Adaptive and non-adaptive group sequential tests

CHRISTOPHER JENNISON
Department of Mathematical Sciences, University of Bath, Bath, BA2 7AY, U. K.
cj@maths.bath.ac.uk

and BRUCE W. TURNBULL
Department of Statistical Science, Cornell University, Ithaca, New York, 14853-3801, U. S. A.
turnbull@orie.cornell.edu

SUMMARY

Methods have been proposed to re-design a clinical trial at an interim stage in order to increase power. 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 schemes 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 that 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.

Some key words: Adaptive re-design; Admissibility; Clinical trial; Conditional power; Complete class theorem; Efficiency; Group sequential test; 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 shall consider changing the alternative at which a specified power is to be attained. This should not be confused with 're-estimating' the sample size needed to meet a fixed power requirement as more is learnt about a nuisance parameter that governs 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 for changing the treatment definition or the primary response, switching between tests for superiority and non-inferiority, or response-dependent randomisation to reduce 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 the withdrawal of a rival treatment so that smaller effect sizes for the new treatment are now of interest and power $1-\beta$ is desired at an alternative $\theta = \delta'$ where $0 < \delta' < \delta$. A similar conclusion might arise from information internal to the study but on a secondary endpoint; for example, good safety results combined with positive efficacy at a level below δ could justify use of the new treatment.

There may, instead, be completely internal reasons to re-design a study, in view of interim data on the primary endpoint. It could be deemed appropriate to increase the remaining sample size 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 expected or hoped for, it still represents a worthwhile improvement and they would like to extend the study to ensure that high power is 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 a type I error probability of α if sample size is changed 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 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, Proschan & Hunsberger (1995), Cui et al. (1999), 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 as 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 maintains the overall type I error rate.

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 principle of sufficiency (Cox & Hinkley, 1974, §2.3) raises questions about both the statistical efficiency of the experimental designs and the credibility of reported results. In an analysis of selected examples, Jennison & Turnbull (2003) show that adaptive sampling rules can be much less efficient than 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 more 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 et al. (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.

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 et al. (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, as estimated from interim data. Our objectives in this paper are to illustrate and critically appraise methods of adaptive re-design for power criteria, and in particular, to answer the following questions.

> Does use of non-sufficient statistics in adaptive designs automatically imply inefficiency?
>
> 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?

We present theoretical results which answer the first question in the affirmative and computations showing that the efficiency gains of the best possible pre-planned adaptive designs are very small. 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 recommend that 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 for preventing 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 rate for patients receiving the drug was only 25% lower. The investigators recognised that a 25% reduction in incidence was still clinically significant, but 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 for mid-study adaptation of sample size which preserves type I error. We describe their proposal in the context of a general group sequential test of a treatment effect θ. Suppose that 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 that 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

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

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 an 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 that each information increment is increased by a factor γ so that, for $k = j+1, \ldots, K$,

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

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 in the newly defined S_k remain independent, $N(0,\, \mathcal{I}_k - \mathcal{I}_{k-1})$ and applying the boundary (2) to these statistics preserves the type I error

probability α exactly. Means of the new increments are multiplied by $\gamma^{1/2}$, so if $\gamma > 1$ this increases power for $\theta > 0$. Cui et al. (1999) suggest that 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

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

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

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

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

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

Suppose external information becomes available at the second analysis, leading 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 that modification could be made without prejudicing the type I error rate. However, it would be difficult to prove that the data revealed at interim analyses had played no part in the decision to re-design. We consider modifications following 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 $1-\beta = 0.9$, but truncate γ to lie in the range 1 to 6 so that sample size is never reduced and the maximum total information is increased by at most a factor of 4. Fig. 1(a) shows that 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 because of 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 $\mathcal{I}_5 = \mathcal{I}_{max} = 3.78\,\mathcal{I}_f$. This choice ensures that the power of the non-adaptive test is everywhere as high as that of the adaptive test, as seen in Fig. 1(a), and the expected information curves of the two tests are of a similar shape. Fig. 1(b) shows the expected information on termination as a function of θ/δ for these two tests; the vertical axis is in units of $\mathcal{I}_f$. Together, Figs 1(a) and 1(b) show that the non-adaptive test dominates the adaptive test in terms of 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 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

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

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

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

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

The plot in Fig. 1(c) of the efficiency ratio between the two tests in our example quantifies the cost of delay in learning the study's objective as an efficiency loss 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 Fig. 1(a).

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 that 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. This time we choose γ so that conditional power given the observed S_2, if θ is equal to $\widehat{\theta}_2$, is $1 - \beta = 0.9$. A decrease in sample size is allowed if $\widehat{\theta}_2$ is sufficiently high to imply that $\gamma < 1$. As in Example 1, γ is truncated to 6 to restrict the maximum information level to at most 4 times its original value; this has the effect that conditional power under $\theta = \widehat{\theta}_2$ is equal to 0.9 for $\widehat{\theta}_2 \geq 0.49\,\delta$ but lower for smaller values of $\widehat{\theta}_2$.

The power curves in Fig. 2(a) show that this adaptation has been effective in increasing power, with a rise from 0.37 to 0.68 at $\theta = \delta/2$. 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 that the above adaptive rule is in accord with such considerations and that the power curve in Fig. 2(a) 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 $\mathcal{I}_5 = \mathcal{I}_{max} = 3.21\,\mathcal{I}_f$. Fig. 2(a) shows that the power of this non-adaptive test exceeds that of the adaptive test at all values of θ and by a substantial margin at the highest θ values.

Fig. 2(b) shows that the non-adaptive test has considerably lower expected information 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 plot of the efficiency ratio in Fig. 2(c) 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 our examples is that adaptive methods do exist for making mid-study design modifications to meet changes in 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 occur.

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. Additional loss of efficiency in Example 2 can be attributed to over-reliance on the highly variable interim estimator $\widehat{\theta}_2$. 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.

We have carried out many more comparisons of adaptive designs and matched non-adaptive error-spending tests with similar qualitative conclusions. In general, allowing a greater increase in maximum sample size leads to higher inefficiency. The examples of §§ 2.2 and 2.3 follow the recommendation of many authors to base sample size revision on conditional power; the initial tests are allowed to stop early to accept H_0 but we chose the stopping rule carefully to reduce the risk of stopping to accept H_0 when θ is in the range $\delta/2$ to δ and the adaptive test could later be required to attain higher power. 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; the two-stage adaptive design studied by Jennison & Turnbull (2003) has a 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 testing H_0: $\theta \leq 0$ against $\theta > 0$. Suppose there are M analysis times to choose from with associated information levels $\mathcal{I}_1, \ldots, \mathcal{I}_M$, 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 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 $M \gg K$;

non-adaptive group sequential tests are covered by the case $M = K$.

Denote the indices of the information levels arising in a particular realisation of the experiment by $m_1, m_2, \ldots$, so that 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 decision rule fixes $m_1 \in \{1, \ldots, M - K + 1\}$, and then for each k and observed data X_k it chooses an action from the following 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. In deriving theoretical results, we allow randomised rules which correspond to probability distributions on the set of deterministic rules. We denote the set of all randomised and nonrandomised rules by $\mathcal{D}$.

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

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

the type II error function

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

and the expected information function $R_3(\theta, d) = E_\theta(\mathcal{I})$. We assume that the preferred decision is to reject H_0 whenever $\theta > 0$ but R_1 and R_2 could be modified to change this threshold. Although a stronger result appears provable, we avoid technical difficulties by considering risk on a finite set $\Theta = \{\theta_1, \ldots, \theta_Q\}$, where $\theta_1 < \ldots < \theta_P \leq 0 < \theta_{P+1} < \ldots < \theta_Q$. This restriction has little practical impact as one can take, say, ten million points over the range of θ values of interest.

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

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

A decision rule $d \in \mathcal{D}$ is said to be inadmissible if there is a rule d' with $R(i, d') \leq R(i, d)$ for all $i = 1, \ldots, 2Q$ and $R(i, d') < R(i, d)$ 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

$$\sum_{q=1}^{Q} \pi_q \sum_{j=1}^{3} c_{qj} R_j(\theta_q, d), \tag{4}$$

where c_{q1} is the cost of rejecting H_0, c_{q2} is the cost of accepting H_0 and c_{q3} is 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

$$w^{\mathrm{T}} R(d) \; = \; \sum_{i=1}^{2Q} w(i) R(i, d), \tag{5}$$

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 admissible rules as Bayes rules, the risk set $\mathcal{S} = \{R(d); d \in \mathcal{D}\}$ plays a central role. The proofs of Theorem 1 and Corollary 1 below are given in Appendix 1.

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:*

(i) $w(i) > 0$ *for some* $i \leq P$,

(ii) $w(i) > 0$ *for some* $P + 1 \leq i \leq Q$,

(iii) $w(i) > 0$ *for some* $i \geq Q + 1$.

Let $\mathcal{D}_{NS}$ denote the set of 'non-sequential' decision rules which terminate at $\mathcal{I}_1$ with probability 1 or terminate at $\mathcal{I}_M$ 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. □

Ferguson (1967, §§ 7.1 and 7.2) and Brown et al. (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 θ value. 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 lower expected information function, as was very nearly the case in Example 1 of § 2. Again in the non-adaptive setting, Chang (1996) considers a risk vector 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 prove that the risk set is closed.

Corollary 1 with $K < M$ characterises admissible adaptive designs. If a design is properly sequential, and so produces a non-degenerate distribution of sample sizes, to be admissible it must be a Bayes rule for a problem satisfying conditions (i) to (iii). Since a Bayes problem has a solution based on sufficient statistics, this establishes the principle that a sequential test should be defined in terms of sufficient statistics. Adaptive designs using non-sufficient statistics, as required in the flexible adaptive approach, are dominated by admissible adaptive designs based on sufficient statistics.

If the group size multiplier γ in Cui et al.'s (1999) method is a one-to-one function of S_j, the adaptive rule defined through non-sufficient statistics (3) can be re-expressed in terms of sufficient statistics. However, many proposals for adaptive designs truncate γ to a maximum value, as in the examples of § 2, and, with the same sequence of future information levels arising for an interval of S_j values, the rule cannot be re-expressed in terms of the sufficient statistics. Even when an adaptive design is a function of sufficient statistics, it is admissible only if its sampling rule, stopping rule and terminal decision rule coincide with those of a Bayes optimal design. In § 4 we examine Bayes optimal designs and note qualitative differences between their sampling rules and those of adaptive designs based on fixed conditional power at a pre-specified or estimated effect size.

The case $K = M$ covers non-adaptive tests and Corollary 1 tells us that non-adaptive group sequential tests with stopping rules or decision rules based on non-sufficient statistics are dominated by non-adaptive Bayes optimal designs defined in terms of sufficient statistics. The variance-spending tests of Shen & Fisher (1999) fall in this category since the sequence of information levels is fixed and it is the weights for each group of observations that are chosen adaptively; unequal weighting implies departure from a Bayes rule, and hence the variance-spending test is inadmissible. The papers by Falissard & Lellouch (1991, 1992, 1993) propose tests which reject the null hypothesis if a boundary is crossed at a set number of successive interim analyses. In discussion reported in the first of these papers, P. Armitage notes that this procedure uses a non-sufficient statistic and T. Louis suggests it might be possible to prove this will imply the test can be dominated, as we have now done; the second paper contains references to earlier proposals of a similar nature.

If $K < M$, increasing the number of analyses above K adds to the available options. Thus, for a Bayes problem, the optimal adaptive test with $K < M$ analyses does 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 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 values of θ in the alternative hypothesis. Our result goes further in showing that error probability and expected information can be maintained or reduced at all values of θ in a composite null hypothesis, whereas Tsiatis & Mehta (2003) consider only a simple null hypothesis and the expected information at this value of θ may increase. Also, the test constructed by Tsiatis & Mehta (2003) is not necessarily admissible.

Calculations for optimal group sequential tests in Eales & Jennison (1992) show that most of the achievable reductions in expected sample size are obtained using 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 the best adaptive tests may have when the maximum number of analyses is restricted to $K = 2$ or 3. Adaptivity extends the class of group 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 sub-optimal tests using non-sufficient statistics are competitive with the best non-adaptive tests. We shall explore the extent of these possible gains from adaptivity in § 4, where we solve Bayes decision problems to find adaptive and non-adaptive designs meeting specific optimality criteria.

4 Computing optimal adaptive designs

The theory of § 3 shows the equivalence between the class of admissible adaptive tests and the set of Bayes optimal adaptive designs. Eales & Jennison (1992) and Barber & Jennison (2002) have exploited this 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 $\theta = \delta$. Designs 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 δ and a desire to stop particularly early if this is the case. We have taken $M = 50$ and $\mathcal{I}_1, \ldots, \mathcal{I}_{50}$ equally spaced between 0 and $R\,\mathcal{I}_f$, as sensitivity analyses indicate there is no significant change if M is increased further. Although these results are for one particular problem, we have reached similar conclusions with other choices of α and β and a variety of optimisation criteria.

To find the optimal tests, we 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 states are indexed by the analysis number k and the index m_k of the information level at which the analysis occurs; further details 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 value of

$$\int E_\theta(\mathcal{I})\, f(\theta)\, d\theta \tag{6}$$

achieved by 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 the 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 (6) for (a) a non-adaptive test with K analyses at information levels $\mathcal{I}_1, \ldots, \mathcal{I}_K$ placed optimally between 0 and $R\,\mathcal{I}_f$, as proposed previously by Eales & Jennison (1992) and Brittain & Bailey (1993), 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 $\mathcal{I}_f$. 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-spaced information levels is within 2% of $\mathcal{I}_f$ of the optimal adaptive test's average $E_\theta(\mathcal{I})$ in most

cases: exceptions when $K = 2$ and $R \geq 1.2$ or $K = 3$ and $R = 1.3$ occur because these values of R are unnecessarily high for these values of K. Optimising $\mathcal{I}_1, \ldots, \mathcal{I}_K$ subject to $\mathcal{I}_K \leq R\mathcal{I}_f$ gives the results in the middle column of 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. These comparisons are much tighter than the conclusions drawn by Tsiatis & Mehta (2003) that an adaptive test with 2 or 3 analyses can be matched by a non-adaptive test with 10 analyses since this is close to continuous monitoring.

The small advantages of adaptivity are in keeping with results of Posch et al. (2003) for $K = 2$. Our results are more far-reaching in that we optimise over completely general sampling rules and stopping boundaries, and consider higher values of K. Even if the limited benefits of adaptive designs are deemed worthwhile, it may be preferable administratively to achieve these in a non-adaptive design with one or two additional analyses. The gap between the best adaptive and best non-adaptive tests is large enough that an adaptive test based on non-sufficient statistics might not be dominated by a non-adaptive test with the same K; however, the margin for error is small.

Sampling rules for optimal adaptive tests follow a consistent pattern. At analysis k, increments in information are smaller when S_{m_k} is close to either stopping boundary and larger when S_{m_k} is in 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 § 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.

We can now re-consider the examples of § 2 in the light of the theory of § 3 and the computational results of this section. In both examples, the adaptive test is inadmissible since it is not a function of sufficient statistics. Theory implies that there is a superior adaptive test, not necessarily a non-adaptive one. Since our computations show that the best non-adaptive tests are almost as efficient as their adaptive counterparts, it should not be a surprise that matched non-adaptive tests can out-perform the inadmissible adaptive tests. We have taken error-spending tests as our matched tests to show there are efficient 'off the shelf' options, but we could have stayed closer to the theoretical reasoning and used designs optimised for appropriately selected Bayes decision problems.

5 Discussion

Non-adaptive group sequential tests are well studied and optimal tests have been derived for a variety of 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 $3\delta/2$. Jennison & Turnbull (2005) consider criteria $\{E_0(\mathcal{I}) + E_\delta(\mathcal{I}) + E_{H\delta}(\mathcal{I})\}/3$ with $H = 2$, 3 and 4 and demonstrate the effectiveness of the ρ-family as long as $\mathcal{I}_1$ is carefully chosen. These error-spending tests are easily implemented and provide flexibility to deal with unpredictable information sequences.

Incorporating adaptivity in pre-planned group sequential designs, as proposed by Schmitz (1993), produces a small benefit. However, similar improvements are often achieved by non-adaptive designs with one extra analysis, avoiding the administrative complications of a pre-planned adaptive design.

Using adaptive methods in an unplanned manner offers flexibility to the organisers of a study but, since the sufficiency principle is contravened, there is an efficiency cost. One argument for flexible adaptive designs is that they allow investigators to choose a study's power curve in response to early estimates of the effect size, θ. This may be appealing when there is uncertainty about the likely effect size and optimistic estimates 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 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 group sequential test that achieves this while giving low expected sample size under larger effects. 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 with a special sequence of information levels seen in § 2, or ρ-family tests with an optimised choice of $\mathcal{I}_1$ investigated by Jennison & Turnbull (2005).

There is a substantial literature on sample size modification in response to estimates of a nuisance parameter, such as the variance of a normal response, which determines the sample size needed to achieve power at a specified effect size. In the 'information-based monitoring' approach described by Mehta & Tsiatis (2001), the maximum information for an error-spending test is fixed but the target sample size is adjusted in the light

of new estimates of the parameter governing the relationship between sample size and information. This process has minimal effect on the type I error rate. In principle, sample size adjustment for a nuisance parameter can be combined with modifications to increase power in view of the observed treatment effect; further study of such schemes would be helpful to assess the effect of any interplay between the two types of update.

A key role that remains for flexible adaptive methods is to help investigators respond to unexpected external events. As Müller & Schäfer (2001) and Posch et al. (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. However, if the unexpected occurs, adaptive methods can be applied. The approach based on maintaining conditional type I error probability put forward by Denne (2001) and 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.

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 that such problems will cease to arise.

ACKNOWLEDGEMENT

This research was supported in part by a grant from the U. S. National Institutes of Health.

APPENDIX 1

Proofs

Proof of Theorem 1. We refer to Chapter 2 of Ferguson (1967) for proofs of the supporting hyperplane and separating hyperplane theorems used below, as well as for background to complete class theorems which show, broadly speaking, that admissible rules are Bayes and vice versa.

We restrict attention to cases with $\pi_q = 1/Q$, $q = 1, \ldots, Q$. Then the Bayes risk $w^{\mathrm{T}}R(d)$ in (5) 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 (4) can be written in the form (5), this does not reduce the class of problems considered. It is helpful to treat the Bayes risk as the expectation, under the prior $\pi_q = 1/Q$, $q = 1, \ldots, Q$, of the loss function $L(w, \mathcal{A}, \mathcal{I}, \theta)$, where

$$L(w, \mathcal{A}, \mathcal{I}, \theta_q) \;=\; \begin{cases} Q\,w(q)\,I(\mathcal{A} = \text{Reject } H_0) \;+\; Q\,w(Q+q)\,\mathcal{I}, & q \le P, \\ Q\,w(q)\,I(\mathcal{A} = \text{Accept } H_0) \;+\; Q\,w(Q+q)\,\mathcal{I}, & q > P. \end{cases} \qquad \text{(A1)}$$

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

We shall prove that $\mathcal{S}$ is closed by taking a general point r_1 on the boundary of $\mathcal{S}$ and showing that it is in $\mathcal{S}$. We use the supporting hyperplane at r_1 to define a Bayes decision problem, and a decision rule solving this Bayes problem can be found by backwards induction. 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 r_1. However, if intersection is at a set of points, further work is required to prove that r_1 is the risk vector of a Bayes rule, and therefore in $\mathcal{S}$. The full proof is by induction. We start by outlining the first two stages 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 $\bar{\mathcal{S}}$. By the supporting hyperplane theorem, there is a hyperplane $P_1 = \{r : w_1^{\mathrm{T}} r = k_1\}$ for which $w_1^{\mathrm{T}} r_1 = k_1$ and $w_1^{\mathrm{T}} r \geq k_1$ for all $r \in \mathcal{S}$. Let $\mathcal{S}_1 = P_1 \cap \mathcal{S}$ and $\mathcal{Q}_1 = P_1 \cap \bar{\mathcal{S}}$. If some elements of w_1 are negative, minimising $w_1^{\mathrm{T}} R(d)$ is an unusual Bayes decision problem, but that is unimportant. A decision rule, d_1 say, minimising $w_1^{\mathrm{T}} R(d)$ can be constructed by backwards induction. Since $w_1^{\mathrm{T}} r \geq k_1$ for all $r \in \mathcal{S}$, we know that $w_1^{\mathrm{T}} R(d_1) \geq k_1$. However, there are decision rules with risk vectors approaching r_1, and therefore $w_1^{\mathrm{T}} R(d_1) \leq w_1^{\mathrm{T}} r_1 = k_1$. Hence $w_1^{\mathrm{T}} R(d_1) = k_1$ and $R(d_1) \in \mathcal{S}_1$, showing that $\mathcal{S}_1$ is non-empty. Considering linear combinations of decision rules shows 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 to consider further is where $\mathcal{S}_1$ is a strict subset of $\mathcal{Q}_1$ and $r_1 \in \mathcal{Q}_1 \backslash \mathcal{S}_1$. We aim to show that (i) $\mathcal{S}_1$ is closed, and (ii) $\mathcal{S}_1 = \mathcal{Q}_1$, from which it follows that $r_1 \in \mathcal{S}_1$. Lemma A1 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 $\Re^{2Q}$.

If we take an arbitrary r_2 on the boundary of $\mathcal{S}_1$, there is a supporting hyperplane within P_1, $P_2 = \{r : w_2^{\mathrm{T}} r = k_2\} \cap P_1$, for which $w_2^{\mathrm{T}} r_2 = k_2$ and $w_2^{\mathrm{T}} r \geq k_2$ for all $r \in \mathcal{S}_1$. Define $\mathcal{S}_2 = P_2 \cap \mathcal{S}$ and $\mathcal{Q}_2 = P_2 \cap \bar{\mathcal{S}}$. Points in $\mathcal{S}_2$ are risk vectors of decision rules solving the following problem: first, minimise $w_1^{\mathrm{T}} R(d)$; then, as a secondary criterion, minimise $w_2^{\mathrm{T}} R(d)$ among rules minimising $w_1^{\mathrm{T}} R(d)$. Such a rule, d_2 say, can be constructed by backwards induction and, following earlier reasoning, must satisfy $w_1^{\mathrm{T}} R(d_2) = k_1$ and

$w_2^{\mathrm{T}} R(d_2) = k_2$. Thus, $\mathcal{S}_2$ is nonempty and, by the usual argument, convex. We now wish to show that (i) $\mathcal{S}_2$ is closed and (ii) $\mathcal{S}_2 = \mathcal{Q}_2$, to deduce $r_2 \in \mathcal{S}_2$.

Further iterations of this process lead eventually to a nonempty, convex $\mathcal{S}_u$ of dimension zero. As this is a singleton set, it is closed and thus (i) holds. It is still necessary to show that (ii) holds at this level and work back to deduce that $\mathcal{S}_1$ 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 $\mathcal{S}_{v-1}$,

$P_v = \{r : w_v^{\mathrm{T}} r = k_v\} \cap P_{v-1}$ is a supporting hyperplane to $\mathcal{S}_{v-1}$ within P_{v-1} at the point r_v, for which $w_v^{\mathrm{T}} r_v = k_v$ and $w_v^{\mathrm{T}} r \geq k_v$ for all $r \in \mathcal{S}_{v-1}$,

$\mathcal{S}_v = P_v \cap \mathcal{S}$ is non-empty and $\mathcal{Q}_v = P_v \cap \bar{\mathcal{S}}$. (A2)

Note that arbitrary choice of the boundary point r_v is allowed in (A2). 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. Each $\mathcal{S}_v$ is non-empty since backwards induction can be used to construct a decision rule d_v minimising $w_1^{\mathrm{T}} R(d)$ first, then minimising $w_2^{\mathrm{T}} R(d)$ among rules that minimise $w_1^{\mathrm{T}} R(d)$, and so forth. Arguments outlined above and given more fully in the proof of Lemma A1 show that $w_1^{\mathrm{T}} R(d_v) = w_1^{\mathrm{T}} r_1 = k_1$, etc., so that $R(d_v)$ 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 A1 states that 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 as follows:

if the dimension of $P_v \leq h$, then $\mathcal{S}_v$ is closed. (A3)

This hypothesis is satisfied for $h = 0$ since $\mathcal{S}_v$ is a singleton set. Inductively, suppose that (A3) 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 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 A1, $\mathcal{S}_v = \mathcal{Q}_v$. Now, r_v is in $\bar{\mathcal{S}}$ and P_v, and 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 is

a general boundary point of $\mathcal{S}_{v-1}$, we see that $\mathcal{S}_{v-1}$ is closed and this establishes (A3) for $h \leq \tilde{h} + 1$. With $v = 0$ and $h = 2Q$ in (A3) we see 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 has multiple solutions. 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 et al. (1980) in proving their Theorem 3.3. Extending this argument to our setting would be nontrivial. Moreover, our proof generalises to discrete distributions where Bayes problems may not have unique solutions.

Lemma A1. *In the setting defined at (A2), for any* $1 \leq v \leq 2Q$*, if* $\mathcal{S}_v$ *is closed, then* $\mathcal{S}_v = \mathcal{Q}_v$.

Proof of Lemma A1. As noted in the proof of the theorem, $\mathcal{S}_v$ is nonempty and convex. Suppose that $\mathcal{S}_v$ 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 $\epsilon > 0$ such that $b^{\mathrm{T}} y \leq b^{\mathrm{T}} r - \epsilon$ 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 prove the lemma by constructing a decision rule $\tilde{d}$ for which $R(\tilde{d}) \in \mathcal{S}_v$ and $b^{\mathrm{T}} R(\tilde{d}) \leq b^{\mathrm{T}} y$, contradicting the assumptions about y. The rule $\tilde{d}$ is defined by the following criteria:

1. Minimise $w_1^{\mathrm{T}} R(d)$.

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

⋮

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

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

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

Here, $\mathcal{L}$ is the following ordered list: (1) Stop, accept H_0; (2) Stop, reject H_0; (3) Continue to an analysis at information level $\mathcal{I}_1$; ... ; ($M + 2$) 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 will be of use later.

A rule satisfying the above criteria can be constructed by finding the optimal actions to be taken 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 for $\theta = \theta_q$ 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 (A1) 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 $w^{\mathrm{T}} R(d)$ from sample paths followed up to at least analysis k can be written as

$$\sum_{(m_1,\ldots,m_k)} \int_{(s_{m_1},\ldots,s_{m_k})} \sum_{q=1}^{Q} \frac{1}{Q} f_{\theta_q}(x_k)\, \alpha(d, x_k)\, E_{\theta_q}\{L(w, \mathcal{A}, \mathcal{I}, \theta_q) \mid x_k; d\}\, ds_{m_1} \ldots ds_{m_k}. \tag{A4}$$

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 $\pi(\theta_q \mid x_k) = Q^{-1} f_{\theta_q}(x_k) / f_\pi(x_k)$. Letting $\oint dx_k$ denote the sum over $(m_1, \ldots, m_k)$ followed by integration over $(s_{m_1}, \ldots, s_{m_k})$, we can re-write (A4) as

$$\oint f_\pi(x_k)\, \alpha(d, x_k) \sum_{q=1}^{Q} \pi(\theta_q \mid x_k)\, E_{\theta_q}\{L(w, \mathcal{A}, \mathcal{I}, \theta_q) \mid x_k; d\}\, dx_k. \tag{A5}$$

In the backwards induction process, the optimal decisions at analyses $k+1, \ldots, K$ are known when analysis k is considered. A Bayes optimal procedure must minimise the expected conditional loss under the posterior distribution of θ. Let $E_{\theta_q}\{L(w, \mathcal{A}, \mathcal{I}, \theta_q) \mid x_k; j, \tilde{d}\}$ 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. If we follow the list of $v+2$ criteria, the optimal choice when in state x_k at analysis k is the action j minimising

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

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

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

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

We wish to show that $b^{\mathrm{T}} R(\tilde{d}) \leq b^{\mathrm{T}} y = b^{\mathrm{T}} \lim_{i\to\infty} R(d_i)$. 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

$$R(d_i) - R(\tilde{d}) \;=\; R(d_i^K) - R(d_i^{-1}) \;=\; \sum_{k=0}^{K} R(d_i^k) - R(d_i^{k-1}). \tag{A6}$$

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

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

Rule d_i^k may take a different action, j, at analysis k, and then proceed as $\tilde{d}$, in which case we write the conditional expected loss under $\theta = \theta_q$ as $E_{\theta_q}\{L(w, \mathcal{A}, \mathcal{I}, \theta_q) \mid x_k; j, \tilde{d}\}$. Let

$$G(w, x_k, j) = \sum_{q=1}^{Q} \pi(\theta_q \mid x_k)\, [E_{\theta_q}\{L(w, \mathcal{A}, \mathcal{I}, \theta_q) \mid x_k; j, \tilde{d}\} - E_{\theta_q}\{L(w, \mathcal{A}, \mathcal{I}, \theta_q) \mid x_k; \tilde{d}\}]. \tag{A7}$$

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 (A5), (A6) and (A7), we obtain

$$w^{\mathrm{T}}R(d_i) - w^{\mathrm{T}}R(\tilde{d}) \;=\; \sum_{k=0}^{K} w^{\mathrm{T}}R(d_i^k) - w^{\mathrm{T}}R(d_i^{k-1}) \;=$$

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

Define $A_1 = \{(x_k,j) : G(w_1,x_k,j) > 0\}$ 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_t, action j is optimal for minimising each of $w_1^{\mathrm{T}}R(d),\ldots,w_{t-1}^{\mathrm{T}}R(d)$ in order, but not then optimal for minimising $w_t^{\mathrm{T}}R(d)$. The functions $G(w,x_k,j)$ are such that $G(w_1,x_k,j) > 0$ for (x_k,j) in A_1 and $G(w_1,x_k,j) = 0$ for (x_k,j) in A_1^c and 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 that $\{d_i\}$ is a sequence of decision rules with $R(d_i) \to y \in \mathcal{Q}_v\backslash\mathcal{S}_v$ where $w_t^{\mathrm{T}}y = w_t^{\mathrm{T}}R(\tilde{d})$ for $t = 1,\ldots,v$ and $b^{\mathrm{T}}y \le b^{\mathrm{T}}r - \epsilon$ for all $r \in \mathcal{S}$. Since $w_1^{\mathrm{T}}R(d_i) - w_1^{\mathrm{T}}R(\tilde{d}) \to 0$,

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

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

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

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

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

$$\sum_{k=0}^{K}\sum_{j=1}^{M+2} \oint I\{(x_k,j) \in A_1\}\,f_\pi(x_k)\,\alpha(d_i,x_k)\,\beta(d_i,x_k,j)\,G(w_t,x_k,j)\,dx_k \;\to\; 0, \quad \text{(A8)}$$

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

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

At the next level, the fact that $w_2^{\mathrm{T}}R(d_i) - w_2^{\mathrm{T}}R(\tilde{d}) \to 0$ and (A8) for $t = 2$ imply that

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

from which we deduce that

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

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 $b^{\mathrm{T}}R(d_i) - b^{\mathrm{T}}R(\tilde{d})$ 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 $w_1^{\mathrm{T}}R(d), \ldots, w_v^{\mathrm{T}}R(d)$ in order and, where this leaves a choice of actions, rule $\tilde{d}$ is defined to minimise the expected contribution to $b^{\mathrm{T}}R(d)$, so that $G(b, x_k, j) \geq 0$. In consequence,

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

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

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

Proof of Corollary 1. Given that $\mathcal{S}$ is closed, arguments of Ferguson (1967, Ch. 2) show that the 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 accepts H_0 with probability 1 under all θ. This can be achieved by stopping 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 $i = Q+1, \ldots, 2Q$. Similarly, 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 $i \in \{Q+1, \ldots, 2Q\}$. These $w(i)$ imply a cost for sampling but not for wrong decisions, so d must stop at $\mathcal{I} = \mathcal{I}_1$ with probability 1. As d is admissible, it is also admissible among 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 $\mathcal{T} = \{R(d) : R(i, d) = \mathcal{I}_1, i = Q+1, \ldots, 2Q\}$. Standard arguments show that $\mathcal{S}'$ is a closed convex set, and $R(d)$ 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 $i \in \{Q+1, \ldots, 2Q\}$, for sufficiently large H. □

APPENDIX 2

The backwards induction algorithm

The Bayes decision problem of § 4 is solved by backwards induction. 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; as this assigns probability zero to $\theta = 0$ and $\theta = \delta$, we can say that sampling cost is one per unit of information for $\theta \notin \{0, \delta\}$ 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

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

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

$$\xi^{(K-1)}(m_{K-1}, S_{m_{K-1}}, m_K) \ = \ (\mathcal{I}_{m_K} - \mathcal{I}_{m_{K-1}}) \int_{-\infty}^{\infty} 1 \,.\, p_2^{(K-1)}(\theta | m_{K-1}, S_{m_{K-1}})\, d\theta$$

$$+ \int_{-\infty}^{\infty} \zeta^{(K)}(m_K, S_{m_K})\, dF^{(K)}(S_{m_K} | m_{K-1}, S_{m_{K-1}}, m_K)$$

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

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

$$+ \int_{-\infty}^{\infty} \eta^{(k+1)}(m_{k+1}, S_{m_{k+1}})\, dF^{(k+1)}(S_{m_{k+1}} | m_k, S_{m_k}, m_{k+1}).$$

Proceeding 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 $k \leq K-2$ we use values for $\eta^{(k+1)}(m_{k+1}, S_{m_{k+1}})$ already computed on a grid of values of $S_{m_{k+1}}$. We divide the range of values for S_{m_k} into intervals within which $\eta^{(k)}(m_k, S_{m_k})$ is attained by one of the following actions: stop now and accept H_0, stop now and reject H_0, continue to $\mathcal{I}_{m_k+1}$, ..., continue to $\mathcal{I}_{M-K+k+1}$. 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 this type of recursive numerical integration.

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. Am. 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. *Statist. Med.* **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.

Falissard, B. & Lellouch, J. (1991). Some extensions to a new approach for interim analysis in clinical trials. *Statist. Med.* **10**, 949–57.

Falissard, B. & Lellouch, J. (1992). A new procedure for group sequential analysis in clinical trials. *Biometrics* **48**, 373–88.

Falissard, B. & Lellouch, J. (1993). The succession procedure for interim analysis: Extensions for early acceptance of H_0 and for flexible times of analysis. *Statist. Med.* **12**, 51–67.

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

Fisher, L. D. (1998). Self-designing clinical trials. *Statist. Med.* **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. R. Statist. Soc.* B **51**, 305–61.

Jennison, C. & Turnbull, B. W. (1997). Group sequential analysis incorporating covariate information. *J. Am. 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. *Statist. Med.* **22**, 971–93.

Jennison, C. & Turnbull, B. W. (2005). Efficient group sequential designs when there are several effect sizes under consideration. *Statist. Med.* **24**, to appear.

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

Mehta, C. R. & Tsiatis, A. A. (2001). Flexible sample size considerations using information-based interim monitoring. *Drug Inform. 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. *Statist. Med.* **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. *Statist. Med.* **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.

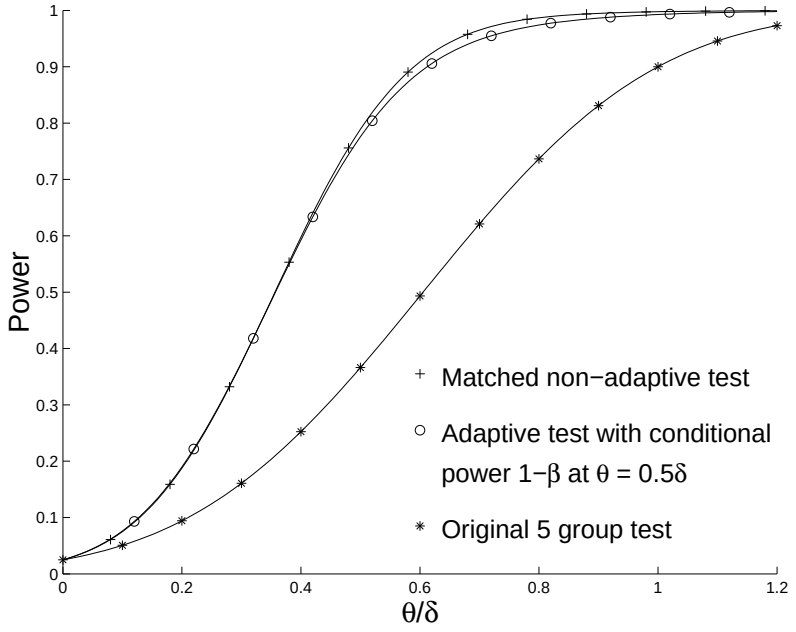
Thach, C. & Fisher, L. D. (2002). Self-designing two-stage trials to minimize expected costs. *Biometrics* **58**, 432–8.

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

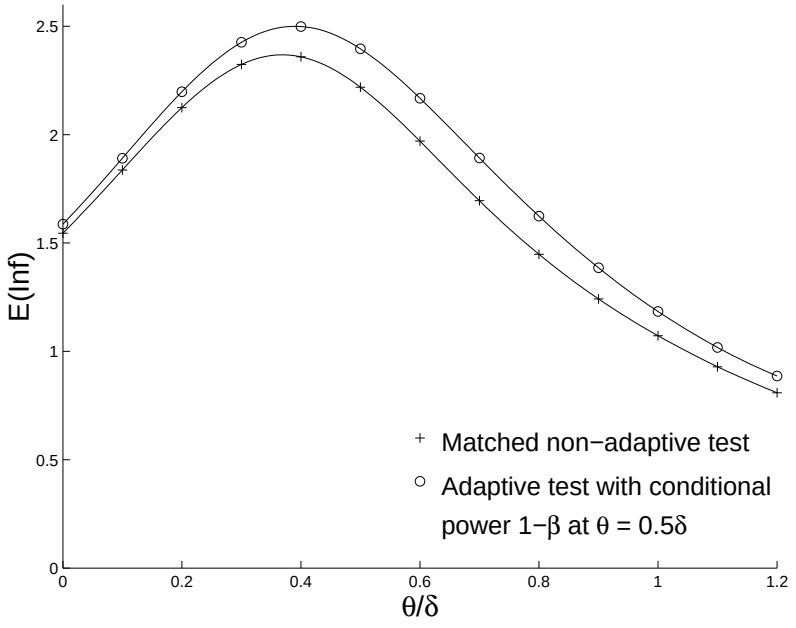
Wittes, J. & Brittain, E. (1990). The role of internal pilot studies in increasing efficiency of clinical trials. *Statist. Med.* **9**, 65–72.

Fig. 1. Example 1. (a) 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 $\theta = 0.59\,\delta$. (b) $E_\theta(\mathcal{I})$ of the Cui et al. adaptive design and the matched non-adaptive design, expressed in units of $\mathcal{I}_f$. (c) Efficiency ratio between the Cui et al. adaptive design and the matched non-adaptive design.

(a)



(b)



(c)

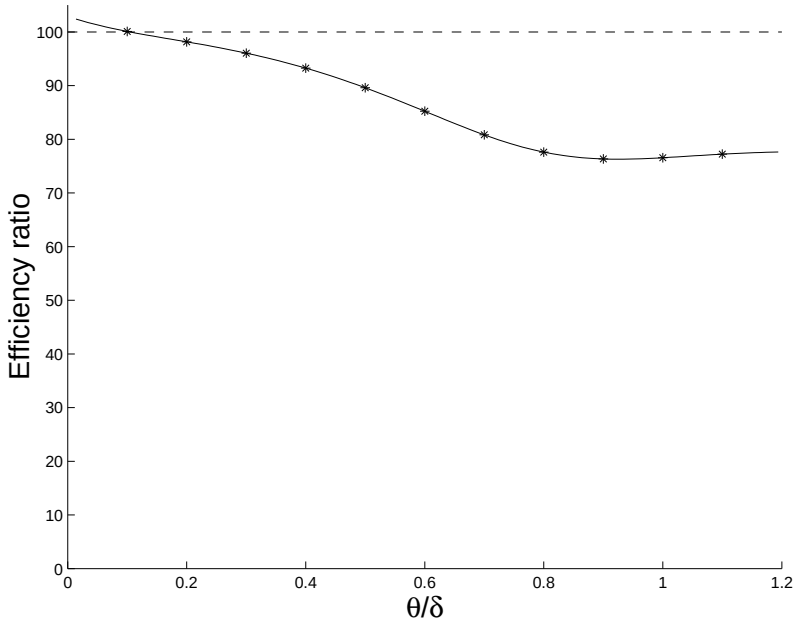
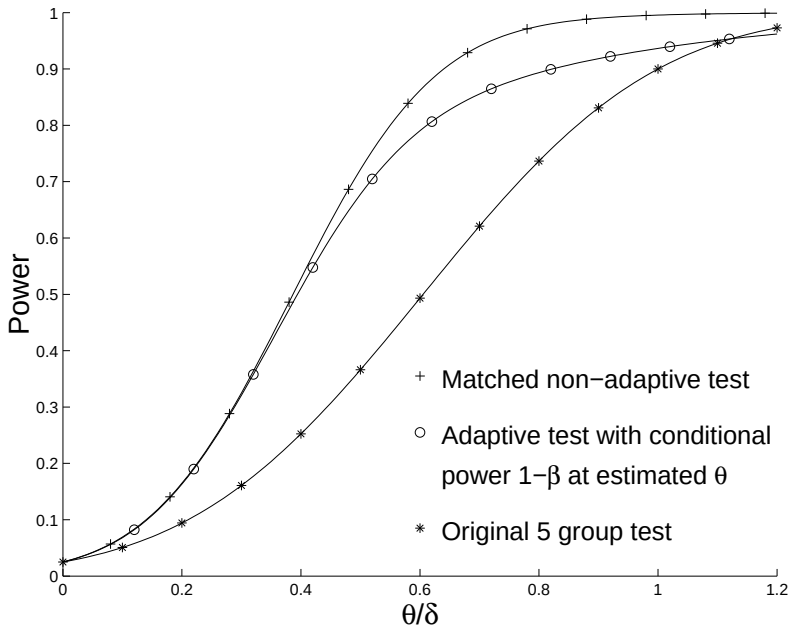
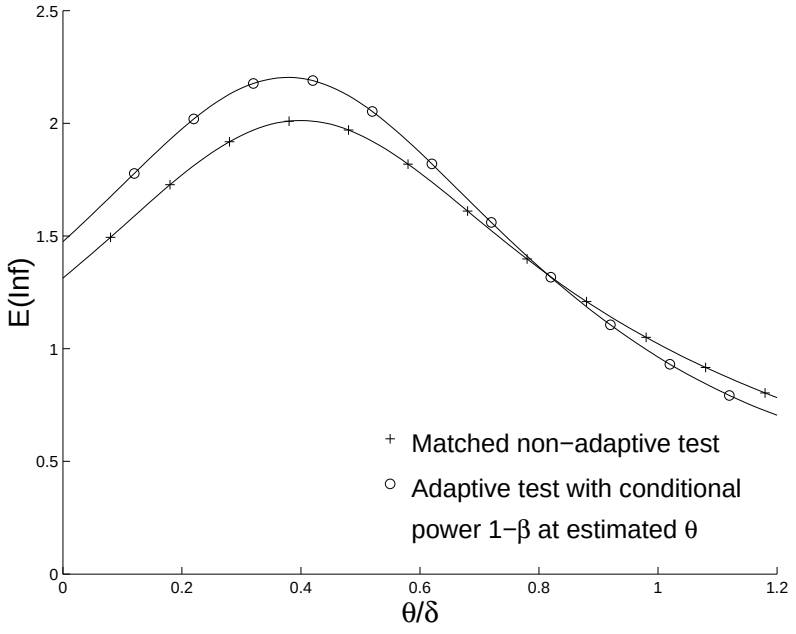

Efficiency ratio
θ/δ
0
10
20
30
40
50
60
70
80
90
100
0
0.2
0.4
0.6
0.8
1
1.2

Fig. 2. Example 2. (a) 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 $\theta = 0.64\,\delta$. (b) $E_\theta(\mathcal{I})$ of the Cui et al. adaptive design and the matched non-adaptive design, expressed in units of $\mathcal{I}_f$. (c) Efficiency ratio between the Cui et al. adaptive design and the matched non-adaptive design.

(a)



(b)



(c)

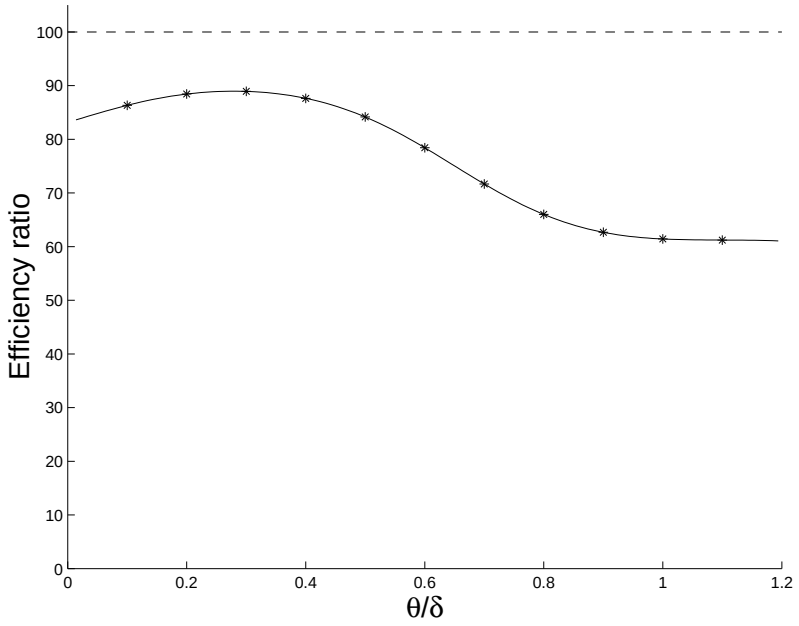

Efficiency ratio
θ/δ
0
10
20
30
40
50
60
70
80
90
100
0
0.2
0.4
0.6
0.8
1
1.2

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