Group sequential tests for delayed responses

Lisa V. Hampson

Medical and Pharmaceutical Statistics Research Unit, Department of Mathematics and Statistics, Lancaster University, Lancaster, UK.

Christopher Jennison

Department of Mathematical Sciences, University of Bath, Bath, UK.

Summary. Group sequential methods are used routinely to monitor clinical trials and to provide early stopping when there is evidence of a treatment effect, lack of an effect, or concerns about patient safety. In many studies, the response of clinical interest is measured some time after the start of treatment and there are subjects at each interim analysis who have been treated but are yet to respond. We formulate a new form of group sequential test which gives a proper treatment of these "pipeline" subjects; these tests can be applied even when the continued accrual of data after the decision to stop the trial is unexpected. We illustrate our methods through a series of examples. We define error spending versions of these new designs which handle unpredictable group sizes and provide an information monitoring framework that can accommodate nuisance parameters, such as an unknown response variance. By studying optimal versions of our new designs, we show how the benefits of lower expected sample size normally achieved by a group sequential test are reduced when there is a delay in response. The loss of efficiency for larger delays can be ameliorated by incorporating data on a correlated short-term endpoint, fitting a joint model for the two endpoints but still making inferences on the original, longer term endpoint. We derive p-values and confidence intervals on termination of our new tests.

Keywords: Adaptive designs; Bayes decision problem; clinical trials; delayed observations; error spending tests; group sequential tests; inference on termination; information monitoring; optimal tests; short-term endpoints

1. Introduction

Randomised controlled trials (RCTs) are the gold standard for evaluating a new treatment. Monitoring of trials to assess recruitment, compliance and safety is standard practice; stopping rules based on accumulating data allow early termination, leading to quicker decisions and lower average sample size. The books of Armitage (1975), Whitehead (1997), Jennison & Turnbull (2000) and Proschan et al. (2006) survey the developing methodology which adapted sequential sampling methods from industrial to medical applications. A major impetus to the use of sequential analysis in RCTs was the concept of group sequential tests (GSTs) introduced by Pocock (1977) and O'Brien & Fleming (1979), which are now widely used and strongly advocated by regulators. Wassmer & Vandemeulebroecke (2006) review software packages that facilitate the implementation of GSTs.

GSTs are typically formulated assuming response is immediate, or at least rapidly observed relative to the length of the trial. However, in many studies, the primary endpoint is defined as a measurement taken some time after treatment has commenced. Even with a rapid response, data cleaning can introduce a delay so at interim analyses there are "responses in the pipeline", i.e., subjects who have commenced treatment but for whom

data on the primary endpoint are not yet available. Investigators are likely to be obliged to follow-up and report these data, even though standard GSTs do not provide a formal framework for doing so. Such a requirement is not usually imposed on survival studies, where it would be unrealistic to wait for complete follow-up of all patients before making a final decision. We shall consider the case where the primary endpoint is determined at a fixed time after treatment but we comment further on survival trials in Section 7.

Methods have been proposed for incorporating data accrued after termination of a GST. In Whitehead's (1992) deletion method, a boundary is calculated ignoring the analysis at which the trial stopped and assuming the final analysis occurred with the pipeline data observed; this boundary is used for hypothesis testing and to determine the final p-value. Hall & Ding (2008) divide the final data into two parts comprising responses observed before and after the stopping decision and then apply a combination test (Bauer & Köhne, 1994). Sooriyarachchi et al. (2003) investigate these methods when the number of responses in the pipeline does not depend on the observed stage of stopping, and find they perform poorly with respect to power. They note that the deletion method is conservative with respect to type I error and, with a modest amount of pipeline data, this leads to lower power than the GST which ignores the additional data; if the number of pipeline subjects is high, the extra data increase power but only by a small amount for the increase in expected sample size. Faldum & Hommel (2007) attribute the conservatism of the deletion method to the "double hurdle" of having to cross the boundary to terminate sampling, then having to remain above the new boundary when the pipeline data are added. With a moderate number of pipeline subjects, the two versions of the Hall & Ding (2008) method, with fixed and random weights in the combination test, show a greater loss of power than the deletion method.

In the above methods, the decision to stop sampling is triggered by a GST appropriate to immediate response. Where the addition of pipeline data leads to loss of power, this can be attributed to reversals in the initial decision of the GST and the higher likelihood of moving from a positive result to a negative one than from negative to positive. We shall avoid this source of conservatism and low power by constructing a delayed response GST which anticipates pipeline data from the outset. Optimising within the class of such designs ensures that as much additional value as possible is gained from the data contributed by pipeline subjects. Anderson (1964) noted the need to accommodate delayed responses in the final analysis of a sequential procedure and showed that the final decision in a test between two hypotheses should be based on a likelihood ratio test using all the final data. He used a sequential probability ratio test to decide when to stop sampling but commented that in the delayed response problem "Determination of optimal stopping rules is an interesting problem. It is difficult, however, and shall not be studied here". We shall determine optimal stopping and decision rules for our delayed response GSTs.

Hall & Liu (2002) analyse data accrued after termination by applying a maximum likelihood estimate ordering to the sample space of final data sets including pipeline subjects. Our methods share the property that final decisions are based on the sufficient statistic and not on the sample path towards this final statistic. Sooriyarachchi et al. (2003) note that Hall & Liu's method does not adapt to situations where group sizes (or, more generally, information levels) are unpredictable. While we initially derive our optimal delayed response GSTs for the case of pre-specified group sizes, we shall go on to derive error spending versions of these designs to handle unpredictable group sizes.

Faldum & Hommel (2007) propose an adaptive group sequential procedure which anticipates pipeline data and builds these into the design. They consider a two stage trial with a conditional type I error function (Proschan & Hunsberger, 1995), constructed to give

type I error rate equal to the pre-specified α. They propose several choices of conditional type I error function along with a rule for choosing the second stage sample size, adapting this to the first stage responses. Mehta & Pocock (2011) use the method of Chen et al. (2004) in defining an adaptive procedure for the case of a delayed response.

Adaptive designs offer flexibility by permitting aspects of a study, including sample size, to be modified at interim analyses while controlling the type I error rate; see, for example, Bauer & Köhne (1994), Fisher (1998), Cui et al. (1999) and Brannath et al. (2009), and the survey of clinical trials using adaptive methodology compiled by Bauer & Einfalt (2006). Jennison & Turnbull (2006) study adaptive GSTs for the case of immediate response and derive optimal rules for data-dependent choice of group size. They compare fully optimal adaptive designs with optimal GSTs with fixed group sizes and conclude that the adaptive approach does not significantly reduce expected sample size below that of a well-chosen non-adaptive GST. Lokhnygina & Tsiatis (2008) report similar findings for adaptive and non-adaptive two stage designs and state that, for their measures of expected sample size, improvements arising from adaptive choice of the second group size are less than 1% of the fixed sample size. In assessing group sequential designs when there is a delay in response, it is important to include in the expected sample size all subjects recruited to the trial, whether or not they contribute significantly to the final decision. Mehta & Pocock (2011) argue that adaptive designs should have an efficiency advantage over standard GSTs. Our delayed response GSTs are similar to adaptive GSTs in that the amount of pipeline data at an interim analysis sets a minimum value for the remaining sample size. We have optimised our two-stage adaptive GSTs with a delayed response over a class of procedures including those of Faldum & Hommel (2007). Our conclusions are similar to the case of immediate response: while adaptive choice of group size can reduce expected sample size slightly, the benefits over a good non-adaptive delayed response GST are insubstantial.

It is useful to have examples and questions in mind as we introduce our new methodology.

Example A: Normally distributed response, moderate delay

Facey (1992) considers a placebo controlled trial of a treatment for hypercholesterolemia. The primary endpoint, reduction in total serum cholesterol level after four weeks of treatment, is assumed to be normally distributed. Under the anticipated recruitment of 16 subjects per month, about 16 pipeline responses are expected at each interim analysis.

Suppose termination occurs at the first interim analysis with a standardised test statistic of $Z = 2.4$, which exceeds the boundary value of 2.3. When pipeline responses are added to the data, the standardised statistic falls to $Z = 2.1$. There is now some confusion as to how to interpret the trial outcome: is it legitimate to claim a positive result or not?

Example B: Ordinal response, longer delay

Whitehead (1993) describes the ASCLEPIOS stroke trial, comparing an experimental drug against placebo. The primary endpoint was the Barthel Index 90 days after randomisation. At the first interim analysis, with 140 responses, the Data and Safety Monitoring Board recommended that recruitment close. However, the 90 day lag in response and delays in data transfer meant a further 89 patients had incomplete records and it was deemed appropriate to continue to treat and follow-up these cases.

The expected sample size attributed to a group sequential design should include all subjects recruited. With this in mind, how can we design efficient trials when a long delay in response will lead to a significant proportion of the total sample size being treated but

not yet observed at each interim analysis? Furthermore, when both time and sample size are considered, how should an efficient design be constructed?

Example C: Adaptive sampling rules

Faldum & Hommel (2007) describe a trial investigating the incidence of ischaemic-type biliary lesions within 6 months of liver transplantation. Patients are randomised to receive in situ perfusion only (group A) or in situ perfusion plus arterial ex situ flushing (group B). Let θ denote the lesion probability in group A minus that in group B. In testing H_0: $\theta \leq 0$ against $\theta > 0$, an interim analysis is to be conducted once 50 responses are available, by which time 130 subjects will have been recruited.

Faldum & Hommel (2007) propose two-stage designs with inferences based on a combination test. The second stage sample size is allowed to depend on first stage responses, with a minimum value of 80, the number of "pipeline" subjects. This adaptive choice of sample size may have benefits for efficiency but how great are these benefits and do they compensate for logistical complications of modifying the sample size in mid-study or the fact that inferences are not based on minimal sufficient statistics?

Example D: Measurements on a short-term secondary endpoint

Todd & Stallard (2005) describe a trial of a treatment intended to prevent bone fracture in postmenopausal women. While the endpoint of clinical interest is occurrence of a fracture within five years, changes in bone mineral density after one year will also be measured. An obvious question is how this short-term response might be used in a group sequential test. More generally, when a short-term response correlated with the primary endpoint is available, what reductions in expected sample size can this provide?

We shall address the issues raised in these examples in the course of this paper. In Section 2, we introduce delayed response GSTs, designed to include analysis of pipeline data in the final decision. In Section 3, we derive optimal versions of these designs that minimise the expected number of patients and time to a conclusion. When the delay in response is small we find group sequential monitoring leads to considerable reductions in expected sample size, just as for an immediate response; however, the savings decrease for larger delays. The benefits of group sequential monitoring are retained more fully when the objective is to minimise the expected time to reach a conclusion. In Section 3.3, we extend our designs to include response-dependent choice of group size, but find there is little to be gained from doing so. In Section 4, we develop additional features for delayed response GSTs. We define error spending versions of our tests to control the type I error rate when group sizes and numbers of pipeline responses are unpredictable. We propose methods for calculating p-values and confidence intervals on termination, using a stage-wise ordering of the sample space so that inferences do not depend on unknown future information levels. We then apply these methods to draw valid inferences after a trial has unexpectedly overrun.

In Section 5, we show how to incorporate data on a correlated short-term endpoint into delayed response GSTs. Inferences still concern the long-term, primary endpoint but we see that information provided by a more rapidly observed short-term endpoint at interim analyses can substantially improve the procedure's overall efficiency. Finally, in Section 6 we apply our methods with unknown nuisance parameters such as the response variance or the correlation between short-term and long-term endpoints considered in Section 5. Our simulations confirm the accuracy of this approach for controlling error rates.

2. Group sequential tests for delayed responses

2.1. A new form of group sequential test for delayed responses

We introduce our methodology for the case of a two treatment comparison with a single long-term endpoint. Suppose responses are independent and distributed as $X_{A,i} \sim N(\mu_A, \sigma^2)$ and $X_{B,i} \sim N(\mu_B, \sigma^2)$, $i = 1, 2, \ldots$, for subjects randomised to treatments A and B respectively. For now, we assume σ^2 is known. Let $\theta = \mu_A - \mu_B$ and consider a test of H_0: $\theta \leq 0$ against $\theta > 0$ with type I error rate α at $\theta = 0$ and power $1 - \beta$ at $\theta = \delta$.

Suppose responses are available a time Δ_t after randomisation. A "delayed response GST" can terminate recruitment at an interim analysis but waits a time Δ_t while pipeline responses become available before conducting a "decision analysis" to reject or accept H_0. A K-stage design has $K - 1$ possible interim analyses. For $k = 1, \ldots, K - 1$, denote the number of responses at interim analysis k by n_k and the number at the subsequent decision analysis, if recruitment stops, by $\tilde{n}_k$. If recruitment continues past interim analysis $K-1$, we wait for responses from the total sample of $\tilde{n}_K$ subjects before conducting the final decision analysis. Note that we do not allow recruitment to be re-started once it has stopped.

For $k = 1, \ldots, K - 1$, let $\hat{\theta}_k$ denote the maximum likelihood estimate (MLE) of θ based on n_k responses at interim analysis k and define $\mathcal{I}_k = \{\mathrm{Var}(\hat{\theta}_k)\}^{-1}$, the Fisher information for θ, and $S_k = \mathcal{I}_k \hat{\theta}_k$, the score statistic. Similarly, for $k = 1, \ldots, K$, let $\tilde{\theta}_k$ be the MLE of θ based on $\tilde{n}_k$ responses at decision analysis k, $\tilde{\mathcal{I}}_k = \{\mathrm{Var}(\tilde{\theta}_k)\}^{-1}$ and $\tilde{S}_k = \tilde{\mathcal{I}}_k \tilde{\theta}_k$. A K-stage, one-sided delayed response GST of H_0: $\theta \leq 0$ against $\theta > 0$ has the following form:

$$
\begin{array}{lll}
\text{At interim analysis } k = 1, \ldots, K-2, & & \\
\quad \text{if } S_k \leq l_k \text{ or } S_k \geq u_k & \text{stop recruitment and proceed to decision analysis } k, & \\
\quad \text{otherwise} & \text{continue recruitment and proceed to interim analysis } k+1. & \\
\text{At interim analysis } K-1, & & \\
\quad \text{if } S_{K-1} \leq l_{K-1} \text{ or } S_{K-1} \geq u_{K-1} & \text{stop accrual and proceed to decision analysis } K-1, & (1) \\
\quad \text{otherwise} & \text{complete recruitment and proceed to decision analysis } K. & \\
\text{At decision analysis } k = 1, \ldots, K, & & \\
\quad \text{if } \tilde{S}_k \geq c_k & \text{reject } H_0, & \\
\quad \text{if } \tilde{S}_k < c_k & \text{accept } H_0. &
\end{array}
$$

Figure 1 illustrates this definition for a 3-stage delayed response GST. The stopping boundaries l_k and u_k are indicated by $\bullet$ and decision boundaries c_k by $\times$.

2.2. Computations for a delayed response GST

For a test of form (1), let $\mathcal{C}_k = (l_k, u_k)$ denote the region in which recruitment continues at interim analysis k. For $k = 1, \ldots, K - 1$, the probability of stopping recruitment at interim analysis k and rejecting H_0 at the subsequent decision analysis is

$$\psi_k(\theta) = \mathbb{P}(S_1 \in \mathcal{C}_1, \ldots, S_{k-1} \in \mathcal{C}_{k-1}, S_k \notin \mathcal{C}_k, \tilde{S}_k \geq c_k; \theta)$$

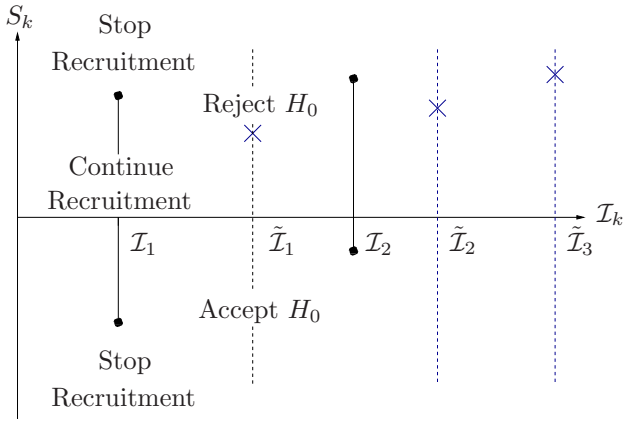


Fig. 1. Boundaries of a three-stage delayed response GST.

and the probability of rejecting H_0 at the final stage is

$$\psi_K(\theta) = \mathbb{P}(S_1 \in \mathcal{C}_1, \ldots, S_{K-1} \in \mathcal{C}_{K-1}, \tilde{S}_K \geq c_K; \theta).$$

The type I error rate at $\theta = 0$ is the sum of the probabilities $\psi_1(0), \ldots, \psi_K(0)$ and power at $\theta = \delta$ is the sum of $\psi_1(\delta), \ldots, \psi_K(\delta)$. Let N denote the total number of subjects recruited when the trial terminates. The expected sample size under treatment effect θ is

$$\begin{aligned}\mathbb{E}(N;\theta) = \sum_{k=1}^{K-1} \mathbb{P}(S_1 \in \mathcal{C}_1, \ldots, S_{k-1} \in \mathcal{C}_{k-1}, S_k \notin \mathcal{C}_k; \theta)\, \tilde{n}_k \\ + \mathbb{P}(S_1 \in \mathcal{C}_1, \ldots, S_{K-1} \in \mathcal{C}_{K-1}; \theta)\, \tilde{n}_K.\end{aligned} \tag{2}$$

Thus, the key requirement in computing properties of a delayed response GST is calculation of probabilities $\psi_k(\theta)$ and the related probabilities in the right hand side of (2).

THEOREM 1. *With normally distributed responses as described above, for each* $k = 1, \ldots, K-1$, $(S_1, \ldots, S_k, \tilde{S}_k)$ *is multivariate normal with* $S_j \sim N(\mathcal{I}_j\theta, \mathcal{I}_j)$, $j = 1, \ldots, k$, $\tilde{S}_k \sim N(\tilde{\mathcal{I}}_k\theta, \tilde{\mathcal{I}}_k)$, *and independent increments. We refer to this as the "canonical joint distribution". The analogous property holds for* $(S_1, \ldots, S_{K-1}, \tilde{S}_K)$. *Furthermore, these joint distributions hold asymptotically when score statistics are formed from MLEs of a parameter* θ *in a general parametric model.*

PROOF. The statistics $S_1, \ldots, S_k, \tilde{S}_k$ are based on nested subsets of the data at decision analysis k and, similarly, $S_1, \ldots, S_{K-1}, \tilde{S}_K$ are functions of nested subsets of the data at decision analysis K. The distributional results of Jennison & Turnbull (1997) for a sequence of statistics based on accumulating data apply to this setting. It follows that $(S_1, \ldots, S_k, \tilde{S}_k)$ has the canonical joint distribution, as does $(S_1, \ldots, S_{K-1}, \tilde{S}_K)$. The general theory of Jennison & Turnbull (1997) and Scharfstein et al. (1997) implies the same results hold asymptotically when score statistics are formed from MLEs in a general parametric model.

Since statistics follow the same form of joint distribution as in a GST with immediate response, the numerical routines of Jennison (1993) and Jennison & Turnbull (2000, Chapter 19) for standard GSTs are also applicable to delayed response GSTs.

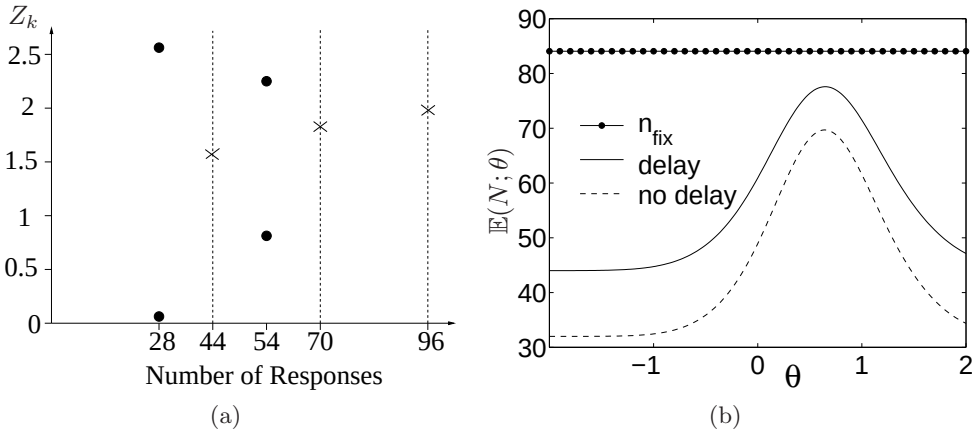


Fig. 2. (a) Three-stage test of H_0: $\theta \leq 0$ against $\theta > 0$ minimising expected sample size criterion F with type I error rate $\alpha = 0.025$ at $\theta = 0$ and power $1-\beta = 0.9$ at $\theta = 1$. Boundaries are plotted on the standardised test statistic scale. (b) Expected sample size curves for: the fixed sample test; optimal delayed response GST with unconstrained choice of c_k; optimal GST for an immediate response.

2.3. Revisiting Example A: Normally distributed response, moderate delay

We illustrate our delayed response GSTs in an application based on Example A of Section 1. Let θ denote the difference in mean response between new treatment and placebo and suppose we wish to test H_0: $\theta \leq 0$ against $\theta > 0$ with type I error rate $\alpha = 0.025$ and power $1 - \beta = 0.9$ at $\theta = 1$. We assume responses are normally distributed with variance $\sigma^2 = 2$. A fixed sample test requires $n_{fix} = 86$ subjects, allocated equally to the two treatments. (Unequal allocation would require a larger total sample size in both fixed sample and group sequential designs.) We consider a group sequential design with a maximum sample size of 96 to compensate for loss of power due to early stopping. Suppose responses are measured $\Delta_t = 4$ weeks after randomisation to treatment and there is no additional delay in preparing data for an analysis (if there were, we would simply increase the value of Δ_t by this amount).

We shall design a 3-stage delayed response GST, planning ahead for responses that will be in the pipeline at each interim analysis. If subjects are recruited at a rate of 4 per week, the maximum sample size of $n_{max} = 96$ subjects will be recruited by $t_{max} = 24$ weeks and all observations will be available at 28 weeks. In this scenario, $\Delta_t / t_{max} = 1/6$ and this fraction of the trial's maximum sample size, amounting to 16 observations, will be in the pipeline at each interim analysis. In scheduling analyses, note that responses first become available after calendar time Δ_t and the last subject is recruited by time t_{max}. Consider a design with three interim analyses equally spaced between Δ_t and t_{max} at times $\Delta_t + (t_{max} - \Delta_t)/3$, $\Delta_t + 2\,(t_{max} - \Delta_t)/3$ and t_{max}. Each interim analysis has a possible decision analysis 4 weeks later. Since the third interim analysis always leads to a decision analysis, it can be omitted. Rounding to even numbers so there are equal numbers on each treatment leads to interim analyses and decision analyses with observed numbers of responses $n_1 = 28$, $\tilde{n}_1 = 44$, $n_2 = 54$, $\tilde{n}_2 = 70$ and $\tilde{n}_3 = 96$.

Boundaries of a delayed response GST for this problem are depicted in Figure 2(a), expressed as critical values for $Z_k = S_k/\sqrt{\mathcal{I}_k}$ and $\tilde{Z}_k = \tilde{S}_k/\sqrt{\tilde{\mathcal{I}}_k}$ (the choice of scale is unimportant since the conversion between statistics Z_k and S_k is straightforward). The

design shown in Figure 2(a) was chosen to optimise the weighted average of $\mathbb{E}(N;\theta)$ over a $N(1/2,(1/2)^2)$ density for θ. So, among all procedures with the same interim analysis and decision times, type I error rate 0.025 and power 0.9 at $\theta = 1$, it has the minimum value of

$$F = \int \mathbb{E}(N;\theta)\,\frac{2}{\delta}\,\phi\left(\frac{\theta-\delta/2}{\delta/2}\right)d\theta \tag{3}$$

for $\delta = 1$. (Here ϕ denotes the standard normal probability density function.) We explain in Section 2.4 how to derive tests meeting this — and other — optimality criteria. The value of F for this design is 68.6, a significant improvement on the fixed sample size of 86.

The solid line in Figure 2(b) is the function $\mathbb{E}(N;\theta)$ for this optimal delayed response GST; reductions below the fixed sample size are evident at all values of θ. The dashed line shows $\mathbb{E}(N;\theta)$ for the conventional GST minimising F for the same problem when response is immediate and analyses are conducted after $n_1 = 32$, $n_2 = 64$ and $n_3 = 96$ responses. Comparing the curves for these two cases, we see the effect of a four week delay in response is to reduce the savings in expected sample size. Both expected sample size curves are approximately symmetric about $\theta = 0.64$ with $\mathbb{E}(N;\theta = 0.64) = 69.7$ when response is immediate, compared to 77.6 when $\Delta_t = 4$. The fixed sample test, standard GST and delayed response GST have type I error rate 0.025 and power 0.9 at $\theta = 1$ and, in consequence, their power curves are closely matched over the whole range of θ values.

When recruitment is terminated with $S_k \geq u_k$ in a test of form (1), one expects the final decision will be to reject H_0. However, it is possible for the pipeline data to cause a "reversal" so $\tilde{S}_k < c_k$ and the eventual decision is to accept H_0. Similarly, termination of recruitment with $S_k \leq l_k$ may be followed by a final decision to reject H_0. In considering the incorporation of data received after a GST has stopped, Sooriyarachchi et al. (2003) note it would be particularly regrettable to stop recruitment based on a positive trend, only to miss out on a significant result at the decision analysis. For our delayed response GST, the total reversal probability under θ is the sum over $k = 1,\ldots,K-1$ of the terms

$$\varphi_k(\theta) = \mathbb{P}(S_1 \in \mathcal{C}_1,\ldots,S_{k-1} \in \mathcal{C}_{k-1}, S_k \geq u_k, \tilde{S}_k < c_k;\theta) \tag{4}$$

and

$$\eta_k(\theta) = \mathbb{P}(S_1 \in \mathcal{C}_1,\ldots,S_{k-1} \in \mathcal{C}_{k-1}, S_k \leq l_k, \tilde{S}_k \geq c_k;\theta). \tag{5}$$

The total reversal probability for the optimal delayed response GST depicted in Figure 2(a) takes a maximum value of 0.01 under $\theta = 0.63$. The small value of this probability gives reassurance that the decision to terminate recruitment is not being made prematurely. However, the fact that the information provided by pipeline data has little impact on the final outcome of the hypothesis test means it is inevitable that pipeline subjects increase the expected sample size over that attainable in the case of immediate response. Inspection of the boundaries in Figure 2(a) shows that for both $k = 1$ and 2, c_k lies some way below u_k. Thus, if recruitment is stopped with $Z_k \geq u_k$, H_0 will still be rejected at the decision analysis for a range of $\tilde{Z}_k$ values less extreme than Z_k. We have, therefore, avoided the difficulties associated with a decrease in Z value referred to in Example A of Section 1.

Examination of Figure 2(a) shows that the decision analyses at stages 1 and 2 have values of c_k below 1.96, the critical value for $\tilde{Z}_k$ in a fixed sample test with one-sided significance level 0.025. Since rejection of H_0 with such a low value of the Z-statistic may not be deemed acceptable in practice, we have also derived optimal designs with the constraint that each $c_k \geq 1.96$. The effects of imposing this constraint are small; the expected sample size function for this design remains very close to the optimal design without this constraint, and the power curves for these two tests are almost indistinguishable.

2.4. Optimal delayed response group sequential tests

We now present the methods used to derive optimal versions of our delayed response GSTs. We consider tests with specified information levels $\{\mathcal{I}_1, \tilde{\mathcal{I}}_1, \ldots, \mathcal{I}_{K-1}, \tilde{\mathcal{I}}_{K-1}, \tilde{\mathcal{I}}_K\}$ that have type I error rate α at $\theta = 0$ and power $1 - \beta$ at $\theta = \delta$. Within this class we seek designs minimising expected sample size at a single value of θ, averaged over several θ values, or integrated against a probability density for θ. The literature on optimal sequential tests goes back to Wald and Wolfowitz's (1948) proof that Wald's (1947) sequential probability ratio test minimises $\mathbb{E}(N;\theta)$ under the null hypothesis $\theta = 0$ and alternative $\theta = \delta$. Minimisation of $\mathbb{E}(N;\theta = \delta/2)$, known as the Kiefer-Weiss problem (Kiefer & Weiss, 1957), has been considered in the fully sequential case by Weiss (1962), Lai (1973) and Lorden (1976). Optimal GSTs for this and other criteria have been found by Jennison (1987), Eales & Jennison (1992, 1995), Barber & Jennison (2002) and Öhrn & Jennison (2010).

For conciseness, we focus on the derivation of delayed response GSTs that minimise the objective function F defined in (3), but note it is just as easy to find designs minimising the integral of $\mathbb{E}(N;\theta)$ over a general $N(\zeta, \omega^2)$ density for θ or other optimality criteria, such as those of Barber & Jennison (2002). The type I error rate and power requirements create a constrained optimisation problem. A key step is to construct a method for solving an unconstrained Bayes sequential decision problem which penalises type I errors under $\theta = 0$ and type II errors under $\theta = \delta$ but does not insist on specific error rates; we then find a combination of prior distribution and costs for which the Bayes procedure satisfies the error rate requirements. It follows that this procedure minimises F among all tests with type I error rate α and power $1 - \beta$ at $\theta = \delta$. This same technique underlies theoretical and numerical results in many of the papers referred to above. Banerjee & Tsiatis (2006), who apply this approach to find optimal adaptive designs, note it is equivalent to using Lagrangian multipliers. The new feature in our application is the delayed response.

In deriving a delayed response GST to minimise F, we first place a prior on θ with point masses $1/3$ at 0 and δ and a continuous component with $1/3$ times a $N(\delta/2, (\delta/2)^2)$ density. Let A_0 denote the action of accepting H_0 and A_1 that of rejecting H_0. We define the loss function $L(A_i, \theta)$ for taking decision A_i when the treatment effect is θ as $L(A_1, 0) = d_1$, $L(A_0, \delta) = d_0$ and all other $L(A_i, \theta) = 0$. We specify a sampling cost of $c(\theta)$ per subject recruited, where $c(\delta) = c(0) = 0$ and $c(\theta) = c_0$ otherwise. The total expected cost is then

$$\begin{aligned}\mathbb{E}(\text{Cost}) &= \mathbb{E}(\text{Cost of making incorrect decisions}) + \mathbb{E}(\text{Sampling cost}) \\ &= \frac{1}{3}\left\{d_1\,\mathbb{P}(\text{Reject } H_0; \theta = 0) + d_0\,\mathbb{P}(\text{Accept } H_0; \theta = \delta) + c_0 F\right\}.\end{aligned}$$

Incorrect decisions under other values of θ do not appear in this formula since the Lagrangian construction only requires penalties for errors under $\theta = 0$ and under $\theta = \delta$. A Bayes test minimising $\mathbb{E}(\text{Cost})$ can be found using backwards induction and numerical integration. Recursive relations for the expected additional cost associated with the actions permitted at each stage are presented in Appendix 1. It remains to search for values of d_0, d_1 and c_0 such that this Bayes test has $\mathbb{P}(\text{Reject } H_0; \theta = 0) = \alpha$ and $\mathbb{P}(\text{Accept } H_0; \theta = \delta) = \beta$. The following argument establishes that there is such a triple with $c_0 = 1$.

Let $\mathcal{D}$ be the set of all non-randomised "decision rules" of the form (1) plus randomised decisions rules defined as probability distributions over these non-randomised rules. For each rule $d \in \mathcal{D}$, define the risk vector

$$R(d) = (R_1(d),\, R_2(d),\, R_3(d)) = (\,\mathbb{P}(\text{Reject } H_0;\, \theta = 0),\, \mathbb{P}(\text{Accept } H_0;\, \theta = \delta),\, F\,). \qquad (6)$$

The proof of Jennison & Turnbull (2006, Theorem 1) can be adapted to the delayed response problem to show that the risk set $\mathcal{S} = \{R(d);\, d \in \mathcal{D}\}$ is closed. Let $\mathcal{I}_{fix}$ denote the information required by the fixed sample test with error rates α and β. We restrict attention to problems with information sequences satisfying $\tilde{\mathcal{I}}_1 < \mathcal{I}_{fix} < \tilde{\mathcal{I}}_K$. It is straightforward to construct examples of tests with error rates α and β, so there are points (α, β, F) in $\mathcal{S}$ for some values of F. Denote the infimum of values of F such that (α, β, F) lies in $\mathcal{S}$ by F_0. The risk vector (α, β, F_0) lies on the boundary of $\mathcal{S}$ and, because the risk set is closed, it also belongs to $\mathcal{S}$. It follows that there exists a delayed response GST with risk vector (α, β, F_0) and this is a solution to our original frequentist problem.

Since the set $\mathcal{S}$ is convex, the supporting hyperplane theorem (see Ferguson (1967), Section 2.7) implies there is a supporting hyperplane to $\mathcal{S}$ at the boundary point (α, β, F_0). We can write the equation of this hyperplane as

$$d_1 R_1 + d_0 R_2 + c_0 R_3 = \xi \tag{7}$$

and note that $d_1 R_1(d) + d_0 R_2(d) + c_0 R_3(d) \geq \xi$ for all risk vectors $R(d) \in \mathcal{S}$. Consider the Bayes decision problem with costs (d_0, d_1, c_0). The risk vectors of solutions to this problem lie in the intersection of $\mathcal{S}$ and the hyperplane. In fact, we show in Appendix 8.2 that this Bayes problem has a unique solution up to the definition of actions on sets of Lebesgue measure zero. Hence, the hyperplane intersects $\mathcal{S}$ at a single point, and this must be (α, β, F_0). It follows that solving the Bayes problem with costs (d_0, d_1, c_0) solves our original frequentist problem and this solution is unique. The sampling cost c_0 must be non-zero, otherwise the optimal rule would automatically continue to $\tilde{\mathcal{I}}_K$ and error rates would be lower than α and β. Thus, without loss of generality, we can choose to write (7) with $c_0 = 1$.

The final step in the derivation of the optimal delayed response GST is a search over pairs of decision costs $d_0 > 0$ and $d_1 > 0$ to find the Bayes problem whose solution has the desired error rates. The uniqueness result of Appendix 8.2 is helpful here as it implies each pair (d_0, d_1) leads to a single pair of error probabilities. It is convenient to work with the transformed costs $\log(d_0)$ and $\log(d_1)$ as this leads to an unconstrained search over $\mathbb{R}^2$.

3. Efficiency of optimal designs

3.1. Tests optimal for expected sample size

With optimal delayed response GSTs to refer to, we can quantify the effect of a delay in response on the sample size savings relative to a fixed sample test. We return to the problem, introduced in Section 2.1, of comparing treatments A and B with normal responses with means μ_A and μ_B. Defining $\theta = \mu_A - \mu_B$, we wish to test H_0: $\theta \leq 0$ against $\theta > 0$ with type I error rate α and power $1 - \beta$ at $\theta = \delta$. The fixed sample test requires information

$$\mathcal{I}_{fix} = \frac{\{\Phi^{-1}(1-\alpha) + \Phi^{-1}(1-\beta)\}^2}{\delta^2}, \tag{8}$$

where Φ denotes the standard normal cumulative distribution function, and we set the maximum information for the delayed response GST at $\mathcal{I}_{max} = R\mathcal{I}_{fix}$. Let n_{max} be the number of subjects needed for information $\mathcal{I}_{max}$, assuming equal allocation between treatments. Suppose c patients are recruited per unit time and define t_{max} to be the time taken to recruit n_{max} subjects. Suppose responses are observed after a delay Δ_t and define $r = \Delta_t / t_{max}$. The parameter r combines the effect of Δ_t, c and n_{max} in an overall measure of the effect of response delay on the information observed at each interim analysis.

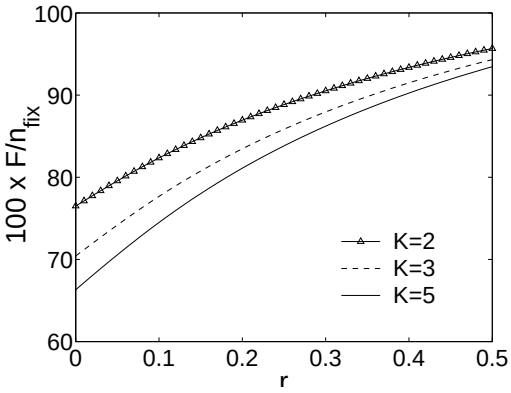


Fig. 3. Minima of F for optimal K-stage delayed response GSTs of H_0: $\theta \leq 0$ against $\theta > 0$. All tests have $\alpha = 0.025$, $\beta = 0.1$, $R = 1.1$ and analyses scheduled at information levels given by (10). Results are expressed as percentages of the corresponding fixed sample size test.

Since the first responses are observed at time Δ_t and the last subject is recruited by t_{max}, we consider designs with interim analyses equally spaced between Δ_t and t_{max} at

$$t_k = \Delta_t + (k/K)(t_{max} - \Delta_t), \quad k = 1, \ldots, K. \tag{9}$$

This gives numbers of observed responses

$$n_k = \frac{k}{K}(1-r)n_{max} \quad \text{and} \quad \tilde{n}_k = n_k + r\, n_{max}, \quad k = 1, \ldots, K$$

and the information sequence

$$\mathcal{I}_k = \frac{k}{K}(1-r)\mathcal{I}_{max} \quad \text{and} \quad \tilde{\mathcal{I}}_k = \mathcal{I}_k + r\, \mathcal{I}_{max}, \quad k = 1, \ldots, K. \tag{10}$$

Although we have included an interim analysis at $\mathcal{I}_K$ to make the pattern clear, the only option at this point is to continue to a decision analysis at $\tilde{\mathcal{I}}_K$. This interim analysis can, therefore, be omitted and doing so gives a design of the form introduced in Section 2.1.

We present results for delayed response GSTs for this problem which minimise the expected sample size criterion F defined in (3). Figure 3 plots the minima of F expressed as a percentage of n_{fix} for tests with $n_{max} = 1.1\, n_{fix}$. Results for $r = 0$, the case of an immediate response, show savings of 23.5%, 29.6% and 33.7% relative to the fixed sample test for GSTs with $K = 2$, 3 and 5 analyses, respectively. Minimum values of F/n_{fix} are invariant to changes in δ and σ^2, and to changes in Δ_t and c which preserve the value of r (see Appendix 8.3), so these results apply to a range of problems.

Figure 3 shows that benefits of group sequential monitoring decrease as r increases. Substantial savings are still present for small values of r, for example, F is 77.7% of n_{fix} when $r = 0.1$ and $K = 3$. However, savings relative to the fixed sample test fall by about half as r increases from 0 to 0.25. One reason for this is that when Δ_t is large, by the time of the first interim analysis recruitment will have progressed so that a large fraction of n_{max} has already been taken, even if accrual stops at this earliest opportunity.

Another reason for reduced efficiency when there is a long delay in response is the lack of impact of information from pipeline subjects on the final decision. Table 1 lists

Table 1. Reversal probabilities for two-stage delayed response GSTs.

	Switch to correct decision		Switch to incorrect decision	
	$\theta = 0$	$\theta = \delta$	$\theta = 0$	$\theta = \delta$
r	$S_1 \geq u_1;\ \tilde{S}_1 < c_1$	$S_1 \leq l_1;\ \tilde{S}_1 \geq c_1$	$S_1 \leq l_1;\ \tilde{S}_1 \geq c_1$	$S_1 \geq u_1;\ \tilde{S}_1 < c_1$
0.01	10^{-10}	3×10^{-13}	10^{-13}	4×10^{-10}
0.1	0.000428	0.000707	0.000175	0.000808
0.2	0.00195	0.00460	0.000885	0.00276
0.3	0.00437	0.0116	0.00176	0.00474
0.4	0.00824	0.0228	0.00276	0.00680
0.5	0.0150	0.0416	0.00395	0.00925

reversal probabilities of the two-stage tests minimising F under values of r between 0.01 and 0.5. These are $\varphi_1(0)$, $\eta_1(\delta)$, $\eta_1(0)$ and $\varphi_1(\delta)$ where φ and η are defined in (4) and (5). For small r, the reversal probabilities are particularly small and the test's power would change very little if the final decision were determined purely by the boundary crossed at an interim analysis $k < K$. Thus, pipeline subjects at analyses 1 to $K - 1$ add to $\mathbb{E}(N; \theta)$ while contributing little towards increased power. Since there is only a decision analysis at stage K, the pipeline subjects do make a contribution at this point. The loss of efficiency is evident in Figure 3, with F increasing most rapidly with r when r is small.

Despite the low overall probabilities of reversal in Table 1, pipeline data can, on occasion, play an important role in providing a final check on the conclusion at a decision analysis. Table 2 shows the conditional probability that the two-stage delayed response GSTs of Table 1 reject H_0 given that recruitment terminates at analysis 1 with S_1 on the boundary at l_1 or u_1. Some of these conditional probabilities are well away from 0 and 1. In particular, a reversal to avoid an incorrect conclusion is quite likely when $\theta = 0$ and $S_1 = u_1$ or $\theta = \delta$ and $S_1 = l_1$. Reversals that would lead to a wrong conclusion are much less likely. To reconcile these numbers with the much smaller overall probabilities in Table 1, remember the conditioning events for columns 1 and 4 of Table 2 are quite unlikely: the boundary crossed is at odds with the value of θ and this boundary is crossed with no overshoot. The results of this section quantify the impact of a delay in response and have implications for study design. Since efficiency benefits from a low value of r, it is important to reduce any delay in the availability of responses due to data cleaning and preparation. Recruitment strategies also affect the value of r. In one example, Mehta (2009) notes that halving the recruitment rate and re-defining the primary response to be measured at 12 rather than 26 weeks reduces the number of pipeline subjects from 208 to 48. However, slower enrolment will imply later analyses and a delay in the final decision. In the next section, we consider trial designs which balance the twin objectives of low sample size and a rapid decision.

3.2. Tests optimal for a combination of objectives

3.2.1. Optimising designs for a combination of objectives

When a phase III clinical trial reaches a positive conclusion, it is in the interest of future patients who may benefit from the new treatment that this should occur as rapidly as possible. Liu et al. (2004) note the importance of the time remaining on patent to the economic benefit for the company developing a new treatment. In our discussion of the stroke trial of Example B, we raised the question of balancing the (possibly conflicting)

Table 2. Conditional probabilities of rejecting H_0 for two-stage delayed response GSTs, given recruitment stops at the first interim analysis with S_1 on the boundary at l_1 or u_1.

	$S_1 = u_1$		$S_1 = l_1$	
r	$\theta = 0$	$\theta = \delta$	$\theta = 0$	$\theta = \delta$
0.01	1.000	1.000	1.7×10^{-11}	1.6×10^{-10}
0.1	0.876	0.987	0.00338	0.0512
0.2	0.648	0.971	0.00973	0.207
0.3	0.459	0.961	0.0141	0.370
0.4	0.314	0.952	0.0175	0.517
0.5	0.208	0.944	0.0207	0.642

objectives of low sample size and a rapid decision. We now present delayed response GSTs which optimise a criterion combining sample size and time of the final decision. With this new objective in mind, we modify our delayed response GSTs by adding the option to stop at an interim analysis with an immediate decision to reject or accept H_0; with this modification, we also add an interim analysis K to choose between stopping with an immediate decision or waiting to observe the final set of pipeline responses. When this new form of test stops with a decision to accept or reject H_0 at an interim analysis, we would still expect the pipeline data to be reported informally, once they become available. However, in some situations, an early conclusion might lead to changes in treatment for some of the subjects remaining in the study (for example, patients randomised to the inferior treatment may switch to the comparator) and there would then be no further pipeline data.

Consider again a comparison of treatments A and B with normally distributed responses with variance σ^2 and difference in means θ. We are to test H_0: $\theta \leq 0$ against $\theta > 0$ with type I error rate α and power $1 - \beta$ at $\theta = \delta$. A fixed sample size design requires $n_{fix} = 4\,\sigma^2\,\mathcal{I}_{fix}$ subjects where $\mathcal{I}_{fix}$ is given by (8). With recruitment at constant rate c, the fixed sample test reaches a conclusion after time $t_{fix} = n_{fix}/c + \Delta_t$. Let T denote the time taken for a delayed response GST to reach a conclusion and, as before, let N denote the total number of subjects recruited when the trial terminates. We scale N by n_{fix} and T by t_{fix} in defining the combined measure of sample size and time to a conclusion

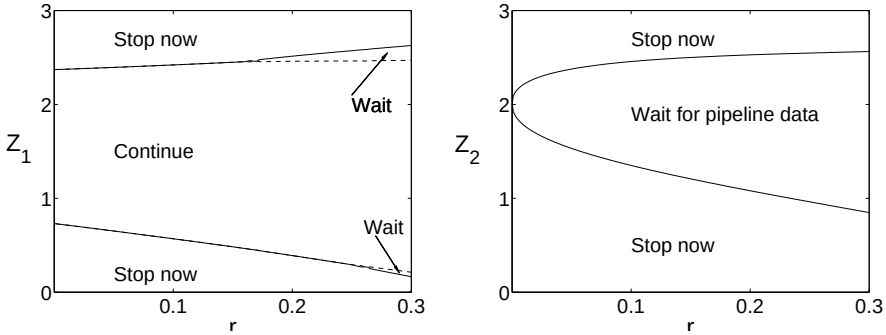
$$H \;=\; a\,N/n_{fix} + b\,T/t_{fix}\,,$$

where the dimensionless weights a and b are chosen to satisfy $a \geq 0$, $b \geq 0$ and $a + b = 1$.

We consider K-stage tests with $\mathcal{I}_{max} = R\,\mathcal{I}_{fix}$ and information levels defined by (10), and seek designs which minimise

$$G = \int \mathbb{E}(H;\theta)\,\frac{2}{\delta}\,\phi\left(\frac{\theta - \delta/2}{\delta/2}\right)d\theta,$$

a generalisation of the criterion F to the new objective. When response is immediate, $T = N/c$ and minimising G is equivalent to minimising F. Optimal designs can be found using the methods described in Section 2.4 with a sampling cost made up of a/n_{fix} per subject and b/t_{fix} per unit of time to reach a conclusion, and with the extra option of stopping with a final decision at each interim analysis. In searching over values of S_k to find the optimal action at interim analysis k, the most complex pattern we have found is five intervals of increasing values of S_k with optimal actions: Accept H_0, Stop recruitment



(a) Interim analysis $k = 1$ (b) Interim analysis $k = 2$

Fig. 4. Stopping rule at interim analyses 1 and 2 of the delayed response GST minimising G for $a = b = 0.5$ among tests with $K = 2$, $\alpha = 0.025$, $\beta = 0.1$, $R = 1.1$ and information sequence (10).

and wait for pipeline data, Continue recruitment, Stop recruitment and wait for pipeline data, and Reject H_0. In other cases, one or both of the intervals where the optimal action is to stop and wait for pipeline data is absent. Thus, these optimal designs have the form:

$$
\begin{array}{ll}
\text{At interim analysis } k = 1, \ldots, K-1, & \\
\quad \text{if } S_k \geq u_k^\star & \text{stop recruitment and reject } H_0, \\
\quad \text{if } S_k \leq l_k^\star & \text{stop recruitment and accept } H_0, \\
\quad \text{if } l_k^\star < S_k \leq l_k \text{ or } u_k \leq S_k < u_k^\star & \text{stop recruitment and proceed} \\
 & \text{to decision analysis } k, \\
\quad \text{otherwise} & \text{continue recruitment and proceed} \\
 & \text{to interim analysis } k+1. \\
\text{At interim analysis } K, & \\
\quad \text{if } S_K \geq u_K^\star & \text{stop recruitment and reject } H_0, \\
\quad \text{if } S_K \leq l_K^\star & \text{stop recruitment and accept } H_0, \\
\quad \text{otherwise} & \text{stop recruitment and proceed} \\
 & \text{to decision analysis } K.
\end{array}
\tag{11}
$$

It is possible that $u_k^\star = u_k$ or $l_k^\star = l_k$ at some interim analyses. Figure 4 illustrates optimal two-stage designs for $r \leq 0.3$. When pipeline data are few, the option to stop at the first interim analysis and wait for them is not used; for larger r, pipeline data can make a greater contribution and the interval on which it is optimal to wait widens. We have found similar features in optimal designs for other examples with (a, b) ranging between $(0, 1)$ and $(1, 0)$.

3.2.2. Efficiency of delayed response GSTs optimised for a combination of objectives
Figure 5 shows the efficiency of delayed response GSTs of form (11) minimising G. Designs have $K = 5$, $\alpha = 0.025$, $\beta = 0.1$ and $R = 1.1$. Information is assumed to be proportional to the number of observed responses. Results are presented as $100\,G$ for comparability with previous values of $100\,F/n_{fix}$. When $(a, b) = (1, 0)$, G is identical to F and the top-most

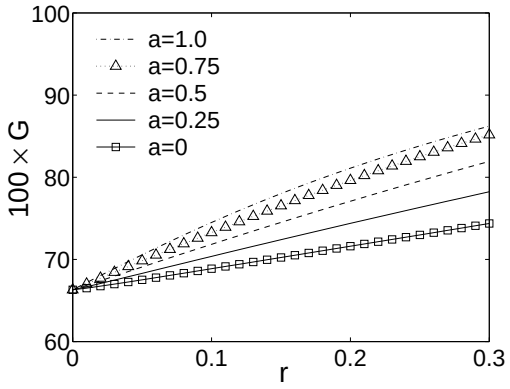


Fig. 5. Minima of $100 \times G$ for delayed response GSTs with $K = 5$, $\alpha = 0.025$, $\beta = 0.1$, $R = 1.1$ and information sequence (10).

curve in Figure 5 is the same as that for $K = 5$ in Figure 3. Lower values for G when $b > 0$ and $a < 1$ show the benefits of group sequential testing are more substantial, even for quite large values of r, when both time to a decision and sample size are considered.

For illustration, we consider an example incorporating features of Example B, the ASCLEPIOS stroke trial. As before, suppose responses on treatments A and B are normally distributed with difference in means θ and we are to test H_0: $\theta \leq 0$ against $\theta > 0$ with type I error rate α. We set power $1-\beta = 0.9$ at $\theta = 0.34$ and assume $\sigma^2 = 1$ so a fixed sample size test requires $n_{fix} = 364$ observations, divided equally between the two treatments, to obtain information $\mathcal{I}_{fix} = 90.9$. Suppose the primary endpoint is measured at 90 days and it takes 30 days to clean the data before each interim and decision analysis, so $\Delta_t = 90 + 30 = 120$ days. If one subject is recruited per day, it will take $t_{fix} = 364 + 120 = 484$ days for the fixed sample test to reach a conclusion. In designing a delayed response GST, we set $\mathcal{I}_{max} = 1.1\,\mathcal{I}_{fix} = 100$, requiring $n_{max} = 400$ subjects to be recruited over $t_{max} = 400$ days. Thus, $r = \Delta_t / t_{max} = 0.3$. In a 5-stage design, analyses are scheduled according to (10), giving $\mathcal{I}_k = 14k$ and $\tilde{\mathcal{I}}_k = 14k + 30$, for $k = 1, \ldots, 5$. In order to achieve $\mathcal{I}_1 = 14$ at the first interim analysis, the data set is "locked" at $t = 90 + 56 = 146$ days, with observed responses from the first 56 patients. Responses observed during the 30 days of data cleaning are not used in this analysis. By the time of the first interim analysis, 176 subjects will have been recruited to the trial. If it is decided to halt recruitment at this point and wait for the pipeline data, it will take 90 days for responses to be observed and 30 days of data cleaning, giving a decision analysis at 296 days. In a similar way, further interim and decision analyses are scheduled at times $t_k = 120 + 56k$ and $\tilde{t}_k = t_k + 120$, for $k = 1, \ldots, 5$, with the data set for interim analysis k locked 30 days prior to t_k. Properties of optimal delayed response GSTs for this problem are those for the case $r = 0.3$ in Figure 5. Varying a and b alters the emphases on sample size and time to a conclusion. If we focus solely on time and set $a = 0$, the optimal design has a mean duration of 360 days (averaging over the normal distribution for θ in the definition of G). This is 124 days less than the fixed sample test's $t_{fix} = 484$ days. In the case of an immediate response, $\Delta_t = 0$, the GST minimising G has an average length of 241 days, 123 days less than the 364 days for a fixed sample design. Thus, there are savings in average duration from using a group sequential design for quite large values of the delay parameter r.

Table 3. Minima of F expressed as a percentage of n_{fix} for optimal two-stage delayed response GSTs with and without adaptive choice of the second group size. Tests are for $\alpha = 0.025$, $\beta = 0.1$, $R = 1.1$ and $r = 0.2$ using a selection of values of $\mathcal{I}_1$. The choice $\mathcal{I}_1/\mathcal{I}_{fix} = 0.425$ is optimal both for adaptive and non-adaptive designs. The case $\mathcal{I}_1/\mathcal{I}_{fix} = 0.555$ is included for comparison with a Faldum & Hommel (2007) design.

$\mathcal{I}_1/\mathcal{I}_{fix}$	Optimal adaptive delayed response GST	Optimal non-adaptive delayed response GST
0.1	97.1	98.7
0.2	92.1	92.8
0.3	88.4	88.7
0.4	86.8	87.0
0.425	86.8	86.9
0.5	87.4	87.6
0.555	88.7	88.8
0.6	90.2	90.3
0.7	94.9	95.0

3.3. Adaptive sampling rules

In Example C of Section 1, we discussed Faldum & Hommel's (2007) adaptive two-stage designs. Delayed response designs of form (1) with $K = 2$ can be extended to allow an adaptive choice of the second group size. With our current definition, after interim analysis 1, the test proceeds to a final analysis with either $\tilde{n}_1$ responses (adding just the pipeline subjects) or $\tilde{n}_2$ responses (recruiting additional subjects). Allowing a data-dependent choice of final sample size in the range $\tilde{n}_1$ to $\tilde{n}_2$ gives a "sequentially planned sequential test", as proposed by Schmitz (1993) for immediate response. We have optimised such designs, minimising the function F, for $\alpha = 0.025$, $\beta = 1$, $R = 1.1$ and $r = 0.2$. Table 3 lists minima of F, expressed as a percentage of n_{fix}, for optimal adaptive and non-adaptive delayed response GSTs. Only minor savings are achieved by the adaptive choice of group size, just as Jennison & Turnbull (2006) found for the case of immediate response.

We have applied the definitions of Faldum & Hommel (2007, Section 4.3) to implement their two-stage test using a conditional error function with "linear level in $[0, \alpha_1]$". In their notation, $\alpha_0 = 0.3$ and $\beta_2 = 0.95$. In our optimal adaptive and non-adaptive designs, we set $n_{max} = 1.1\, n_{fix}$ and constrain the second group size to lie in the interval $[\tilde{n}_1 - n_1, n_{max} - n_1]$. With $r = 0.2$, the power of Faldum & Hommel's test is 0.9 at $\theta = \delta$ when $\mathcal{I}_1/\mathcal{I}_{fix} = 0.555$ and this design has F equal to 90.4% of n_{fix}. It is evident from Table 3 that, for this value of $\mathcal{I}_1/\mathcal{I}_{fix}$, little is gained by introducing adaptivity. We also see that values of $\mathcal{I}_1/\mathcal{I}_{fix}$ around 0.4 give the most efficient designs. Our overall conclusion is that efficient designs can be found by exploring the options offered by simpler, non-adaptive tests.

4. Practical implementation of delayed response GSTs

4.1. Error spending tests for delayed responses

Lan & DeMets (1983) introduced error spending group sequential tests to deal with unpredictable information sequences arising from random variation in accrual rates or loss to follow-up. For a one-sided test, Jennison & Turnbull (2000, Section 7.3) describe how to construct efficacy and futility boundaries by spending type I and type II error probability

according to two pre-specified functions of observed information. We now define error spending versions of our delayed response GSTs to provide the same flexibility.

Consider a test of H_0: $\theta \leq 0$ against $\theta > 0$ with type I error rate α at $\theta = 0$ and power $1 - \beta$ at $\theta = \delta$. The error spending functions involve a maximum information level, $\mathcal{I}_{max}$. It is assumed that, with no early stopping, recruitment will continue until there are sufficient subjects to generate information level $\mathcal{I}_{max}$ and cease at the first interim analysis k at which it is anticipated $\tilde{\mathcal{I}}_k \geq \mathcal{I}_{max}$. The value of $\mathcal{I}_{max}$ should be chosen to meet the power requirement and we shall return to this topic after explaining how the design with a particular $\mathcal{I}_{max}$ is implemented. In order that the sequence $\{S_1, \ldots, S_k, \tilde{S}_k\}$ has the usual canonical joint distribution given $\{\mathcal{I}_1, \ldots, \mathcal{I}_k, \tilde{\mathcal{I}}_k\}$, as defined in Section 2.2, future information levels must not be influenced by previously observed outcomes: this precludes, for example, changing the schedule of analyses when data are close to a testing boundary.

The type I and type II error spending functions $f(t)$ and $g(t)$ are increasing functions of $t = \mathcal{I}/\mathcal{I}_{max}$, where $\mathcal{I}$ is the current observed information, with $f(0) = g(0) = 0$ and $f(t) = \alpha$ and $g(t) = \beta$ for $t \geq 1$. At interim analysis k, we know $\mathcal{I}_1, \ldots, \mathcal{I}_k$ and also the number of pipeline subjects who will contribute to $\tilde{\mathcal{I}}_k$. We need critical values l_k and u_k for S_k and c_k for $\tilde{S}_k$ such that the cumulative probability under $\theta = 0$ of rejecting H_0 by analysis k is $f(\mathcal{I}_k/\mathcal{I}_{max})$ and the probability under $\theta = \delta$ of accepting H_0 by analysis k is $g(\mathcal{I}_k/\mathcal{I}_{max})$. Here, we use $\mathcal{I}_k$ rather than $\tilde{\mathcal{I}}_k$ in f and g since the low reversal probabilities discussed in Section 3.1 indicate that the statistic S_k at interim analysis k plays the major role in reaching a decision at this stage. However, on reaching an analysis with $\tilde{\mathcal{I}}_k \geq \mathcal{I}_{max}$, the value of $\tilde{\mathcal{I}}_k$ is used, giving $f(\tilde{\mathcal{I}}_k/\mathcal{I}_{max}) = \alpha$ and $g(\tilde{\mathcal{I}}_k/\mathcal{I}_{max}) = \beta$ since $\tilde{\mathcal{I}}_k/\mathcal{I}_{max} \geq 1$.

We shall describe two ways to obtain values for l_k, u_k and c_k. In the first method, we calculate l_k and u_k at interim analysis k and only consider c_k on reaching decision analysis k. In the second method, we assume the number of subjects who have been treated but have not yet responded is known at interim analysis k and $\tilde{\mathcal{I}}_k$ can be predicted; then we use this value for $\tilde{\mathcal{I}}_k$ in calculating l_k, u_k and c_k together at interim analysis k.

Method 1

At the first analysis, set u_1 and l_1 to satisfy

$$\mathbb{P}(S_1 \geq u_1; \theta = 0) = f(\mathcal{I}_1/\mathcal{I}_{max}) \quad \text{and} \quad \mathbb{P}(S_1 \leq l_1; \theta = \delta) = g(\mathcal{I}_1/\mathcal{I}_{max}).$$

For $k > 1$, choose u_k such that

$$\mathbb{P}(S_1 \in \mathcal{C}_1, \ldots, S_{k-1} \in \mathcal{C}_{k-1}, S_k \geq u_k; \theta = 0) = f(\mathcal{I}_k/\mathcal{I}_{max}) - f(\mathcal{I}_{k-1}/\mathcal{I}_{max}) \tag{12}$$

and l_k to satisfy

$$\mathbb{P}(S_1 \in \mathcal{C}_1, \ldots, S_{k-1} \in \mathcal{C}_{k-1}, S_k \leq l_k; \theta = \delta) = g(\mathcal{I}_k/\mathcal{I}_{max}) - g(\mathcal{I}_{k-1}/\mathcal{I}_{max}). \tag{13}$$

If $S_k \leq l_k$ or $S_k \geq u_k$, recruitment ceases and the study proceeds to decision analysis k. With l_k and u_k as defined in (12) and (13), the allocated increments in type I and type II error probability would be "spent" if the final decision were simply to reject H_0 if $S_k \geq u_k$ and to accept H_0 if $S_k \leq l_k$. Pipeline data increase the observed information to $\tilde{\mathcal{I}}_k$ at the decision analysis, where the critical value c_k is chosen to satisfy

$$\begin{aligned} &\mathbb{P}(S_1 \in \mathcal{C}_1, \ldots, S_{k-1} \in \mathcal{C}_{k-1}, S_k \geq u_k, \tilde{S}_k < c_k; \theta = 0) = \\ &\mathbb{P}(S_1 \in \mathcal{C}_1, \ldots, S_{k-1} \in \mathcal{C}_{k-1}, S_k \leq l_k, \tilde{S}_k \geq c_k; \theta = 0). \end{aligned} \tag{14}$$

This choice balances the probability under $\theta = 0$ of sample paths that cross the lower boundary at interim analysis k and end above c_k against that of paths crossing the upper boundary u_k which fall back below c_k. Thus, the null probability of rejecting H_0 at the stage k decision analysis remains the same as the probability of observing $S_k \geq u_k$, so the cumulative type I error probability up to stage k remains equal to $f(\mathcal{I}_k/\mathcal{I}_{max})$.

Incorporating information from the pipeline subjects reduces the type II error probability spent by stage k, so power is higher than originally specified. (The extent to which this occurs could be reduced by replacing $g(\mathcal{I}_{k-1}/\mathcal{I}_{max})$ in (13) by the type II error actually accruing up to analysis $k-1$.) Method 1 has the advantage that l_k and u_k can be calculated at interim analysis k without knowing the value of $\tilde{\mathcal{I}}_k$, which may not be completely predictable due, for example, to slow information flow in a multi-centre trial.

When recruitment is terminated at interim analysis k in anticipation of $\tilde{\mathcal{I}}_k \geq \mathcal{I}_{max}$, we proceed directly to the decision analysis and spend all the remaining type I error probability by setting c_k as the solution to

$$\mathbb{P}(S_1 \in \mathcal{C}_1, \ldots, S_{k-1} \in \mathcal{C}_{k-1}, \tilde{S}_k \geq c_k; \theta = 0) = \alpha - f(\mathcal{I}_{k-1}/\mathcal{I}_{max}). \tag{15}$$

Formally, the final error probability α comes from evaluating $f(\tilde{\mathcal{I}}_k/\mathcal{I}_{max})$. This procedure based on $\tilde{S}_k$ but not S_k has the same form as the final analysis K of a delayed response GST in our original definition (1). Even if, in fact, $\tilde{\mathcal{I}}_k < \mathcal{I}_{max}$, a final decision must be made and c_k is chosen to satisfy (15), bringing the final type I error probability up to α.

Method 2

Suppose it is possible to predict the value of $\tilde{\mathcal{I}}_k$ at the time of interim analysis k. Then, we can find values l_k, u_k and c_k to bring the cumulative type I and II error probabilities up to exactly $f(\mathcal{I}_k/\mathcal{I}_{max})$ and $g(\mathcal{I}_k/\mathcal{I}_{max})$, respectively. To do this, we set u_k to be the solution to (12), then search for the value of l_k that satisfies

$$\mathbb{P}(S_1 \in \mathcal{C}_1, \ldots, S_{k-1} \in \mathcal{C}_{k-1}, S_k \notin \mathcal{C}_k, \tilde{S}_k < c_k; \theta = \delta) = g(\mathcal{I}_k/\mathcal{I}_{max}) - g(\mathcal{I}_{k-1}/\mathcal{I}_{max}) \tag{16}$$

where c_k denotes the solution to (14) for each candidate value of l_k. As in Method 1, the final analysis, where $\tilde{\mathcal{I}}_k \geq \mathcal{I}_{max}$, is treated differently and c_k is obtained as the solution to (15) to give total type I error probability α.

In both Methods 1 and 2, the attained power depends primarily on $\mathcal{I}_{max}$ and, to a lesser extent, on the sequence of information levels actually observed. At the design stage, error spending functions and target information level $\mathcal{I}_{max}$ can be chosen so that power $1-\beta$ at $\theta = \delta$ is achieved under an assumed pattern of information levels $\mathcal{I}_k$ and $\tilde{\mathcal{I}}_k$, $k = 1, \ldots, K$. Departures from this pattern will perturb the attained power but one can expect power to be close to the desired value as long as the target $\mathcal{I}_{max}$ is eventually reached.

We illustrate our error spending designs and assess their efficiency in the example introduced in Section 2.1 and developed in Section 3.1. The values of α, β and δ determine $\mathcal{I}_{fix}$ and the inflation factor R determines $\mathcal{I}_{max} = R\mathcal{I}_{fix}$. The sample size needed to achieve $\mathcal{I}_{max}$ is $n_{max} = 4\,\sigma^2\,\mathcal{I}_{max}$. With accrual rate c, these subjects can be recruited in time $t_{max} = n_{max}/c$ and, with response delay Δ_t, the delay parameter is $r = \Delta_t/t_{max}$.

We consider error spending, delayed response GSTs with spending functions

$$f(t) = \alpha \min\{t^\rho, 1\} \quad \text{and} \quad g(t) = \beta \min\{t^\rho, 1\} \tag{17}$$

based on the ρ-family of one-sided error spending functions (Jennison & Turnbull, 2000, Section 7.3) and we shall find the value of ρ appropriate to the given choice of R and $\mathcal{I}_{max}$.

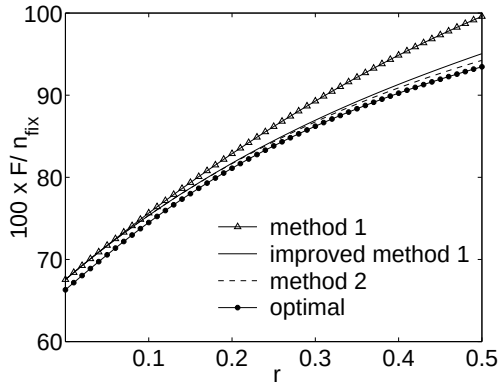


Fig. 6. Performance for objective function F of error spending and optimal delayed response GSTs. Tests are designed and implemented for information sequences of the form (10) with $K = 5$ and $R = 1.1$. Tests have type I error probability $\alpha = 0.025$ and power $1 - \beta = 0.9$ at $\theta = \delta$.

We assume analysis times as in (9) so information levels follow the pattern (10), reaching $\tilde{\mathcal{I}}_K = \mathcal{I}_{max}$ at analysis K. The stopping boundaries and decision rule for an error spending design following either Method 1 or Method 2 can be derived as described above. For the test to conclude properly at stage K attaining power $1 - \beta$ at $\theta = \delta$, we require the value of c_K obtained from (15) also to satisfy

$$\mathbb{P}(S_1 \in \mathcal{C}_1, \ldots, S_{K-1} \in \mathcal{C}_{K-1}, \tilde{S}_K \leq c_K; \theta = \delta) = \beta - g(\mathcal{I}_{K-1}/\mathcal{I}_{max}). \tag{18}$$

This will not be the case for a general choice of ρ, but we can search for the value of ρ for which this occurs. (An alternate strategy would be to fix ρ and search for an appropriate value of $\mathcal{I}_{max}$, noting that in this case the value of r varies with $\mathcal{I}_{max}$.)

We assess the efficiency of error spending designs by comparing them with delayed response GSTs optimised for the objective function F defined in (3). In Figure 6, for the case $K = 5$, error spending tests with boundaries derived using Method 1 require values of ρ between 1.3 and 2.0 for $0 \leq r \leq 0.5$, with tests for larger delay parameter r needing smaller ρ. For tests designed according to Method 2, ρ ranges from 0.9 to 2.0. Calculations are for the case of observed sample sizes in agreement with planning assumptions but error spending GSTs will, of course, adapt to other eventualities. The efficiency of the error spending designs is impressive. Values of F for tests derived using Method 2 exceed the minimum possible by at most 2% of n_{fix} for all $r \leq 0.5$. This performance is matched for small r by the more flexible Method 1 designs, although their efficiency diverges from the optimum as r increases beyond 0.2. This divergence is related to the extra power of the Method 1 designs: as r increases to 0.5, attained power increases from 0.9 to 0.913. As suggested previously, we can adjust these designs by replacing $g(\mathcal{I}_{k-1}/\mathcal{I}_{max})$ in (13) and $g(\mathcal{I}_{K-1}/\mathcal{I}_{max})$ in (18) by the cumulative type II error probability actually spent, to ensure the error spent approximates that specified by g more closely over the course of the trial and type II error probability β is fully spent at the final analysis. This adjustment improves the Method 1 designs and the values of F shown in Figure 6 are within 1% of n_{fix} of the values achieved by Method 2. We conclude that the proposed error spending methods offer a very effective parallel to current methodology for the case of immediate response.

4.2. P-values and confidence intervals on termination of a delayed response GST

A delayed response GST answers the single question whether $\theta \leq 0$ or $\theta > 0$. It is often desirable to provide a more complete analysis on termination of a trial; FDA and EMA guidelines 'E9: Statistical Principles for Clinical Trials' recommend presenting confidence intervals for treatment effects, and it is common to present a p-value for testing $H_0 : \theta \leq 0$.

Setting $\mathcal{C}_K = \emptyset$, define $T = \min\{k\colon S_k \notin \mathcal{C}_k\}$. Following Jennison & Turnbull (2000, Section 8.2), we can write the probability density of the sequence of responses up to termination of a test of form (1) as a product of a term involving $\tilde{\mathcal{I}}_T$, $\tilde{S}_T$ and θ and a term that does not depend on θ. It follows that the pair $(\tilde{\mathcal{I}}_T, \tilde{S}_T)$ is a sufficient statistic for θ; moreover the conditional distribution of the sample path given its end point $(\tilde{\mathcal{I}}_T, \tilde{S}_T)$ does not depend on θ. The sample space for this sufficient statistic is

$$\Omega = \bigcup_{k=1}^{K} \{(\tilde{\mathcal{I}}_k, \tilde{s})\colon\ \tilde{s} \in \mathbb{R}\}. \tag{19}$$

We can define tests of H_0: $\theta \leq 0$ against $\theta > 0$ at a continuum of significance levels by specifying the rejection region of each test as a set of outcomes in Ω. On termination of a trial, the one-sided upper p-value for testing H_0: $\theta \leq 0$ is the minimum significance level at which H_0 is rejected. An ordering of the points in Ω is needed to define tests with nested rejection regions so that this definition of a p-value is well-founded. We write $(\tilde{\mathcal{I}}_{k'}, \tilde{s}') \succeq (\tilde{\mathcal{I}}_k, \tilde{s})$ to denote that $(\tilde{\mathcal{I}}_{k'}, \tilde{s}')$ is placed higher than $(\tilde{\mathcal{I}}_k, \tilde{s})$ in such an ordering. The one-sided p-value for testing H_0: $\theta \leq 0$ against $\theta > 0$ based on outcome $(\tilde{\mathcal{I}}_{k^\star}, s^\star)$ is

$$p^+ = \mathbb{P}\{\,(\tilde{\mathcal{I}}_T, \tilde{S}_T) \succeq (\tilde{\mathcal{I}}_{k^\star}, s^\star)\,; \theta = 0\}. \tag{20}$$

When response is immediate, Rosner & Tsiatis (1988) note there is no uniformly most powerful test of H_0: $\theta = 0$ on the sample space of a standard GST, and hence no single natural ordering of outcomes. Jennison & Turnbull (2000, Section 8.4) survey proposed orderings, including: the stage-wise ordering of Armitage (1957) in which outcomes are ordered first by the boundary crossed, then by the analysis at which stopping occurs, and lastly by the value of the test statistic; the MLE ordering (Armitage, 1958); the signed likelihood ratio test ordering (Chang and O'Brien, 1986), which is equivalent to ordering by Z_T; and the score test ordering (Rosner & Tsiatis, 1988). The stage-wise ordering has been used by Siegmund (1978), Fairbanks & Madsen (1982), Jennison & Turnbull (1983) and Tsiatis et al. (1984), and is the method preferred by Proschan et al. (2006). In this ordering, the position of an observed outcome $(\tilde{\mathcal{I}}_k, \tilde{s})$ does not depend on the values of unobserved information levels $\mathcal{I}_{k+1}, \ldots, \mathcal{I}_K$. This is not the case for other orderings, so only the stage-wise ordering can be used to make inferences on termination of error spending GSTs when future information levels beyond the observed stage of stopping are unknown. We shall base our methods for delayed response GSTs on an adaptation of the stage-wise ordering so they may be used with the error spending versions of delayed response GSTs.

For consistency between the outcome of a standard GST and the upper p-value p^+, the smallest values of p^+ should occur when the test rejects H_0. While a standard GST has a continuation region separating outcomes with a particular $\mathcal{I}_k$ into upper and lower sections, Figure 1 illustrates that $\tilde{S}_k$ can take values on the whole real line at decision analysis k. Hence, in ordering the sample space Ω for a delayed response GST, we need to partition the range of outcomes with $\tilde{\mathcal{I}}_T = \tilde{\mathcal{I}}_k$ into a "high end" with $\tilde{S}_k \geq c_k$ and a "low end" with $\tilde{S}_k < c_k$. In the spirit of the stage-wise ordering, we create an overall ordering of Ω by

defining the relation $(\tilde{\mathcal{I}}_{k_1}, \tilde{s}_{k_1}) \succeq (\tilde{\mathcal{I}}_{k_2}, \tilde{s}_{k_2})$ to hold if

$$\begin{array}{ll} \text{(a)} & \tilde{\mathcal{I}}_{k_1} = \tilde{\mathcal{I}}_{k_2} \text{ and } \tilde{s}_{k_1} \geq \tilde{s}_{k_2}, \text{ or} \\ \text{(b)} & \tilde{\mathcal{I}}_{k_1} < \tilde{\mathcal{I}}_{k_2} \text{ and } \tilde{s}_{k_1} \geq c_{k_1}, \text{ or} \\ \text{(c)} & \tilde{\mathcal{I}}_{k_1} > \tilde{\mathcal{I}}_{k_2} \text{ and } \tilde{s}_{k_2} < c_{k_2}. \end{array} \tag{21}$$

Since $(\tilde{\mathcal{I}}_{k_1}, \tilde{s}_{k_1}) \succeq (\tilde{\mathcal{I}}_{k_2}, \tilde{s}_{k_2})$ whenever $\tilde{s}_{k_1} \geq c_{k_1}$ and $\tilde{s}_{k_2} < c_{k_2}$, outcomes for which the delayed response GST rejects H_0 are highest in the ordering, so the p-value is consistent with the test outcome in that $p^+ \leq \alpha$ if and only if H_0 is rejected.

Although partitioning outcomes $\tilde{\mathcal{I}}_T = \tilde{\mathcal{I}}_k$ about c_k leads to a discontinuous treatment of these outcomes, results in Table 1 indicate that, at least when r is small, the density around $\tilde{S}_k = c_k$ is low. Alternatives, such as ordering by the MLE of θ, would avoid such discontinuities. However, p^+ would then depend on information levels beyond stage T, so the method would not extend to error spending versions of our designs.

Given an ordering of Ω, it is straightforward to create a $(1-\alpha)100\%$ confidence set for θ by inverting a family of hypothesis tests for each possible θ value. To test $H\colon\ \theta = \theta'$, we partition Ω into sets $R_{L,\theta'}$, $A_{\theta'}$ and $R_{U,\theta'}$ such that $(\tilde{\mathcal{I}}_{k_1}, \tilde{s}_{k_1}) \prec (\tilde{\mathcal{I}}_{k_2}, \tilde{s}_{k_2}) \prec (\tilde{\mathcal{I}}_{k_3}, \tilde{s}_{k_3})$ for outcomes $(\tilde{\mathcal{I}}_{k_1}, \tilde{s}_{k_1}) \in R_{L,\theta'}$, $(\tilde{\mathcal{I}}_{k_2}, \tilde{s}_{k_2}) \in A_{\theta'}$ and $(\tilde{\mathcal{I}}_{k_3}, \tilde{s}_{k_3}) \in R_{U,\theta'}$,

$$\mathbb{P}(R_{L,\theta'};\, \theta') = \mathbb{P}(R_{U,\theta'};\, \theta') = \alpha/2 \quad \text{and} \quad \mathbb{P}(A_{\theta'};\, \theta') = 1 - \alpha.$$

For each $\theta' \in \mathbb{R}$, a level α, two-tailed test of $H\colon\ \theta = \theta'$ accepts its null hypothesis for outcomes in $A_{\theta'}$. Inverting this family of tests yields a $(1-\alpha)100\%$ confidence set for θ. If the observed outcome is $(\tilde{\mathcal{I}}_{k^\star}, \tilde{s}^\star)$, the confidence set is

$$\{\, \theta\colon\ (\tilde{\mathcal{I}}_{k^\star}, \tilde{s}^\star) \in A_\theta \,\}. \tag{22}$$

If $\mathbb{P}\{\, (\tilde{\mathcal{I}}_T, \tilde{S}_T) \succeq (\tilde{\mathcal{I}}_{k^\star}, \tilde{s}^\star);\, \theta\}$ is monotone increasing in θ for each $(\tilde{\mathcal{I}}_{k^\star}, \tilde{s}^\star) \in \Omega$, the set (22) is the interval $(\theta_L,\, \theta_U)$ whose endpoints satisfy

$$\mathbb{P}\{\, (\tilde{\mathcal{I}}_T, \tilde{S}_T) \succeq (\tilde{\mathcal{I}}^\star_k, \tilde{s}^\star)\,;\, \theta_L) \,\} = \alpha/2 \quad \text{and} \quad \mathbb{P}\{\, (\tilde{\mathcal{I}}_T, \tilde{S}_T) \preceq (\tilde{\mathcal{I}}^\star_k, \tilde{s}^\star)\,;\, \theta_U) \,\} = \alpha/2. \tag{23}$$

We have checked a wide variety of optimal delayed response GSTs with $0 < r \leq 0.5$ and $2 \leq K \leq 10$ and found no evidence of non-monotonicity in these examples. However, if the monotonicity property does not hold, a conservative confidence interval can still be defined by evaluating $\xi(\theta) = \mathbb{P}\{\, (\tilde{\mathcal{I}}_T, \tilde{S}_T) \preceq (\tilde{\mathcal{I}}_{k^\star}, \tilde{s}^\star);\, \theta\}$ on a grid of θ values and taking the lowest solution for θ_L and the highest solution for θ_U in (23).

4.3. Methods of inference to deal with unexpected overrunning

Whitehead (1992) refers to the arrival of additional data after stopping according to a group sequential rule as "overrunning". In Section 1 we noted that Whitehead's (1992) deletion method for making inferences after a standard GST overruns is conservative and the methods proposed by Hall & Ding (2008) do not improve on the power achieved by the deletion method for constant amounts of overrunning. Hall & Liu (2002) order the sample space by the MLE of θ, so their method requires knowledge of unobserved future information levels. We shall draw on the methods of Sections 4.1 and 4.2 to create a new method for making decisions and inferences after a GST overruns.

Consider a standard one-sided GST with information level $\mathcal{I}_k$ and boundary values l_k and u_k at analyses $k = 1, \ldots, K$, where $l_K = u_K$. The information sequence and boundary

values are chosen to give type I error rate α under $\theta = 0$ and the design stipulates that the trial should stop at analysis k to accept H_0: $\theta \leq 0$ if $S_k \leq l_k$ and to reject H_0 if $S_k \geq u_k$. Suppose the trial stops at analysis $k^\star$ with $S_{k^\star} \leq l_{k^\star}$ or $S_{k^\star} \geq u_{k^\star}$ but additional responses are then observed. Although this eventuality had not been anticipated, it is required that these overrun data be included in a final analysis. We define $\tilde{\mathcal{I}}_{k^\star}$ as the information and $\tilde{S}_{k^\star}$ as the score statistic incorporating the overrun data. If we allow for the possibility of overrunning at analysis K, or otherwise set $\tilde{\mathcal{I}}_K = \mathcal{I}_K$ and $\tilde{S}_K = S_K$, we see the sample space has the same form as Ω defined in (19) when delayed responses are anticipated.

We are faced with the task of specifying a constant $c_{k^\star}$ such that H_0 is rejected if $\tilde{S}_{k^\star} \geq c_{k^\star}$ and accepted if $\tilde{S}_{k^\star} < c_{k^\star}$. Suppose first that $k^\star < K$. Since information levels $\tilde{\mathcal{I}}_1, \ldots, \tilde{\mathcal{I}}_{k^\star-1}$ involving overrun data that would have been observed after stopping at an earlier analysis are typically unknown, as are information levels at analyses $k > k^\star$, we find $c_{k^\star}$ by application of Method 1 for constructing error spending tests. Choosing $c_{k^\star}$ to satisfy (14) preserves the null probability of rejecting H_0 at stage $k^\star$. This methodology is valid if the amount of overrun data is unpredictable and $\tilde{\mathcal{I}}_k$ is regarded as a random variable, as long as the amount of additional data is not related to the value of $S_{k^\star}$. A simple adjustment is required if the trial terminates at analysis K with $\tilde{\mathcal{I}}_K > \mathcal{I}_K$. In this case, $c_{K^\star}$ is chosen so that $\mathbb{P}(S_1 \in C_1, \ldots, S_{K-1} \in C_{K-1}, \tilde{S}_K \geq c_{K^\star}; \theta = 0)$ is equal to the probability $\mathbb{P}(S_1 \in C_1, \ldots, S_{K-1} \in C_{K-1}, S_K \geq u_K; \theta = 0)$ under the original trial design.

We can use the same construction to define a p-value after a one-sided GST has overrun. Suppose first that each information level $\tilde{\mathcal{I}}_{k^\star}$ that would arise after stopping at interim analysis $k^\star$ with $S_{k^\star} \notin \mathcal{C}_{k^\star}$ and observing overrun data is known. We can compute critical values $c_{k^\star}$ satisfying (14) and combining these with the stopping boundary values l_k and u_k gives a delayed response GST of form (1). We now apply the stage-wise ordering defined in Section 4.2 to create the one-sided p-value for testing $H_0 : \theta = 0$ against $\theta > 0$ when $\tilde{S}_{k^\star} = s^\star$ after the GST stops at analysis $k^\star$. The equality of reversal probabilities stipulated by definition (14) implies the null probability of stopping at stage $k < k^\star$ with $S_k \notin C_k$ and then observing $\tilde{S}_k \geq c_k$ is equal to the probability that the original GST stops at stage k with $S_k \geq u_k$. Likewise, the null probability of stopping at stage $k < k^\star$ with $S_k \notin C_k$ and then observing $\tilde{S}_k \leq c_k$ is equal to the probability that the original GST stops at stage k with $S_k \leq l_k$. It follows that p^+ has no dependence on $\tilde{\mathcal{I}}_1, \ldots, \tilde{\mathcal{I}}_{k^\star-1}$ and $c_1, \ldots, c_{k^\star-1}$, so our earlier assumption that these information levels are known is unnecessary.

The above method makes assumptions about how inferences would have been made had the trial stopped at a different stage. It is important for methods of inference on termination of a GST to be specified in advance to avoid concerns that another method might have been chosen if this would have given a more impressive p-value. It is not difficult to explain how unexpected overrunning will be dealt with: protocols can include a standard statement that the above construction based on the stage-wise ordering will be used if there are overrun data. In the absence of overrunning this reduces to the usual stage-wise ordering for a standard GST. If the trial overruns unexpectedly and the scenario discussed in Example A arises, a p-value can be calculated enabling a clear interpretation of the trial data.

5. Incorporating data on a short-term endpoint

5.1. Formulation of GSTs incorporating data on a short-term endpoint

We have considered the case of a single, long-term endpoint measured after time Δ_t. Suppose that a correlated short-term endpoint is also observed at time $\Delta_{1,t}$. An example is

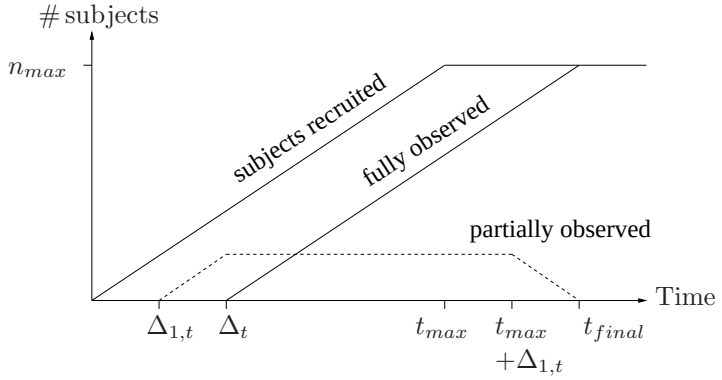


Fig. 7. Pattern of variation of numbers of fully and partially observed subjects during a delayed response GST with short-term and long-term endpoints. During the interval $(\Delta_t, t_{max} + \Delta_{1,t})$, the total number of partially observed subjects remains constant at $n_{max} r(1 - \Delta_{1,t}/\Delta_t)$.

the study in Example D, where the long-term endpoint is incidence of fracture within five years but the change in bone mineral density is also measured at one year.

Let $Y_{A,i}$ and $Y_{B,i}$, $i = 1, 2, \ldots$, denote the short-term responses and $X_{A,i}$ and $X_{B,i}$, $i = 1, 2, \ldots$, the long-term responses for subjects allocated to treatments A and B, respectively. Suppose responses on different subjects are independent and the pair of responses for subject i on treatment $T \in \{A, B\}$ follows a bivariate normal distribution

$$\begin{pmatrix} Y_{T,i} \\ X_{T,i} \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_{T,1} \\ \mu_{T,2} \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \tau\sigma_1\sigma_2 \\ \tau\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \right). \tag{24}$$

We assume for now that σ_1^2, σ_2^2 and τ are known but we shall show in Section 6 that it is possible to proceed adaptively, estimating these parameters during the course of a trial. Suppose we wish to test H_0: $\theta \leq 0$ against $\theta > 0$, where $\theta = \mu_{A,2} - \mu_{B,2}$, using a K-stage delayed response GST with type I error rate α at $\theta = 0$ and power $1 - \beta$ at $\theta = \delta$.

At each analysis, subjects are unobserved, partially observed (with just the short-term endpoint available), or fully observed (with both endpoints). Figure 7 shows how the numbers of partially and fully observed subjects develop as the trial proceeds. Let $\boldsymbol{\beta} = (\mu_{A,1}, \mu_{B,1}, \mu_{A,2}, \mu_{B,2})^T$, then at each interim analysis $k = 1, \ldots, K-1$, we fit the model (24) to all available data to obtain the maximum likelihood estimator $\hat{\boldsymbol{\beta}}_k$, and hence $\hat{\theta}_k$ and $\mathcal{I}_k = \{\mathrm{Var}(\hat{\theta}_k)\}^{-1}$. If recruitment is terminated at interim analysis k, we wait to observe long-term endpoints for all subjects, then re-fit model (24) to obtain $\tilde{\boldsymbol{\beta}}_k$, $\tilde{\theta}_k$ and $\tilde{\mathcal{I}}_k$. The sets of data used to obtain $\hat{\theta}_1, \ldots, \hat{\theta}_k, \tilde{\theta}_k$ are nested and it follows from the general theory of Jennison & Turnbull (1997) that these estimates have the standard joint distribution and the related score statistics follow the canonical joint distribution of Section 2.2. Thus, with $S_j = \hat{\theta}_j \mathcal{I}_j$, $j = 1, \ldots, k$, and $\tilde{S}_k = \tilde{\theta}_k \tilde{\mathcal{I}}_k$, the sequence $(S_1, \ldots, S_k, \tilde{S}_k)$ is multivariate normal with $S_j \sim N(\mathcal{I}_j\theta, \mathcal{I}_j)$, $j = 1, \ldots, k$, $\tilde{S}_k \sim N(\tilde{\mathcal{I}}_k\theta, \tilde{\mathcal{I}}_k)$ and independent increments. Since this distribution is exactly that arising for a single delayed response, GSTs of form (1) can be applied and their properties computed by the methods described previously.

Note that the final decision here concerns only the long-term endpoint. The short-term endpoint improves efficiency by increasing the information $\mathcal{I}_k$ for the long-term endpoint at interim analyses. While the short-term endpoint may itself be of clinical interest (as in

Example D), it is actually sufficient for it to be correlated with the long-term endpoint.

The approach extends to repeated measurements where a model can be fitted to all data on partially observed subjects at each interim analysis. The ASTIN stroke trial described by Grieve & Krams (2005) made similar use of a longitudinal model for intermediate data in implementing a response-adaptive dose allocation scheme. Fu & Manner (2010) describe uses of intermediate data in dose-finding studies with early stopping and adaptive treatment allocation. Use of repeated measurement data in GSTs has been considered by Galbraith & Becker (2001) and Galbraith & Marschner (2003). Their aim is to reduce study duration and, in their formulation, data collection ceases when the GST reaches a conclusion, so there is no "pipeline data" to wait for. As in our application, the benefit of short-term data is an increase in the precision of the estimate for the long-term endpoint.

5.2. Implementation of delayed response GSTs using short-term data

Let $n_{1,k}$ and $n_{2,k}$ denote the numbers of partially and fully observed cases out of the $\tilde{n}_k$ subjects recruited at interim analysis k, for $k = 1, \ldots, K-1$. For planning purposes, it is convenient to assume equal numbers of subjects are randomised to each treatment. Then, at interim analysis k, there are $n_{2,k}/2$ subjects on each treatment for whom both short-term and long-term responses are observed and a further $n_{1,k}/2$ subjects per treatment with just a short-term response. For a given recruitment rate c and planned interim analysis time t_k, the values of $n_{1,k}$, $n_{2,k}$ and $\tilde{n}_k$ can be predicted as

$$n_{1,k} = (\Delta_t - \Delta_{1,t})c, \quad n_{2,k} = (t_k - \Delta_t)c, \quad \text{and} \quad \tilde{n}_k = t_k\, c. \tag{25}$$

The full set of data at interim analysis k follows a normal linear model with parameter vector $\boldsymbol{\beta}$ and covariances are present between short-term and long-term responses for the same subject. Fitting this model by maximum likelihood, we find

$$\hat{\mu}_{A,1} - \hat{\mu}_{B,1} = \frac{2}{n_{2,k} + n_{1,k}} \sum_{i=1}^{(n_{2,k}+n_{1,k})/2} (Y_{A,i} - Y_{B,i}) \tag{26}$$

and

$$\hat{\theta}_k = \frac{2}{n_{2,k}} \sum_{i=1}^{n_{2,k}/2} (X_{A,i} - X_{B,i}) - \frac{\tau\sigma_2}{\sigma_1} \left[\frac{2}{n_{2,k}} \sum_{i=1}^{n_{2,k}/2} (Y_{A,i} - Y_{B,i}) - (\hat{\mu}_{A,1} - \hat{\mu}_{B,1}) \right]. \tag{27}$$

The information for θ at interim analysis k is

$$\mathcal{I}_k = \{\mathrm{Var}(\hat{\theta}_k)\}^{-1} = \left[\frac{4\sigma_2^2}{n_{2,k}} \left(1 - \tau^2 \frac{n_{1,k}}{(n_{1,k} + n_{2,k})} \right) \right]^{-1}. \tag{28}$$

If recruitment ceases at analysis k, short-term and long-term responses are observed for all pipeline subjects and the full data on the $\tilde{n}_k$ subjects at decision analysis k yield

$$\tilde{\theta}_k = \frac{2}{\tilde{n}_k} \sum_{i=1}^{\tilde{n}_k/2} (X_{A,i} - X_{B,i}), \tag{29}$$

and

$$\tilde{\mathcal{I}}_k = \{\mathrm{Var}(\tilde{\theta}_k)\}^{-1} = \frac{\tilde{n}_k}{4\,\sigma_2^2}. \tag{30}$$

Here, with long-term endpoints recorded for all subjects, short-term endpoints play no role in estimating θ.

Given a sequence of planned analysis times, we can substitute $n_{1,k}$, $n_{2,k}$ and $\tilde{n}_k$ from (25) into (28) and (30) to obtain the anticipated information levels $\mathcal{I}_k$ and $\tilde{\mathcal{I}}_k$. We could then use the methods of Section 2.4 to find the delayed response GST with type I error rate α and power $1-\beta$ at $\theta = \delta$ which minimises an expected sample size criterion.

Alternatively, we can use the error spending approach to implement a flexible design with given type I error rate and power. We plan such a design assuming a certain accrual rate and pattern of analysis times, for which (25), (28) and (30) give the sequence of information levels. For given $\mathcal{I}_{max}$, $f(t)$ and $g(t)$, we can apply, say, Method 2 of Section 4.1 to find the stopping and decision boundaries by solving (12), (14) and (16) for $k = 1, \ldots, K-1$ and (15) for $k = K$. We need to tailor this design to achieve the desired power $1-\beta$ at $\theta = \delta$. Keeping the analysis times fixed, we can search for the parameter ρ in the error spending functions (17) that leads to a final boundary point c_K satisfying (18). Or, if we fix ρ and assume analysis times will follow the pattern (9), we can search for the value of t_{max}, and hence n_{max} and $\mathcal{I}_{max}$, that will give a boundary satisfying (18).

5.3. *Efficiency of delayed response GSTs incorporating data on a short-term endpoint*

We illustrate the methods of this section in a testing problem with type I error rate $\alpha = 0.025$, power $1-\beta = 0.9$ at $\theta = \delta$, and inflation factor $R = 1.1$. The delay Δ_t in observing the long-term response leads to a delay parameter $r = \Delta_t / t_{max}$ and we define $\kappa = \Delta_{1,t}/\Delta_t$, so a small value of κ means the short-term endpoint is seen much more rapidly than the long-term endpoint. We suppose the $K-1$ interim analyses are scheduled to follow the pattern (9) with information levels given by(28) and (30). We have optimised delayed response GSTs incorporating a short-term endpoint for the objective function F.

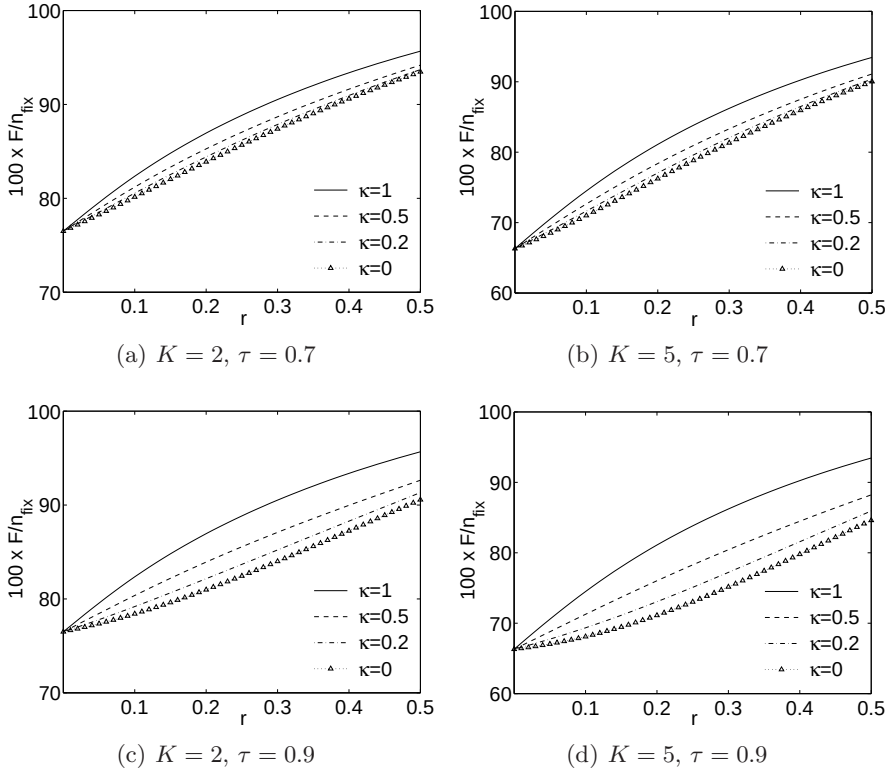
Figures 8(a) to 8(d) plot values of F for these optimised designs. Results for $\kappa = 1$ correspond to the case of no short-term endpoint, while $r = 0$ implies an immediate response. For r up to 0.2 or 0.3 say, using a short-term endpoint helps recover many of the benefits of early stopping associated with GSTs for an immediate response. Even for larger r, the benefits of interim monitoring are increased and group sequential testing becomes worthwhile where the results of Section 3.1 would have indicated the contrary.

In the bone fracture trial of Example D, the short-term endpoint is measured after one year and the long-term endpoint after five years, so $\kappa = 0.2$. If the study duration is expected to be 10 years in the absence of early stopping, then $r = 5/10 = 0.5$. Suppose a GST is planned with $K = 5$ analyses. If the endpoints are negatively correlated with correlation $\tau = -0.7$, a delayed response GST using both endpoints will have an average $\mathbb{E}(N;\theta)$, with weights as in F, equal to 90.4% of n_{fix}. This compares with 93.5% of n_{fix} if no short-term measurement is made. However, if the correlation is stronger, with $\tau = -0.9$, the average $\mathbb{E}(N;\theta)$ decreases further to 85.9% of n_{fix}. It is noteworthy that savings in sample size are possible with such a long delay in observing the primary endpoint.

The efficiency gained by incorporating data on the short-term endpoint stem from the effect on the information levels at interim analyses, which can be written as

$$\mathcal{I}_k = \frac{(n_{1,k} + n_{2,k})}{4\sigma_2^2} \left(1 + \frac{n_{1,k}}{n_{2,k}}(1-\tau^2)\right)^{-1}, \quad k = 1, \ldots, K-1. \tag{31}$$

For high values of τ^2, there can be a significant increase on the information level $n_{2,k}/(4\sigma_2^2)$ from long-term responses alone; as τ^2 tends to 1, $\mathcal{I}_k$ approaches $(n_{1,k} + n_{2,k})/(4\sigma_2^2)$, the



(a) $K = 2$, $\tau = 0.7$ (b) $K = 5$, $\tau = 0.7$

(c) $K = 2$, $\tau = 0.9$ (d) $K = 5$, $\tau = 0.9$

Fig. 8. Values of F, expressed as a percentage of n_{fix}, for delayed response GSTs incorporating measurements on a short-term endpoint. Tests have type I error rate $\alpha = 0.025$, power $1 - \beta = 0.9$ at $\theta = \delta$ and $R = 1.1$, and are designed to minimise F. The parameter $\kappa = \Delta_{1,t}/\Delta_t$ indicates how rapidly the short-term endpoint becomes available.

information if long-term responses are available for all $n_{1,k} + n_{2,k}$ subjects. As a guide, if we write $\mathcal{I}_k = (n_{2,k} + \xi\, n_{1,k})/(4\sigma_2^2)$, then for $\tau^2 = 0.5$ the value of ξ ranges from 0.5 to 0.33 as n_1/n_2 increases from zero to one, while ξ ranges from 0.8 to 0.67 for $\tau^2 = 0.8$.

Intuitively, the short-term response provides a prediction of a patient's long-term response which is, somehow, used in $\hat{\theta}_k$. The formulae in equations (26) and (27) offer a different explanation. We see there that the short-term responses provide an estimate of $\mu_{A,1} - \mu_{B,1}$ which is used in (27) to adjust the contribution of subjects with both short-term and long-term responses to the estimate of θ. This is most easily appreciated when $n_{1,k}$ is very large and $\hat{\mu}_{A,1} - \hat{\mu}_{B,1}$ can be regarded as equal to $\mu_{A,1} - \mu_{B,1}$. Then, (27) becomes

$$\hat{\theta}_k = \frac{2}{n_{2,k}} \sum_{i=1}^{n_{2,k}/2} (X_{A,i} - X_{B,i}) - \frac{\tau\sigma_2}{\sigma_1} \left[\frac{2}{n_{2,k}} \sum_{i=1}^{n_{2,k}/2} (Y_{A,i} - Y_{B,i}) - (\mu_{A,1} - \mu_{B,1}) \right],$$

the formula for an estimate of the mean of $X_{A,i} - X_{B,i}$ when correlated observations $Y_{A,i} - Y_{B,i}$ with known mean $\mu_{A,1} - \mu_{B,1}$ are also observed.

The examples of Figure 8 show the ideal short-term endpoint should be rapidly available and highly correlated with the primary endpoint. In practice, a balance will need to be

sought between these properties. Figure 8 shows that the effects of this trade-off on expected sample size are quite complex. Further gains may be made by measuring the short-term endpoint on several occasions, or even using more than one type of short-term endpoint.

A natural choice for a short-term endpoint is an early measurement of the endpoint of clinical interest, which may be made as a matter of course in many clinical trials. In closing, we re-iterate that the final testing decision concerns the primary, long-term endpoint and we do not assume the short-term endpoint to be a surrogate for this.

6. Dealing with unknown nuisance parameters

6.1. Information monitoring

So far, we have calculated testing boundaries assuming we know the values of nuisance parameters such as the response variance, σ^2, or correlation between short-term and long-term endpoints. However, this is unlikely to be the case in practice. For a single endpoint with unknown σ^2, internal pilot studies use a first stage estimate of σ^2 to update the sample size target; for a review of methods see Proschan (2009) and references therein. If early stopping to reject or accept the null hypothesis is also desired, the group sequential t-tests proposed by Denne & Jennison (2000) and Timmesfeld et al. (2007) combine sample size re-estimation with interim monitoring of the primary endpoint.

Mehta & Tsiatis (2001) adapt error spending designs to the case of unknown variance, monitoring accruing information rather than sample size. At each analysis, $\mathcal{I}_k$ is calculated using the current estimate of σ^2 and error spending boundaries for a sequence of Z-statistics are derived, based on the estimated information levels. These boundaries are expressed in terms of significance levels for testing the null hypothesis, which are applied to the t-statistics observed at interim analyses. When sampling continues, future group sizes are planned to reach the target information level assuming the current estimate of σ^2 to be the true value. This "significance level" approach for converting boundaries for Z-statistics to other applications, first proposed by Pocock (1977), maintains the marginal error probability at each analysis and Jennison & Turnbull (2000, Section 3.8) show this is an effective way to control the overall error rate. Mehta & Tsiatis (2001) show their tests control error probabilities at close to nominal levels in a range of scenarios.

We shall present "information monitoring" versions of our error spending designs for delayed responses with short-term and long-term endpoints. We allow the response variances σ_1^2 and σ_2^2 as well as the correlation coefficient τ to be unknown and show by simulation that error probabilities are accurately achieved.

6.2. Designing and implementing a trial with information monitoring

6.2.1. Specifying $\mathcal{I}_{max}$ and error spending functions $f(t)$ and $g(t)$

Consider a test of $H_0 : \theta \leq 0$ against $\theta > 0$ with type I error rate α at $\theta = 0$ and power $1 - \beta$ at $\theta = \delta$. Short-term and long-term measurements with distribution (24) are to be made on each patient, but σ_1^2, σ_2^2 and τ are unknown. In applying error spending designs for delayed responses, as described in Section 4.1, we specify error spending functions $f(t)$ and $g(t)$. Here t denotes the ratio $\mathcal{I}/\mathcal{I}_{max}$ but an estimate of $\mathcal{I}_k$ using current estimates of σ_2^2 and τ in (28) will be used in calculating the testing boundaries at each analysis k.

For design purposes, we follow the procedure for planning an error spending, delayed response GST described in Section 5.2. There, we assumed constant accrual at a certain rate c and a sequence of analysis times t_k. Now, we also need initial estimates $\sigma_{2,0}^2$ of σ_2^2

and τ_0 of τ to substitute into (28) and (30). The output from this planning stage is a target information level $\mathcal{I}_{max}$ and the two error spending functions $f(t)$ and $g(t)$. The information monitoring approach requires a commitment to reaching $\mathcal{I}_{max}$ unless the test stops with an early decision. Hence, $n_{max,0} = 4\,\sigma^2_{2,0}\,\mathcal{I}_{max}$ is the preliminary estimate of the total sample size that may be required and the time of the first interim analysis is chosen accordingly.

6.2.2. Estimating nuisance parameters and calculating test statistics

In the notation of Section 5, let $n_{1,k}$ and $n_{2,k}$ be the number of partially and fully observed cases out of the $\tilde{n}_k$ subjects recruited at interim analysis k, for $k = 1, \ldots, K-1$. For simplicity, we shall refer to the equations of Section 5.2 which assume equal numbers of responses on each treatment but we note the extension to the general case is straightforward.

Let $s^2_{1,k}$ denote the usual pooled estimator of σ_1^2 calculated from the $n_{1,k}+n_{2,k}$ subjects with short-term responses at interim analysis k. Likewise, let $s^2_{2,k}$ denote the pooled estimator of σ_2^2 from the $n_{2,k}$ subjects with long-term responses. Define $\hat{\tau}_{A,k}$ and $\hat{\tau}_{B,k}$ to be the Pearson product-moment correlation estimators based on the responses of the $n_{2,k}/2$ fully observed subjects on treatment arms A and B respectively. We follow Donner & Rosner (1980) in estimating the common correlation coefficient by the weighted average of $\hat{\tau}_{A,k}$ and $\hat{\tau}_{B,k}$ and denote this estimate by $\hat{\tau}_k$.

Let $\mathcal{I}_k(s^2_{2,k}, \hat{\tau}_k)$ be the estimate of the observed information obtained by substituting $s^2_{2,k}$ and $\hat{\tau}_k$ into (28). Define $\hat{\theta}_k(s^2_{1,k}, s^2_{2,k}, \hat{\tau}_k)$ to be the estimate of θ at interim analysis k calculated by substituting current estimates of the nuisance parameters into formula (27). The Wald statistic for testing H_0 at interim analysis k is

$$T_k = \hat{\theta}_k(s^2_{1,k}, s^2_{2,k}, \hat{\tau}_k)\sqrt{\mathcal{I}_k(s^2_{2,k}, \hat{\tau}_k)}.$$

Since $\hat{\theta}_k(s^2_{1,k}, s^2_{2,k}, \hat{\tau}_k)$ and $\mathcal{I}_k(s^2_{2,k}, \hat{\tau}_k)$ depend on $s^2_{1,k}$, $s^2_{2,k}$ and $\hat{\tau}_k$ in a complex way, T_k does not follow a standard distribution.

At decision analysis k, the long-term response is observed for all $\tilde{n}_k$ subjects. The estimated treatment effect $\tilde{\theta}_k$ is given by (29) and the observed information is estimated by $\tilde{\mathcal{I}}_k(\tilde{s}^2_{2,k}) = \tilde{n}_k/(4\tilde{s}^2_{2,k})$ where $\tilde{s}^2_{2,k}$ denotes the pooled estimate of σ_2^2 based on the $\tilde{n}_k$ long-term responses. Marginally, $\tilde{s}^2_{2,k} \sim \sigma_2^2\,\chi^2_{\tilde{n}_k-2}/(\tilde{n}_k-2)$ and the Wald statistic,

$$\tilde{T}_k = \tilde{\theta}_k\sqrt{\tilde{\mathcal{I}}_k(\tilde{s}^2_{2,k})}, \tag{32}$$

for testing H_0 at decision analysis k follows a $t_{\tilde{n}_k-2}$ distribution under H_0.

6.2.3. Calculating and applying error spending boundaries

We derive the stopping rule and decision rule for the error spending design following the methods described in Section 4.1 with pre-specified $f(t)$, $g(t)$ and $\mathcal{I}_{max}$ and using estimates $\mathcal{I}_k(s^2_{2,k}, \hat{\tau}_k)$ and $\tilde{\mathcal{I}}_k(\tilde{s}^2_{2,k})$ in place of $\mathcal{I}_k$ and $\tilde{\mathcal{I}}_k$. A parallel result to the canonical joint distribution of the score statistics S_k stated in Section 2.2 is that the distribution of standardised statistics $(T_1, \ldots, T_k, \tilde{T}_k)$ is approximately that of a sequence of Z-statistics $(Z_1, \ldots, Z_k, \tilde{Z}_k)$ based on accumulating data. Specifically, the Z-statistics are multivariate normal, each $Z_i \sim N(\theta\sqrt{\mathcal{I}_i}, 1)$ and $\mathrm{Cov}(Z_i, Z_j) = \sqrt{(\mathcal{I}_i/\mathcal{I}_j)}$ for $i < j$. These properties imply that the sequence $(Z_1, \ldots, Z_k, \tilde{Z}_k)$ is Markov.

We recommend following the construction in Method 2 of Section 4.1, but working on the Z-scale. Critical values for Z_k are then applied to the Wald statistic T_k. At a general interim analysis k, the cumulative type I and type II error probabilities are $f(t)$ and $g(t)$ evaluated at $t = \mathcal{I}_k(s^2_{2,k}, \hat{\tau}_k)/\mathcal{I}_{max}$. In calculating critical values for the Z-statistics at analysis k, we substitute estimates $\mathcal{I}_{k-1}(s^2_{2,k-1}, \hat{\tau}_{k-1})$ of $\mathcal{I}_{k-1}$, $\mathcal{I}_k(s^2_{2,k}, \hat{\tau}_k)$ of $\mathcal{I}_k$ and $\tilde{\mathcal{I}}_k(s^2_{2,k})$ of $\tilde{\mathcal{I}}_k$ into formulae for the conditional distribution of Z_k given Z_{k-1} and of $\tilde{Z}_k$ given Z_k. It remains to deal with the case where recruitment terminates at interim analysis k, either because $k = K$, the pre-specified maximum number of analyses, or because $\tilde{\mathcal{I}}_k(s^2_{2,k}) \geq \mathcal{I}_{max}$. Here, we set the critical value for $\tilde{Z}_k$ to spend all the remaining type I error probability. In carrying out this computation under $\theta = 0$, we use the conditional distribution of $\tilde{Z}_k$ given Z_{k-1}, which depends on $\tilde{\mathcal{I}}_k/\mathcal{I}_{k-1}$. This ratio is obtained from (28) with $k-1$ for k and (30) and, since σ_2^2 cancels, there is no need to substitute either estimate $s^2_{2,k-1}$ or $\tilde{s}^2_{2,k}$.

Since the estimators used for σ_1^2, σ_2^2 and τ are consistent, the joint distribution stated above holds asymptotically. For small sample sizes, we can adjust for this approximation through the "significance level" approach referred to in Section 6.1. As T_k does not follow a standard distribution, we simply compare T_k to the upper and lower critical values for Z_k. However, $\tilde{T}_k$ has a $t_{\tilde{n}_k-2}$ distribution under H_0 and we maintain the marginal probability that $\tilde{Z}_k \geq d_k$, say, when $\tilde{Z}_k \sim N(0, 1)$ by rejecting H_0 at decision analysis k if

$$\tilde{T}_k \geq t_{\tilde{n}_k-2,1-\Phi(d_k)},$$

where $t_{\nu,p}$ denotes the $100\,p$ percentile of the t-distribution with ν degrees of freedom.

Fluctuations in variance estimates can lead to $\mathcal{I}_k(s^2_{2,k}, \hat{\tau}_k) < \mathcal{I}_{k-1}(s^2_{2,k-1}, \hat{\tau}_{k-1})$. Jennison & Turnbull (2007) note such decreases in estimated information occur surprisingly often when monitoring a single immediate response. We follow their pragmatic solution and do not permit early stopping at an interim analysis when estimated information decreases. As explained above, our treatment of the analysis k at which recruitment terminates does not require an estimate of σ_2^2, so there is no risk of problems arising from $\tilde{s}^2_{2,k} > s^2_{2,k-1}$ and a decrease in estimated information between interim analysis $k-1$ and decision analysis k.

In an information monitoring design, interim analyses can be conducted at a sequence of calendar times until the target information level $\mathcal{I}_{max}$ is attained or until a time limit or maximum number of analyses is reached. One may also update the estimate of the sample size needed to achieve $\mathcal{I}_{max}$ and re-schedule interim analyses so as to reach this sample size after a certain number of analyses, K. The following example illustrates this approach.

6.3. Properties of information monitoring designs

We have simulated the information monitoring designs defined in Section 6.2 to assess their type I error rates and power. In our examples, procedures are designed to test $H_0 : \theta \leq 0$ against $\theta > 0$ with nominal type I error probability $\alpha = 0.025$ and power 0.9 at $\theta = \delta$. We use ρ-family error spending functions $f(t)$ and $g(t)$, as defined in (17), with $\rho = 2$ and schedule analyses so that the target information level $\mathcal{I}_{max}$ is reached at the fifth analysis. We assume a patient accrual rate c and take the ratio of delays to be $\kappa = \Delta_{1,t}/\Delta_t = 0.6$.

The maximum information level $\mathcal{I}_{max}$ is derived under initial estimates $\sigma^2_{2,0}$ and τ_0. Hence, we obtain preliminary values $n_{max,0} = 4\,\sigma^2_{2,0}\,\mathcal{I}_{max}$ for the maximum sample size and $t_{max,0} = n_{max,0}/c$ for the time needed to recruit this number of patients. With $r_0 = \Delta_t/t_{max,0}$ and anticipating $K-1$ interim analyses, equally spaced between Δ_t and

Table 4. Empirical type I error rates and power of information monitoring designs with nuisance parameters $\sigma_1^2 = 1$, $\sigma_2^2 = 2$ and $\tau = 0.7$. Initial estimates $\sigma_{2,0}^2 = 2.3$ and $\tau_0 = 0.6$ or 0.8 are used at the design stage. The delay ratio is $\kappa = 0.6$. Tests have a maximum of 5 stages and are designed to achieve type I error rate 0.025 and power 0.9 at $\theta = \delta$ under delay parameter λ_0. Results are based on 500,000 simulations.

		$\tau_0 = 0.6$			$\tau_0 = 0.8$		
	λ_0	$\mathcal{I}_{max}$	Type I error rate	Power	$\mathcal{I}_{max}$	Type I error rate	Power
$\delta = 0.5$	0.1	45.9	0.0259	0.8957	46.0	0.0260	0.8962
	0.2	45.3	0.0257	0.8949	45.5	0.0262	0.8969
	0.3	44.6	0.0256	0.8936	44.9	0.0253	0.8955
$\delta = 1.0$	0.1	11.5	0.0299	0.8803	11.5	0.0298	0.8809
	0.2	11.3	0.0282	0.8731	11.4	0.0278	0.8744
	0.3	11.1	0.0265	0.8640	11.2	0.0266	0.8661

$t_{max,0}$, we conduct the first interim analysis when

$$\tilde{n}_1 = \frac{n_{max,0}}{K}\{1 + (K-1)r_0\}$$

subjects have been recruited. At this point, there are $n_{1,1} = n_{max,0} r_0 (1-\kappa)$ partially observed and $n_{2,1} = n_{max,0}(1-r_0)/K$ fully observed cases. Before proceeding to the second interim analysis, the new estimate $s_{2,1}^2$ of σ_2^2 can be used to update the required maximum sample size and plan future analyses. In the general step, if recruitment continues beyond interim analysis k, the target sample size is updated to $n_{max,k} = 4\, s_{2,k}^2\, \mathcal{I}_{max}$ and the next interim analysis is performed after recruiting a total of

$$\tilde{n}_{k+1} = \tilde{n}_k + (n_{max,k} - \tilde{n}_k)/(K-k)$$

subjects. Since a maximum of 5 stages is specified, if the trial continues past interim analysis 4 we proceed directly to a decision analysis with $\tilde{n}_5 = n_{max,4} = 4\, s_{2,4}^2\, \mathcal{I}_{max}$ fully observed subjects and spend all the remaining type I error probability at this point.

Table 4 reports simulations under $\sigma_1^2 = 1$, $\sigma_2^2 = 2$ and $\tau = 0.7$, using initial estimates $\sigma_{2,0}^2 = 2.3$ and $\tau_0 = 0.6$ or 0.8. The delay in the long-term response is specified through the parameter $\lambda_0 = c\,\Delta_t / n_{fix,0}$ from which we can obtain $r_0 = \lambda_0\, \mathcal{I}_{fix}/\mathcal{I}_{max}$. At each interim analysis, the number of pipeline subjects with no long-term response is $\lambda_0\, n_{fix,0}$. Results are based on $500,000$ replicates so 95% reference ranges are $(0.0246, 0.0254)$ for estimates of a type I error rate of 0.025 and $(0.8992, 0.9008)$ for estimates of power equal to 0.9.

Table 4 shows attained type I error rates and power close to their intended values. There is a small inflation of type I error rate when $\delta = 0.5$, where $\mathcal{I}_{max} \approx 45$ and $n_{max} = 4\,\sigma_2^2\,\mathcal{I}_{max} \approx 360$, so there are reasonable numbers of observations for estimating σ_1^2, σ_2^2 and τ, even at the first few analyses. Lower sample sizes when $\delta = 1$, where $n_{max} \approx 90$, lead to higher increases in the type I error rate. These results agree with findings for sample size re-estimation procedures when a single response variance is estimated: reducing sample size when the current variance estimate is low and increasing it when the current estimate is high leads to a downward bias in the final variance estimate and an increase in the type I error rate; see, for example, Wittes et al. (1999) and Jennison & Turnbull (2007).

Tests achieve the desired power accurately, particularly for the higher sample sizes when $\delta = 0.5$. We can also report that expected sample sizes are close to those when the values of

Table 5. Values of F expressed as a percentage of n_{fix} attained by a standard GST optimised for the case of immediate response and optimal delayed response GSTs. Tests have type I error rate $\alpha = 0.025$, power $1 - \beta = 0.9$ at $\theta = \delta$ and $K = 3$ analyses with final information level $\mathcal{I}_{max} = 1.1\, I_{fix}$. The standard GST is designed and implemented for information levels $\mathcal{I}_k = (k/K)\, \mathcal{I}_{max}$, $k = 1, \ldots, K$. Delayed response tests are derived for this information sequence and also for the sequence given by (10).

r	Optimal standard GST with equally spaced $\mathcal{I}_k$	Optimal delayed response GST with equally spaced $\mathcal{I}_k$	Optimal delayed response GST based on information sequence (10)
0.01	71.2	71.2	71.2
0.1	78.2	78.0	77.7
0.2	86.0	84.5	83.5
0.3	93.9	89.9	88.0
0.4	99.2	93.6	91.5

σ_1^2, σ_2^2 and τ are known. In some cases, designs for the case of unknown nuisance parameters have lower expected sample size, attributable to some under-estimation of the sample size needed for the target information level $\mathcal{I}_{max}$ and slightly higher error rates.

We have conducted further simulations to investigate the separate effects of uncertainty about variance and correlation parameters. In repeating the simulations of Table 4 when τ is known, we obtained almost identical results. When we treated σ_1^2 and σ_2^2 as known but τ as unknown, differences between the observed type I error rates and the target of 0.025 were consistent with the sampling error for 500,000 replicates. We conclude that this methodology is robust to uncertainty about τ. Furthermore, in the context of a phase III trial, when sample sizes will usually be at least as large as those for cases with $\delta = 0.5$ in Table 4, we believe one can be confident the information monitoring approach will provide adequate control of the type I error rate and deliver power close to its target value.

7. Discussion and extensions

We have presented a new framework for group sequential testing when there is a delay in observing the primary response. This framework offers a systematic alternative to existing methods which Hall & Liu (2002, Section 4.4) speculate may sometimes be used in an ad hoc, and not necessarily appropriate, manner. Since our designs are optimised, they represent the most efficient option given the constraints in our formulation. We can re-cast the optimality property to state that our optimised delayed response GSTs maximise power at $\theta = \delta$ among all designs with the same stopping rule and type I error rate α. Thus, these designs resolve the question raised in Section 1 of how to incorporate data accrued after terminating a group sequential design. The probabilities that these designs "reverse" the decision anticipated on stopping recruitment are of the correct magnitude: higher values would imply some early stopping decisions are premature while lower values would indicate too conservative an approach with unnecessary delays in terminating recruitment.

One might ask how standard GSTs perform if pipeline data are disregarded. Column 1 in Table 5 shows average expected sample size, F, for GSTs designed for an immediate response but with responses observed after a delay. The three analyses are conducted after equal increments in the number of observed responses; pipeline data are ignored in deciding to reject or accept H_0 but counted in F. Column 2 gives results for delayed response GSTs

with the same information sequence, optimised for F. Results in column 3 are for optimised delayed response GSTs with analysis schedules of the form (10). The benefit gained from analysing the pipeline data under higher values of r is evident in lower values of F.

In planning a trial, designs with different numbers of analyses, K, can be compared to find the most appropriate choice (our numerical routines can handle values as high as $K = 200$). As K increases, group sequential testing approaches fully sequential monitoring. However, the benefit from each additional analysis decreases as K increases and values $K \leq 5$ are likely to be sufficient in practice; see, for example, the discussion in Section 4 of Eales & Jennison (1992) for the case of an immediate response.

Our proposed methods may be extended to other forms of hypothesis test and early stopping. A first step is to define a one-sided test of H_0: $\theta \leq 0$ against $\theta > 0$ with the intention to stop early only for efficacy. Such a test would have the form (1) but with $l_k = -\infty$ for all k. We shall refer to such a test as a one-sided delayed response GST with no futility boundary although, since the decision to accept H_0 if $S_k \geq u_k$ is followed by $\tilde{S}_k < c_k$ is binding, there is a small probability of early stopping for futility. It is not difficult to define such a test with the required type I error rate and optimisation should be possible for a suitably defined objective. An error spending version of such a test might also be defined, the challenge being to find a definition with robust efficiency.

A level α, two-sided test of H_0: $\theta = 0$ vs $\theta \neq 0$ can be formed by combining boundaries from two level $\alpha/2$ one-sided GSTs without futility boundaries, one of H_0: $\theta \leq 0$ vs $\theta > 0$, the other of H_0: $\theta \geq 0$ vs $\theta < 0$. This construction is in the spirit of Cox & Hinkley (1974) who consider a two-sided test as two one-sided tests. There is current interest in one-sided tests with non-binding futility boundaries, which protect the type I error rate even if a study may continue after the lower, futility boundary has been crossed; see, for example, Liu & Anderson (2008). Such a design can be created for the case of delayed response by adding a futility boundary $\{u_k; k = 1, \ldots, K-1\}$ to a level α one-sided GST with no futility boundary, testing H_0: $\theta \leq 0$ vs $\theta > 0$. In order to control type I error, observing $S_k \leq u_k$ would have to imply termination with acceptance of H_0 regardless of the value of $\tilde{S}_k$.

Our results quantify the difference between the expected sample size achievable when response is immediate and that attained when there is a delay in response. Understanding the origins of this efficiency loss can help design more efficient trials. The impact of delays due to data processing and cleaning indicates that any investment which speeds up this process is likely to be worthwhile. An option that we have not considered is for recruitment to be halted and then re-started: if the logistical difficulties in this approach could be overcome, there are interesting challenges in specifying such a design. It is important to remember the twin goals of low sample size and a rapid decision, particularly when the trial outcome is positive. Although our formulation in Section 3.2 of a problem with a combined objective may be somewhat idealised, the results in Figure 5 show this would be a useful option, meeting the combined objective well, even with a significant delay in response. We hope these results will spur discussion and prompt alternative problem formulations, with input from drug developers, ethicists and regulators.

We have demonstrated the benefits of using a short-term endpoint to improve efficiency when there is a delay in observing the primary endpoint. Stallard (2010) employs a short-term endpoint in a seamless phase II/III design both for treatment selection at the first analysis and in implementing a GST to test the effect of the selected treatment on the primary endpoint. Like Galbraith & Marschner (2003), Stallard does not formally include pipeline data in the GST, but suggests using Whitehead's (1992) method to adjust for data on the primary endpoint received after the GST terminates. It is not necessary here for the

short-term endpoint to be a surrogate for the long-term endpoint, only that responses on the two endpoints are correlated.

Our delayed response GSTs assume there is a single piece of information on the primary endpoint obtainable from each subject a certain time after treatment. In contrast, long-term survival studies continue to generate information as long as some subjects remain alive. It is common to apply standard GSTs in survival studies, with an informal treatment of information accruing after a decision is reached. Hampson (2008, Chapter 9) considers the design of such studies with the aim of minimising expected time to a conclusion; she finds that in many cases, forms of GST which give low expected sample size for other data types lead to efficient GSTs for survival data. Even in this setting, it may be desirable to treat a certain, well-defined set of information as "pipeline data". Suppose the database is locked at time t_k for cleaning and processing in advance of interim analysis k at time $t'_k > t_k$. If a stopping decision is made, one may wish to incorporate data from newly recruited subjects and further follow-up of existing subjects between times t_k and t'_k. Our delayed response GSTs could be adapted to this situation, making use of the sequential distribution theory for log-rank tests and Cox (1972) proportional hazards regression models provided by Jennison & Turnbull (1997) and Scharfstein et al. (1997).

We have treated the outcomes in Examples C and D of Section 1 as binary, but they could just as well be regarded as time-to-event data. Thus, in Example D, we might calculate the Kaplan-Meier estimates (Kaplan & Meier, 1958) of the distribution of time to a fracture and compare values at 5 years for treatment and control groups. The relevant sequential distribution theory is presented by Jennison & Turnbull (1985). This approach handles censoring caused by loss to follow-up. It also gives a simple way to use the partial information at the time of each interim analysis on subjects with less than five years of follow-up.

Acknowledgement

The first author was supported by the U.K. Engineering and Physical Sciences Research Council during this work. The authors thank Dr Simon Kirby for helpful discussions.

8. Appendix

8.1. Appendix 1: The backwards induction algorithm

In the Bayes decision problem of Section 2.4 let the random variable L indicate which of the three prior scenarios occurs with $\theta = 0$ when $L = 1$, $\theta = \delta$ when $L = 2$, and $\theta \sim N(\delta/2, (\delta/2)^2)$ when $L = 3$. Denote the prior distribution of L by π_L and set $\pi_L(1) = \pi_L(2) = \pi_L(3) = 1/3$. Since $S_k \sim N(\mathcal{I}_k\,\theta, \mathcal{I}_k)$, we have $S_k \sim N(0, \mathcal{I}_k)$ under $L = 1$, while $S_k \sim N(\mathcal{I}_k\,\delta, \mathcal{I}_k)$ for $L = 2$ and $S_k \sim N(\mathcal{I}_k\,\delta/2, \mathcal{I}_k + \mathcal{I}_k^2\,\delta^2/4)$ if $L = 3$. Let $h_{1,k}(s_k)$, $h_{2,k}(s_k)$ and $h_{3,k}(s_k)$ denote the densities of S_k under $L = 1$, 2 and 3, respectively. We denote the posterior distribution of L given $S_k = s_k$ by $\pi_L^{(k)}(l|s_k)$ and note that

$$\pi_L^{(k)}(l|s_k) \;=\; \frac{h_{l,k}(s_k)}{h_{1,k}(s_k) + h_{2,k}(s_k) + h_{3,k}(s_k)} \quad \text{for } l = 1,\, 2 \text{ and } 3.$$

Conditional on $L = 3$ and $S_k = s_k$, θ has the posterior distribution

$$\theta \sim N\left(\frac{s_k + 2/\delta}{\mathcal{I}_k + 4/\delta^2},\, (\mathcal{I}_k + 4/\delta^2)^{-1}\right).$$

In a similar manner, we denote the posterior distribution of L given $\tilde{S}_k = \tilde{s}_k$ at decision analysis k by $\tilde{\pi}_L^{(k)}(l|\tilde{s}_k)$ and note this can be found from the densities $\tilde{h}_{1,k}(\tilde{s}_k)$, $\tilde{h}_{2,k}(\tilde{s}_k)$ and $\tilde{h}_{3,k}(\tilde{s}_k)$ of $\tilde{S}_k$ under $L = 1$, 2 and 3, respectively.

In finding the Bayes procedure minimising F, we consider the expected additional cost at interim analyses and decision analyses. This cost includes the loss associated with an incorrect decision and the sampling cost of subjects who have not yet been recruited, but not the cost of sampling "pipeline" subjects as this relates to a commitment that has already been made. At decision analysis k, $1 \le k \le K$, with $\tilde{S}_k = \tilde{s}_k$, the optimal test has expected additional cost

$$\eta^{(k)}(\tilde{s}_k) = \min\{d_1\,\tilde{\pi}_L^{(k)}(1|\tilde{s}_k),\, d_0\,\tilde{\pi}_L^{(k)}(2|\tilde{s}_k)\}.$$

Now,

$$\frac{\tilde{\pi}_L^{(k)}(2|\tilde{s}_k)}{\tilde{\pi}_L^{(k)}(1|\tilde{s}_k)} = \frac{\tilde{h}_{2,k}(\tilde{s}_k)}{\tilde{h}_{1,k}(\tilde{s}_k)}$$

and by the monotone likelihood ratio property of the normal distribution there is a value c_k such that

$$d_1\,\tilde{\pi}_L^{(k)}(1|\tilde{s}_k) \;<\; d_0\,\tilde{\pi}_L^{(k)}(2|\tilde{s}_k) \quad \text{for } \tilde{s}_k > c_k$$

and

$$d_1\,\tilde{\pi}_L^{(k)}(1|\tilde{s}_k) \;>\; d_0\,\tilde{\pi}_L^{(k)}(2|\tilde{s}_k) \quad \text{for } \tilde{s}_k < c_k.$$

Hence, the optimal decision is to Reject H_0 if $\tilde{s}_k > c_k$, to Accept H_0 if $\tilde{s}_k < c_k$, and to make either decision if $\tilde{s}_k = c_k$.

At interim analysis k, $1 \le k \le K-1$, we define $\beta^{(k)}(s_k)$ to be the expected additional cost for continuing recruitment and proceeding optimally thereafter; we define $\rho^{(k)}(s_k)$ to be the expected additional cost if recruitment terminates at interim analysis k and the optimal decision is made at the ensuing decision analysis with responses from pipeline subjects. For each $k = 1, \ldots, K-2$, let $f_{k+1}(s_{k+1}|s_k)$ be the conditional density of S_{k+1} given $S_k = s_k$ and recruitment continues past interim analysis k. Also define $f_K(\tilde{s}_K|s_{K-1})$ to be the conditional density of $\tilde{S}_K$ given $S_{K-1} = s_{K-1}$ and recruitment continues past interim analysis $K-1$, leading to the final decision analysis K. For $k = 1, \ldots, K-1$, let $g_k(\tilde{s}_k|s_k)$ be the conditional density of $\tilde{s}_k$ at decision analysis k given $S_k = s_k$ and recruitment is terminated at interim analysis k.

From the definitions of $\rho^{(k)}(s_k)$ and $\eta^{(k)}(\tilde{s}_k)$, we have

$$\rho^{(k)}(s_k) = \int_{\tilde{s}_k} \eta^{(k)}(\tilde{s}_k)\, g_k(\tilde{s}_k|s_k)\, d\tilde{s}_k \quad \text{for } k = 1, \ldots, K-1.$$

Similarly,

$$\beta^{(K-1)}(s_{K-1}) = \pi_L^{(K-1)}(3|s_{K-1})\,(\tilde{n}_K - \tilde{n}_{K-1})\,c_0 \;+ \int_{\tilde{s}_K} \eta^{(K)}(\tilde{s}_K)\, f_K(\tilde{s}_K|s_{K-1})\, d\tilde{s}_K$$

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

$$\begin{aligned} \beta^{(k)}(s_k) \;&=\; \pi_L^{(k)}(3|s_k)\,(\tilde{n}_{k+1} - \tilde{n}_k)\,c_0 \\ &\quad + \int_{s_{k+1}} \min\{\,\beta^{(k+1)}(s_{k+1}),\, \rho^{(k+1)}(s_{k+1})\,\}\, f_{k+1}(s_{k+1}|s_k)\, ds_{k+1}. \end{aligned}$$

The preceding relationships allow derivation of the optimal Bayes procedure by backwards induction. The critical values $c_1, \ldots, c_K$ can be calculated first, and hence the functions $\eta^{(k)}(\tilde{s}_k)$ and $\rho^{(k)}(s_k)$. The functions $\beta^{(k)}(s_k)$ can be evaluated, starting with $k = K - 1$ and proceeding backwards to $k = 1$. Since the functions $\beta^{(k)}(s_k)$ and $\rho^{(k)}(s_k)$ appear in the integrands of other functions, they need to be evaluated on grids of points for use in a numerical integration routine. In the backward induction process, the critical values l_k and u_k defining the stopping and continuation regions of the optimal test at interim analysis k are found as the solutions to $\beta^{(k)}(s_k) = \rho^{(k)}(s_k)$.

8.2. Appendix 2: Uniqueness of the solution to the Bayes problem

THEOREM 2. *The Bayes problem defined by decision and sampling costs d_0, d_1 and c_0 has a unique solution up to actions on sets of Lebesgue measure zero.*

PROOF. We have seen in Appendix 1 that the optimal rule at decision analyses 1 to K is uniquely defined apart from a choice of action for $s_k = c_k$. Arbitrary changes may also be made on any set of measure zero as these will not alter the expected cost.

Following the reasoning of Brown et al. (1980, Theorem 3.3), we note that, for $k = 1, \ldots, K-1$, $\beta^{(k)}(s_k)$ and $\rho^{(k)}(s_k)$ are analytic functions of s_k. Therefore, by the result of Farrell (1968, Lemma 4.2), the set of values s_k for which $\beta^{(k)}(s_k) = \rho^{(k)}(s_k)$ is either of measure zero or equal to the whole real line. However, $\pi_L^{(k)}(3|s_k) \to 1$ as $s_k \to \infty$ and so $\beta^{(k)}(s_k) > \rho^{(k)}(s_k)$ for sufficiently large s_k. It follows that $\beta^{(k)}(s_k) = \rho^{(k)}(s_k)$ only on a set of measure zero and the result is proved. □

8.3. Appendix 3: Invariance of the minimum of F/n_{fix} to changes in δ and σ^2

THEOREM 3. *For fixed values of K, α, β, R and delay parameter r, minimum values of F/n_{fix} for delayed response GSTs with information levels following (10) are invariant to changes in the response variance σ^2 and the treatment effect δ at which power is specified.*

PROOF. Consider applying a delayed response GST of form (1) to the problem defined in Section 2.1. Denote by Problem 1 the case where the response variance is σ_1^2 and power $1 - \beta$ is specified at $\theta = \delta_1$. The information needed by a fixed sample size test is $\mathcal{I}_{1,fix} = \{\Phi^{-1}(1-\alpha) + \Phi^{-1}(1-\beta)\}^2/\delta_1^2$, which requires a sample size of $n_{1,fix} = 4\,\sigma_1^2\,\mathcal{I}_{1,fix}$ divided equally between the two treatments. A delayed response GST has $\mathcal{I}_{1,max} = R\,\mathcal{I}_{1,fix}$ and $n_{1,max} = 4\,\sigma_1^2\,\mathcal{I}_{1,max}$. Suppose the patient recruitment rate is c_1 and the delay in observing each response is $\Delta_{1,t}$, then $t_{1,max} = n_{1,max}/c_1$ and

$$r_1 = \frac{\Delta_{1,t}}{t_{1,max}} = \frac{\Delta_{1,t}\; c_1\; \delta_1^2}{4\,\sigma_1^2\,R\,\{\Phi^{-1}(1-\alpha) + \Phi^{-1}(1-\beta)\}^2}\,.$$

The values r_1 and $\mathcal{I}_{1,max}$ define the information levels $\mathcal{I}_{1,k}$ at interim analyses and $\tilde{\mathcal{I}}_{1,k}$ at decision analyses specified by equation (10). Let $\mathcal{I}_{1,(1)}, \ldots, \mathcal{I}_{1,(2K-1)}$ denote the values $\mathcal{I}_{1,1}, \tilde{\mathcal{I}}_{1,1}, \ldots, \mathcal{I}_{1,K-1}, \tilde{\mathcal{I}}_{1,K-1}, \tilde{\mathcal{I}}_{1,K}$ arranged in increasing order and let $S_{1,(k)}$ be the score statistic, an S_k or $\tilde{S}_k$, associated with $\mathcal{I}_{1,(k)}$.

The sequence $\{S_{1,(1)}, \ldots, S_{1,(2K-1)}\}$, is generated by an accumulating body of data, hence the results of Jennison & Turnbull (1997) imply this sequence has the canonical joint

distribution. Setting $Z_{1,(k)} = S_{1,(k)}/\sqrt{\mathcal{I}_{1,(k)}}$, $k = 1, \ldots, 2K-1$, it follows that the sequence $\{Z_{1,(1)}, \ldots, Z_{1,(2K-1)}\}$ is multivariate normal with

$$Z_{1,(k)} \sim N(\theta \sqrt{\mathcal{I}_{1,(k)}}, 1)$$

and $\mathrm{Cov}(Z_{1,(k_1)}, Z_{1,(k_2)}) = \sqrt{\{\mathcal{I}_{1,(k_1)}/\mathcal{I}_{1,(k_2)}\}}$ for $k_1 < k_2$.

Now consider Problem 2, with response variance σ_2^2, power $1-\beta$ set at $\theta = \delta_2$, recruitment rate c_2 and response delay $\Delta_{2,t}$. Define $\mathcal{I}_{2,fix}$, $n_{2,fix}$, $\mathcal{I}_{2,max}$, $n_{2,max}$, $t_{2,max}$ and r_2 in the analogous manner to Problem 1. If $r_2 = r_1$, scheduling analyses according to (10) generates the information levels of Problem 1 multiplied by δ_1^2/δ_2^2. Hence, the information levels $\mathcal{I}_{2,k}$ and $\tilde{\mathcal{I}}_{2,k}$ have the same ordering and the ordered sequence satisfies

$$\frac{\mathcal{I}_{2,(k)}}{\mathcal{I}_{1,(k)}} \;=\; \frac{\delta_1^2}{\delta_2^2}, \quad k = 1, \ldots, 2K-1.$$

Let $S_{2,(k)}$ be the score statistic associated with $\mathcal{I}_{2,(k)}$ and set $Z_{2,(k)} = S_{2,(k)}/\sqrt{\mathcal{I}_{2,(k)}}$, $k = 1, \ldots, 2K-1$. It is easily checked that the joint distribution of $\{Z_{2,(1)}, \ldots, Z_{2,(2K-1)}\}$ under $\theta = \xi\,\delta_2$ is the same as that of $\{Z_{1,(1)}, \ldots, Z_{1,(2K-1)}\}$ under $\theta = \xi\,\delta_1$.

Consider a delayed response GST for Problem 1 specified in terms of critical values for the standardised statistics $Z_{1,k}$ and $\tilde{Z}_{1,k}$. The same critical values can be applied to the $\tilde{Z}_{2,k}$ in Problem 2. In the cases $\theta = \xi\,\delta_1$ in Problem 1 and $\theta = \xi\,\delta_2$ in Problem 2, these two tests have the same distribution of stopping times and decisions to reject or accept H_0, but on termination at decision analysis k information levels are in the ratio δ_2^2/δ_1^2 and sample sizes in the ratio $(\sigma_1^2\,\delta_2^2)/(\sigma_2^2\,\delta_1^2)$. The criteria F in equation (3) can be written as

$$F = \int \mathbb{E}(N; \xi\,\delta)\; 2\;\phi(2\xi - 1)\; d\xi$$

and substituting $\delta = \delta_i$, $i = 1$ and 2, gives the criteria to be applied in Problems 1 and 2, respectively. Denote the values attained by tests for Problems 1 and 2 defined by a common set of critical values on the Z scale by F_1 and F_2. Then the above results imply

$$\frac{F_1}{n_{1,fix}} \;=\; \frac{F_2}{n_{2,fix}}.$$

If we have a delayed response GST minimising F_1 in Problem 1 we can apply this test, on the Z scale, to Problem 2 to attain the same value of F/n_{fix}, and vice versa. It follows that the minimum of $F_1/n_{1,fix}$ in Problem 1 and the minimum of $F_2/n_{2,fix}$ in Problem 2 are equal, and the result is proved. □

References

Anderson, T.W. (1964). Sequential analysis with delayed observations. *Journal of the American Statistical Association 59*, 1006–1015.

Armitage, P. (1957). Restricted sequential procedures. *Biometrika 44*, 9–56.

Armitage, P. (1958). Numerical studies in the sequential estimation of a binomial parameter. *Biometrika 45*, 1–15.

Armitage, P. (1975). *Sequential Medical Trials* (2nd ed.). Oxford: Blackwell.

Banerjee, A. and Tsiatis, A.A. (2006). Adaptive two-stage designs in phase II clinical trials. *Statistics in Medicine 25*, 3382–3395.

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

Bauer, P. and Einfalt, J. (2006). Application of adaptive designs - a review. *Biometrical Journal 48*, 493–506.

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

Brannath, W., Zuber, E., Branson, M., Bretz, F., Gallo, P., Posch, M., and Racine-Poon, A. (2009). Confirmatory adaptive designs with Bayesian decision tools for a targeted therapy in oncology. *Statistics in Medicine 28*, 1445–1463.

Brown, L.D., Cohen, A., and Strawderman, W.E. (1980). Complete classes for sequential tests of hypotheses. *Annals of Statistics 8*, 377–398.

Chang, M.N. and O'Brien, P.C. (1986). Confidence intervals following group sequential tests. *Controlled Clinical Trials 7*, 18–26.

Chen, Y.H.J., DeMets, D.L., and Lan, K.K.G. (2004). Increasing the sample size when the unblinded interim result is promising. *Statistics in Medicine 23*, 1023–1038.

Cox, D.R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society B 34*, 187– 220.

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

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

Denne, J.S. and Jennison, C. (2000). A group sequential t-test with updating of sample size. *Biometrika 87*, 125–134.

Donner, A. and Rosner, B. (1980). On inferences concerning a common correlation coefficient. *Applied Statistics 29*, 69–76.

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

Eales, J.D. and Jennison, C. (1995). Optimal two-sided group sequential tests. *Sequential Analysis 14*, 273–286.

Facey, K.M. (1992). A sequential procedure for a phase II efficacy trial in hypercholesterolemia. *Controlled Clinical Trials 13*, 122–133.

Fairbanks, K. and Madsen, R. (1982). P-values using a repeated significance test design. *Biometrika 69*, 69–74.

Faldum, A. and Hommel, G. (2007). Strategies for including patients recruited during interim analysis of clinical trials. *Journal of Biopharmaceutical Statistics 17*, 1211–1225.

Farrell, R.H. (1968). Towards a theory of generalized Bayes tests. *Annals of Mathematical Statistics 39*, 1–22.

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

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

Fu, H. and Manner, D. (2010). Bayesian adaptive dose-finding studies with delayed responses. *Journal of Biopharmaceutical Statistics 20*, 1055–1070.

Galbraith, S. and Marschner, I.C. (2003). Interim analysis of continuous long-term endpoints in clinical trials with longitudinal outcomes. *Statistics in Medicine 22*, 1787–1805.

Grieve, A.P. and Krams, M. (2005). ASTIN: A Bayesian adaptive dose-response trial in acute stroke. *Clinical Trials 2*, 340–351.

Hall, W.J. and Ding, K. (2008). Sequential tests and estimates after overrunning based on p-value combination. In *Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh*. Beachwood, Ohio: Institute of Mathematical Statistics.

Hall, W.J. and Liu, A. (2002). Sequential tests and estimators after overrunning based on maximum-likelihood ordering. *Biometrika 89*, 699–707.

Hampson, L.V. (2008). *Group Sequential Tests for Delayed Responses*. Ph. D. thesis, University of Bath.

Jennison, C. (1987). Efficient group sequential tests with unpredictable group sizes. *Biometrika 74*, 155–165.

Jennison, C. (1993). Numerical computations for group sequential tests. *Computing Science and Statistics 25*, 263–272.

Jennison, C. and Turnbull, B.W. (1983). Confidence intervals for a binomial parameter following a multi-stage test with application to MIL-STD 105D and medical trials. *Technometrics 25*, 49–58.

Jennison, C. and Turnbull, B.W. (1985). Repeated confidence intervals for the median survival time. *Biometrika 72*, 619–625.

Jennison, C. and Turnbull, B.W. (1997). Group sequential analysis incorporating covariate information. *Journal of the American Statistical Association 92*, 1330–1341.

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

Jennison, C. and Turnbull, B.W. (2006). Adaptive and nonadaptive group sequential tests. *Biometrika 93*, 1–21.

Jennison, C. and Turnbull, B.W. (2007). Adaptive seamless designs: Selection and prospective testing of hypotheses. *Journal of Biopharmaceutical Statistics 17*, 1135–1161.

Kaplan, E. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association 53*, 457–481.

Kiefer, J. and Weiss, L. (1957). Some properties of generalized sequential probability ratio tests. *Annals of Mathematical Statistics 28*, 57–74.

Lai, T.S. (1973). Optimal stopping and sequential tests which minimize the maximum sample size. *Annals of Statistics 1*, 659–673.

Lan, K.K.G. and DeMets, D.L. (1983). Discrete sequential boundaries for clinical trials. *Biometrics 70*, 659–663.

Liu, Q. and Anderson, K.M. (2008). On adaptive extensions of group sequential trials for clinical investigations. *Journal of the American Statistical Association 103*, 1621–1630.

Liu, Q., Anderson, K.M., and Pledger, G.W. (2004). Benefit-risk evaluation of multi-stage adaptive designs. *Sequential Analysis 23*, 317–331.

Lokhnygina, Y. and Tsiatis, A.A. (2008). Optimal two-stage group-sequential designs. *Journal of Statistical Planning and Inference 138*, 489–499.

Lorden, G. (1976). 2-SPRTs and the modified Kiefer-Weiss problem of minimising an expected sample size. *Annals of Statistics 4*, 281–291.

Marschner, I.C. and Becker, S.L. (2001). Interim monitoring of clinical trials based on long-term binary endpoints. *Statistics in Medicine 20*, 177–192.

Mehta, C.R. (2009). Adaptive design for confirmatory clinical trials. Cambridge, MA. Cytel Corporation and Harvard School of Public Health: Available at http://www.cytel.com/science-technology/publications.

Mehta, C.R. and Pocock, S.J. (2011). Adaptive increases in sample size when interim results are promising: A practical guide with examples. *Statistics in Medicine 30*, 3267–3284.

Mehta, C.R. and Tsiatis, A.A. (2001). Flexible sample size considerations using information based monitoring. *Drug Information Journal 35*, 1095–1112.

O'Brien, P.C. and Fleming, T.R. (1979). A multiple testing procedure for clinical trials. *Biometrics 35*, 549–556.

Öhrn, F. and Jennison, C. (2010). Optimal group sequential designs for simultaneous testing of superiority and non-inferiority. *Statistics in Medicine 29*, 743–759.

Pocock, S.J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika 64*, 191–199.

Proschan, M.A. (2009). Sample size re-estimation in clinical trials. *Biometrical Journal 51*, 348–357.

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

Proschan, M.A., Lan, K.K.G., and Wittes, J.T. (2006). *Statistical Monitoring of Clinical Trials: A Unified Approach*. New York: Springer.

Rosner, G.L. and Tsiatis, A.A. (1988). Exact confidence intervals following a group sequential trial: A comparison of methods. *Biometrika 75*, 723–729.

Scharfstein, D.O., Tsiatis, A.A., and Robins, J.M. (1997). Semiparametric efficiency and its implication on the design and analysis of group-sequential studies. *Journal of the American Statistical Association 92*, 1342–1350.

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

Siegmund, D. (1978). Estimation following sequential tests. *Biometrika 65*, 341–349.

Sooriyarachchi, M.R., Whitehead, J., Matsushita, T., Bolland, K., and Whitehead, A. (2003). Incorporating data received after a sequential trial has stopped into the final analysis: Implementation and comparison of methods. *Biometrics 59*, 701–709.

Stallard, N. (2010). A confirmatory seamless phase II/III clinical trial design incorporating short-term endpoint information. *Statistics in Medicine 29*, 959–971.

Timmesfeld, N., Schäfer, H., and Müller, H.-H. (2007). Increasing the sample size during clinical trials with t-distributed test statistics without inflating the type I error rate. *Statistics in Medicine 26*, 2449–2464.

Todd, S. and Stallard, N. (2005). A new clinical trial design combining phases 2 and 3: Sequential designs with treatment selection and a change of endpoint. *Drug Information Journal 39*, 109–118.

Tsiatis, A.A., Rosner, G.L., and Mehta, C.R. (1984). Exact confidence intervals following a group sequential test. *Biometrics 40*, 797–803.

Wald, A. (1947). *Sequential Analysis*. New York: Wiley.

Wald, A. and Wolfowitz, J. (1948). Optimum character of the sequential probability ratio test. *Annals of Mathematical Statistics 51*, 326–339.

Wassmer, G. and Vandemeulebroecke, M. (2006). A brief review on software developments for group sequential and adaptive designs. *Biometrical Journal 48*, 732–737.

Weiss, L. (1962). On sequential tests which minimise the maximum expected sample size. *Journal of the American Statistical Society 57*, 551–566.

Whitehead, J. (1992). Overrunning and underrunning in sequential clinical trials. *Controlled Clinical Trials 13*, 106–121.

Whitehead, J. (1993). Application of sequential methods to a phase III clinical trial in stroke. *Drug Information Journal 27*, 733–740.

Whitehead, J. (1997). *The Design and Analysis of Sequential Clinical Trials (Revised 2nd Edition)*. Chichester: Wiley.

Wittes, J., Schabenberger, O., Zucker, D., Brittain, E., and Proschan, M. (1999). Internal pilot studies 1: type I error rate of the naive t-test. *Statistics in Medicine 18*, 3481–3491.