\!\!\!\!\!\!\sum I\{(x_k, j) \in A_1\}\, f_\pi(x_k)\, \alpha(d_i, x_k)\, \beta(d_i, x_k, j)\, G(w_1, x_k, j)\, dx_k \;\to\; 0.$$

As $\int\!\!\!\!\!\sum f_\pi(x_k)$ is finite, $\alpha(d_i, x_k) \leq 1$, $\beta(d_i, x_k, j) \leq 1$, and $G(w_1, x_k, j) > 0$ on A_1, it follows that

$$\sum_{k=0}^{K} \sum_{j=1}^{M+2} \int\!\!\!\!\!\!\sum I\{(x_k, j) \in A_1\}\, f_\pi(x_k)\, \alpha(d_i, x_k)\, \beta(d_i, x_k, j)\, dx_k \;\to\; 0$$

and, since all the functions $G(w_t, x_k, j)$ and $G(b, x_k, j)$ are bounded,

$$\sum_{k=0}^{K} \sum_{j=1}^{M+2} \int\!\!\!\!\!\!\sum I\{(x_k, j) \in A_1\}\, f_\pi(x_k)\, \alpha(d_i, x_k)\, \beta(d_i, x_k, j)\, G(w_t, x_k, j)\, dx_k \;\to\; 0 \qquad (15)$$

for $t = 2, \ldots, v$, and

$$\sum_{k=0}^{K} \sum_{j=1}^{M+2} \int\!\!\!\!\!\!\sum I\{(x_k, j) \in A_1\}\, f_\pi(x_k)\, \alpha(d_i, x_k)\, \beta(d_i, x_k, j)\, G(b, x_k, j)\, dx_k \;\to\; 0.$$

At the next level, the fact that $w_2^T R(d_i) - w_2^T R(\tilde{d}) \to 0$ and (15) for $t = 2$ imply

$$\sum_{k=0}^{K} \sum_{j=1}^{M+2} \int\!\!\!\!\!\!\sum I\{(x_k, j) \in A_2\}\, f_\pi(x_k)\, \alpha(d_i, x_k)\, \beta(d_i, x_k, j)\, G(w_2, x_k, j)\, dx_k \;\to\; 0,$$

from which we deduce

$$\sum_{k=0}^{K} \sum_{j=1}^{M+2} \int\!\!\!\!\!\!\sum I\{(x_k, j) \in A_2\}\, f_\pi(x_k)\, \alpha(d_i, x_k)\, \beta(d_i, x_k, j)\, G(w_t, x_k, j)\, dx_k \;\to\; 0$$

for $t = 3, \ldots, v$, and

$$\sum_{k=0}^{K} \sum_{j=1}^{M+2} \oint I\{(x_k, j) \in A_2\} \, f_\pi(x_k) \, \alpha(d_i, x_k) \, \beta(d_i, x_k, j) \, G(b, x_k, j) \, dx_k \ \to \ 0.$$

Continuing this process up to $t = v$ shows that in the limit, there is no contribution to $b^T R(d_i) - b^T R(\tilde{d})$ from sets A_1 to A_v. For $(x_k, j) \in (A_1 \cup \ldots \cup A_v)^c$, action j is optimal for each of $w_1^T R(d), \ldots, w_v^T R(d)$ in order, and where this leaves a choice of actions, rule $\tilde{d}$ is defined to minimise the expected contribution to $b^T R(d)$, so $G(b, x_k, j) \geq 0$. In consequence,

$$b^T y - b^T R(\tilde{d}) \ = \ \lim_{i \to \infty} b^T R(d_i) - b^T R(\tilde{d}) \ =$$

$$\sum_{k=0}^{K} \sum_{j=1}^{M+2} \oint f_\pi(x_k) \, \alpha(d_i, x_k) \, \beta(d_i, x_k, j) \, G(b, x_k, j) \, dx_k \ \geq \ 0.$$

This contradicts the assumed properties of y and the lemma is proved. □

Extension of these results to the case where the finite set of $\mathcal{I}$ values is replaced by an interval $[0, \mathcal{I}_{\max}]$ appears relatively straightforward. In obtaining a Bayes rule by backwards induction, it is necessary to show that the infimum of expected loss under possible continuation points is attained, but this follows from showing it is the infimum of a continuous function over a finite interval. In expressions in Lemma 1, the sum over sequences $(m_1, \ldots, m_k)$ becomes a multiple integral; this raises technical points regarding measurability of sets of information sequences which need a careful treatment.

Proof of Corollary 1. Given that $\mathcal{S}$ is closed, arguments of Ferguson (1967, Ch. 2) can be applied directly to show the that risk vector of an admissible test, d, lies on the lower boundary of $\mathcal{S}$. This point can be separated from the origin by a supporting hyperplane which defines a Bayes problem with $w(i) \geq 0$ for all $i = 1, \ldots, 2Q$ and at least one $w(i) > 0$. The rule d is a Bayes rule for this problem.

Suppose $w(i) > 0$ only for some indices $i \in \{1, \ldots, P\}$. As there is no penalty for accepting H_0 when θ is positive, a Bayes rule must accept H_0 with probability 1 under all θ. This can be achieved stopping as early as possible, i.e., with $\mathcal{I} = \mathcal{I}_1$, and since d is admissible, it must do this. Hence, d is also a Bayes rule for problems where $w(i) > 0$ for all $i = 1, \ldots, P$ and $i = Q + 1, \ldots, 2Q$. Similar reasoning shows that if $w(i) > 0$ only for indices $i \in \{P + 1, \ldots, Q\}$, then d is a Bayes rule for problems where $w(i) > 0$ for all $i = P + 1, \ldots, Q$ and $i = Q + 1 \ldots, 2Q$.

Now suppose $w(i) > 0$ only for some indices $i \in \{Q+1, \ldots, 2Q\}$. Since these $w(i)$ imply a cost of sampling but no costs for wrong decisions, d must stop at $\mathcal{I} = \mathcal{I}_1$ with probability 1. As d is admissible, it is also admissible in the class of decision rules for the fixed sample problem with data $S_1 \sim N(\theta\mathcal{I}_1, \mathcal{I}_1)$, which has risk set $\mathcal{S}' = \mathcal{S} \cap \mathcal{T}$ where

$$\mathcal{T} = \{R(d) \colon R(i, d) = \mathcal{I}_1, i = Q+1, \ldots, 2Q\}.$$

Standard arguments show this is a closed convex set, $R(d)$ is on the boundary and lies on a supporting hyperplane within $\mathcal{T}$ which defines a Bayes problem for the fixed sample test with $w(i) > 0$ for at least one $i \in \{1, \ldots, Q\}$. It follows that d is Bayes for the sequential problem which combines $w(i)$, $i = 1 \ldots, Q$, from this fixed sample problem and $w(i) = H$ for all $i \in \{Q+1, \ldots, 2Q\}$, where H is sufficiently large that stopping always occurs at $\mathcal{I} = \mathcal{I}_1$. This case demonstrates that the converse to the last statement of the corollary does not hold since there are Bayes rules for problems in which conditions 1 to 3 hold that fall in the set $\mathcal{D}_{NS}$. □

APPENDIX 2

The backwards induction algorithm

The Bayes decision problem of Section 4 is solved by a backwards induction algorithm. The prior on θ comprises point probability masses at $\theta = 0$ and δ, which we write as $\pi_1(0) = 1/3$ and $\pi_1(\delta) = 1/3$, plus a density $\pi_2(\theta) = f(\theta)/3$ for $\theta \in \Re$, where $f(\theta)$ is the density of a $N(\delta, \delta^2/4)$ random variable. We must choose between decisions $\mathcal{A}_0 =$ "Accept H_0" and $\mathcal{A}_1 =$ "Reject H_0" with cost function $\mathcal{C}(\mathcal{A}_1, 0) = c_1$, $\mathcal{C}(\mathcal{A}_0, \delta) = c_2$ and $\mathcal{C}(\mathcal{A}, \theta) = 0$ otherwise. The sampling cost is one per unit of observed information under the continuous part of the prior distribution; since this assigns probability zero to $\theta = 0$ and $\theta = \delta$, we can simply say sampling cost is one per unit of information at all $\theta \notin \{0, \delta\}$ and 0 otherwise.

Up to K analyses are allowed at an increasing sequence of information levels from the set $\{\mathcal{I}_1, \ldots, \mathcal{I}_M\}$. Denote the information level at analysis k by $\mathcal{I}_{m_k}$, the test statistic by S_{m_k}, and the posterior distribution for θ by $p^{(k)}(\theta | m_k, S_{m_k})$, comprising point masses $p_1^{(k)}(0 | m_k, S_{m_k})$ and $p_1^{(k)}(\delta | m_k, S_{m_k})$ plus a continuous density $p_2^{(k)}(\theta | m_k, S_{m_k})$.

The minimum additional expected loss incurred by stopping at analysis k with information $\mathcal{I}_{m_k}$ and statistic S_{m_k} is

$$\zeta^{(k)}(m_k, S_{m_k}) \;=\; \min\{c_1\, p_1^{(k)}(0|m_k, S_{m_k}),\; c_2\, p_1^{(k)}(\delta|m_k, S_{m_k})\}.$$

For analyses $k = 1, \ldots, K-1$, $m_k \in \{k, \ldots, M-K+k\}$ and $m_{k+1} \in \{m_k+1, \ldots, M-K+k+1\}$, define $\xi^{(k)}(m_k, S_{m_k}, m_{k+1})$ to be the expected additional cost when the observed statistic is S_{m_k} of continuing to analysis $k+1$ at information level $\mathcal{I}_{m_{k+1}}$ and proceeding optimally thereafter. The minimum additional expected cost given m_k and S_{m_k} is thus

$$\eta^{(k)}(m_k, S_{m_k}) \;=\; \min\,[\zeta^{(k)}(m_k, S_{m_k}),\; \min_{m_{k+1}} \{\xi^{(k)}(m_k, S_{m_k}, m_{k+1})\}].$$

Denoting by $F^{(k+1)}(S_{m_{k+1}}|m_k, S_{m_k}, m_{k+1})$ the conditional cumulative distribution function of $S_{m_{k+1}}$ given m_k, S_{m_k} and m_{k+1}, we have

$$\begin{aligned}\xi^{(K-1)}(m_{K-1}, S_{m_{K-1}}, m_K) \;=\; (\mathcal{I}_{m_K} - \mathcal{I}_{m_{K-1}}) \int_{-\infty}^{\infty} 1\,.\,p_2^{(K-1)}(\theta|m_{K-1}, S_{m_{K-1}})\\ +\int_{-\infty}^{\infty} \zeta^{(K)}(m_K, S_{m_K})\, dF^{(K)}(S_{m_K}|m_{K-1}, S_{m_{K-1}}, m_K)\end{aligned}$$

and, for $k = 1, \ldots, K-2$,

$$\begin{aligned}\xi^{(k)}(m_k, S_{m_k}, m_{k+1}) \;=\; (\mathcal{I}_{m_{k+1}} - \mathcal{I}_{m_k}) \int_{-\infty}^{\infty} 1\,.\,p_2^{(k)}(\theta|m_k, S_{m_k})\\ +\int_{-\infty}^{\infty} \eta^{(k+1)}(m_{k+1}, S_{m_{k+1}})\, dF^{(k+1)}(S_{m_{k+1}}|m_k, S_{m_k}, m_{k+1}).\end{aligned}$$

Proceeding by backwards induction through $k = K-1, \ldots, 2$ and all permissible pairs m_k and m_{k+1}, the above expressions for $\xi^{(k)}(m_k, S_{m_k}, m_{k+1})$ are calculated numerically. In the case $k = K-1$, we apply knowledge of $\zeta^{(K)}(m_K, S_{m_K})$ and for $k \le K-2$ we use values for $\eta^{(k+1)}(m_{k+1}, S_{m_{k+1}})$ already computed on a grid of values of $S_{m_{k+1}}$. The range of values for S_{m_k} is divided into intervals within which the minimum additional expected cost, $\eta^{(k)}(m_k, S_{m_k})$, is attained by just one of the actions: stop now and accept H_0, stop now and reject H_0, continue to information level $\mathcal{I}_{m_k+1}$, …, continue to information level $\mathcal{I}_{M-K+k+1}$. Then, within each interval, $\eta^{(k)}(m_k, S_{m_k})$ is calculated at a grid of points suitable for numerical integration over the distribution of S_{m_k}. Jennison & Turnbull (2000, Ch. 19) provide further details of methods for recursive numerical integration to derive and evaluate group sequential tests.

REFERENCES

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

Bauer, P. & Köhne, K. (1994). Evaluation of experiments with adaptive interim analyses. *Biometrics* **50**, 1029–41.

Brannath, W., Posch, M. & Bauer, P. (2002). Recursive combination tests. *J. Amer. Statist. Assoc.* **97**, 236–44.

Brittain, E. H. & Bailey, K. R. (1993). Optimization of multistage testing times and critical values in clinical trials. *Biometrics* **49**, 763–72.

Brown, L. D., Cohen, A. & Strawderman, W. E. (1980). Complete classes for sequential tests of hypotheses. *Ann. Statist.* **8**, 377–98.

Chang, M. N. (1996). Optimal designs for group sequential clinical trials. *Commun. Statist. A* **25**, 361–79.

Cox, D. R. & Hinkley, D. V. (1974). *Theoretical Statistics*. London: Chapman & Hall.

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

Denne, J. S. (2001). Sample size recalculation using conditional power. *Statistics in Medicine* **20**, 2645–60.

Denne, J. S. & Jennison, C. (2000). A group sequential t-test with updating of sample size. *Biometrika* **87**, 125–34.

Eales, J. D. & Jennison, C. (1992). An improved method for deriving optimal one-sided group sequential tests. *Biometrika* **79**, 13–24.

Ferguson, T. S. (1967). *Mathematical Statistics: A Decision Theoretic Approach.* New York: Academic Press.

Fisher, L. D. (1998). Self-designing clinical trials. *Statistics in Medicine* **17**, 1551–62.

Fisher, R. A. (1932). *Statistical Methods for Research Workers*. London: Oliver & Boyd.

Jennison, C. & Turnbull, B. W. (1989). Interim analyses: the repeated confidence interval approach (with discussion). *J. Roy. Statist. Soc. B* **51**, 305–61.

Jennison, C. & Turnbull, B. W. (1997). Group sequential analysis incorporating covariate information. *J. Amer. Statist. Assoc.* **92**, 1330–41.

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

Jennison, C. & Turnbull, B. W. (2003). Mid-course sample size modification in clinical trials based on the observed treatment effect. *Statistics in Medicine* **22**, 971–93.

Jennison, C. & Turnbull, B. W. (2004). Efficient group sequential designs when there are several effect sizes under consideration. Submitted.

Lehmacher, W. & Wassmer, G. (1999). Adaptive sample size calculation in group sequential trials. *Biometrics* **55**, 1286–90.

Li, G., Shih, W. J., Xie, T. & Lu. J. (2002). A sample size adjustment procedure for clinical trials based on conditional power. *Biostatistics* **3**, 277–87.

Mehta, C. R. & Tsiatis, A. A. (2001). Flexible sample size considerations using information-based interim monitoring. *Drug Information J.* **35**, 1095–112.

Müller, H-H. & Schäfer, H. (2001). Adaptive group sequential designs for clinical trials: Combining the advantages of adaptive and of classical group sequential procedures. *Biometrics* **57**, 886–91.

Nelder, J. A. & Mead, R. (1965). A simplex method for function minimization. *Computer J.* **7**, 308–13.

Posch, M., Bauer, P. & Brannath, W. (2003). Issues in designing flexible trials. *Statistics in Medicine* **22**, 953–69.

Proschan, M. A. & Hunsberger, S. A. (1995). Designed extension of studies based on conditional power. *Biometrics* **51**, 1315–24.

Schäfer, H. & Müller, H-H. (2004). Construction of group sequential designs in clinical trials on the basis of detectable treatment differences. *Statistics in Medicine* **23**, 1413–24.

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

Shen, Y. & Fisher, L. (1999). Statistical inference for self-designing designing clinical trials with a one-sided hypothesis. *Biometrics* **55**, 190–7.

Thach, C. & Fisher, L. D. (2002). Self-designing two-stage trials to minimize expected costs. *Biometrics* **58**, 432–8.

Tsiatis, A. A. & Mehta, C. (2003). On the inefficiency of the adaptive design for monitoring clinical trials. *Biometrika* **90**, 367–78.

Wittes, J. & Brittain, E. (1990). The role of internal pilot studies in increasing efficiency of clinical trials. *Statistics in Medicine* **9**, 65–72.