Optimal group sequential designs for simultaneous testing of superiority and non-inferiority

Fredrik Öhrn
AstraZeneca R & D Mölndal,
SE-431 83 Mölndal, Sweden
and
Christopher Jennison
Department of Mathematical Sciences,
University of Bath, Bath, BA2 7AY U.K

November 5, 2009

Abstract

Confirmatory clinical trials comparing the efficacy of a new treatment with an active control typically aim at demonstrating either superiority or non-inferiority. In the latter case, the objective is to show that the experimental treatment is not worse than the active control by more than a pre-specified non-inferiority margin. We consider two classes of group sequential designs that combine the superiority and non-inferiority objectives: non-adaptive designs with fixed group sizes and adaptive designs where future group sizes may be based on the observed treatment effect. For both classes, we derive group sequential designs meeting error probability constraints which have the lowest possible expected sample size averaged over a set of values of the treatment effect. These optimised designs provide an efficient means of reducing expected sample size under a range of treatment effects, even when the separate objectives of proving superiority and non-inferiority would require quite different fixed sample sizes. We also present error spending versions of group sequential designs which are easily implementable and can handle unpredictable group sizes or information levels. We find the adaptive choice of group sizes to yield some modest efficiency gains; alternatively, expected sample size may be reduced by adding another interim analysis to a non-adaptive group sequential design.

KEY WORDS: adaptive re-design; clinical trial; decision theory; group sequential test; non-inferiority; superiority

1 Introduction

The primary objective in many clinical trials is to demonstrate superiority of the experimental treatment. With an active control treatment, it may also be of interest to show the experimental treatment is not worse than the control by more than a pre-specified margin. Proving "non-inferiority" is particularly appropriate if the new treatment is safer than the control.

Let θ denote the treatment difference between the new treatment and control, with positive values of θ indicating superiority of the new treatment. Superiority can be established by rejecting the null hypothesis $H_{S,0}$: $\theta \leq 0$ in favour of the alternative $\theta > 0$. Suppose it is agreed that the new treatment may be regarded as non-inferior if $\theta > -\delta_N$, where δ_N is a positive quantity referred to as the non-inferiority margin. We shall conclude the new treatment is non-inferior if the null hypothesis $H_{N,0}$: $\theta \leq -\delta_N$ is rejected in favour of $\theta > -\delta_N$.

Morikawa and Yoshida [1] note that tests for superiority and non-inferiority involve nested hypotheses and, hence, overall type I error probability will be controlled if both tests are conducted simultaneously without any adjustment for multiplicity. The same is true if the two hypotheses are tested group sequentially in a closed testing procedure [2]. However, the sample sizes required for tests of superiority and non-inferiority may be quite different. Suppose the test for non-inferiority is to have type I error probability α at $\theta = -\delta_N$ and power $1-\beta$ at $\theta = 0$, while the test for superiority has type I error probability α at $\theta = 0$ and power $1-\beta$ at $\theta = \delta_S$. The value of δ_N is typically set as a fraction of the estimated treatment difference in an earlier comparison of the active control treatment and placebo, and is liable to be quite small. The value of δ_S may be chosen to reflect expectations of a substantial treatment effect and when this is significantly larger than δ_N the sample size needed for the test of non-inferiority will be considerably higher than that required to test for superiority.

The need for different sample sizes to test the two hypotheses has led to quite complex proposals for group sequential designs testing both superiority and non-inferiority. Wang et al. [2] describe an adaptive group sequential procedure in which sample size is initially set for a test of superiority but, if interest shifts to showing non-inferiority, group sizes are increased. When this data-dependent change occurs, the type I error rate is preserved by down-weighting later groups of observations in the manner of Cui et al. [3].

In the two-stage procedures of Shih et al. [4] and Koyama et al. [5], first stage data are used to decide whether to continue and, if so, to select superiority or non-inferiority as the primary objective. The second stage sample size is chosen to give power for the chosen objective: Shih et al. [4] set sample size as a function of first stage data to achieve a given conditional power, while Koyama et al. [5] use a sample size function attaining a specified unconditional power.

Lai et al. [6] describe non-adaptive group sequential designs with fixed group sizes and three possible decisions on termination: superiority, non-inferiority (but not superiority) and inferiority. Reaching the third decision, inferiority, is sometimes referred to as stopping for futility since there is little prospect of reaching either positive decision. When δ_S is greater than δ_N, the study can

terminate at an early stage with a decision of superiority or inferiority then, later on, the options switch to non-inferiority and inferiority.

We shall present general classes of group sequential procedures which build on existing proposals. We first discuss designs with fixed group sizes, extending the options considered by Lai et al. [6] by allowing a choice of all three terminal decisions at each analysis. In our formulation of the testing problem in Section 2, values for δ_S and δ_N are stipulated along with a type I and type II error probability for each hypothesis test. For a given sequence of group sizes, we derive designs with the lowest possible expected sample sizes averaged over a range of values of the treatment effect, θ, while meeting the error probability constraints. Although group sizes are fixed, these procedures still exhibit a form of adaptation: when δ_S is significantly greater than δ_N, the upper continuation region for testing between superiority and non-inferiority comes to an end first, while the lower region continues to allow differentiation between non-inferiority and inferiority.

In Section 3 we generalise these designs to let group sizes depend on previously observed data. The resulting class includes the adaptive group sequential designs of Wang et al. [2] and the adaptive two-stage procedures of Shih et al. [4] and Koyama et al. [5]. In the two-decision problem of a one-sided test for superiority, Jennison and Turnbull [7] found adaptive choice of group sizes provided only a slight efficiency gain over non-adaptive designs. In our three-decision problem, when different fixed sample sizes are appropriate to the two separate hypothesis tests, it seems plausible there could be more substantial gains from using interim data both to choose the null hypothesis on which to focus and to adjust sample size accordingly. We assess previously proposed designs and new, optimised two-stage procedures to investigate the reduction in expected sample size that can be achieved by such adaptation. Our conclusion from the examples we have studied is that little is gained by choosing the second group size based on the observed treatment effect.

2 Optimal non-adaptive designs

2.1 Framework

Suppose observations X_{Aj} and X_{Bj}, $j = 1, 2, \ldots$, on treatments A and B respectively are independent and normally distributed with $X_{Aj} \sim N(\mu_A, \sigma^2)$ and $X_{Bj} \sim N(\mu_B, \sigma^2)$. We assume for now that σ^2 is known but we shall explain in Section 2.5 how unknown variance can be handled. The parameter of interest is the treatment effect $\theta = \mu_A - \mu_B$. We wish to test simultaneously $H_{N,0}$: $\theta \leq -\delta_N$ against $\theta > -\delta_N$ and $H_{S,0}$: $\theta \leq 0$ against $\theta > 0$, where the non-inferiority margin δ_N is established prior to the start of the trial.

Gould [8] and Koyama et al. [5] recognise this is a three-decision problem with outcomes: superiority, non-inferiority (only) and inferiority. Error rate requirements, including power for the two hypothesis tests at $\theta = 0$ and $\theta = \delta_S$, can be expressed through a pair of power curves. The curves displayed in

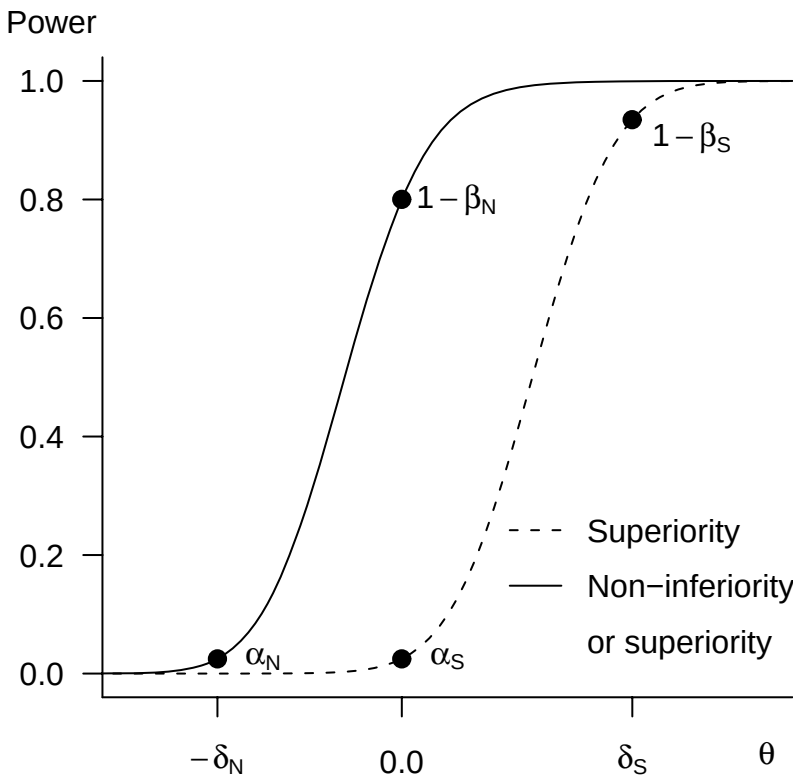

Figure 1: *Power curves for "non-inferiority or superiority" and superiority*

Figure 1 show the probabilities of concluding "Non-inferiority or Superiority" or "Superiority" as functions of θ. Formally, we specify type I and type II error rates α_N and β_N for testing $H_{N,0}$ and error rates α_S and β_S for testing $H_{S,0}$ as:

$$P_{\theta=-\delta_N}(\text{Declare "Non-inferiority" or "Superiority"}) = \alpha_N, \quad (1)$$

$$P_{\theta=0}(\text{Declare "Superiority"}) = \alpha_S, \quad (2)$$

$$P_{\theta=0}(\text{Conclude "Inferiority"}) = \beta_N, \quad (3)$$

$$P_{\theta=\delta_S}(\text{Conclude "Inferiority" or "Non-inferiority"}) = \beta_S. \quad (4)$$

In Appendix I we prove these conditions imply control of type I error for $H_{N,0}$ and $H_{S,0}$ over all values $\theta \leq -\delta_N$ and $\theta \leq 0$, respectively, and of type II error over $\theta \geq 0$ and $\theta \geq \delta_S$. This result holds for all non-adaptive designs satisfying certain minimal criteria, and it also applies to the adaptive designs we shall introduce in Section 3.

For many designs, fixing two points on the power curve results in the whole curve being indistinguishable from that of a fixed sample test. Hence, we do

not consider power under other parameter values when comparing designs. The exceptions are some adaptive designs for which power approaches unity rather slowly: see, for example, the power curves in Figure 11.

If the tests of the two null hypotheses were carried out in separate fixed sample trials, the number of observations per treatment required would be

$$n_{Nf} = 2\{\Phi^{-1}(1-\alpha_N) + \Phi^{-1}(1-\beta_N)\}^2\sigma^2/\delta_N^2$$

for testing $H_{N,0}$ and

$$n_{Sf} = 2\{\Phi^{-1}(1-\alpha_S) + \Phi^{-1}(1-\beta_S)\}^2\sigma^2/\delta_S^2$$

for testing $H_{S,0}$, where Φ is the cumulative distribution function of a standard normal variate.

We shall consider group sequential procedures with a maximum of K analyses, denoting the cumulative sample size per treatment at analysis k by n_k and the maximum sample size per treatment by $n_{max} = n_K$. Let Z_k be the standardised test statistic for testing $\theta = 0$ at analysis k. Allowing early stopping for all possible decisions at each analysis leads to a rule at analysis k of the form:

if $Z_k \leq a_k$,	stop and conclude inferiority,
if $a_k < Z_k < b_k$,	continue sampling,
if $b_k \leq Z_k \leq c_k$,	stop and declare non-inferiority,
if $c_k < Z_k < d_k$,	continue sampling,
if $Z_k \geq d_k$,	stop and declare superiority.

Here, $a_k \leq b_k \leq c_k \leq d_k$ and termination by analysis K is ensured by setting $a_K = b_K$ and $c_K = d_K$. When $n_{Sf} < n_{Nf}$, we may have $c_k = d_k$ in later stages so the upper continuation region is not present. In such cases, we denote the first value of k at which $c_k = d_k$ by K_S and the planned group size at this analysis by $n_{max,S}$. Although the lower boundary a_k, is present throughout, it can be helpful to think of the design as focusing on the test for superiority up to analysis K_S and concentrating on the choice between non-inferiority and inferiority thereafter. In order to meet the error probability constraints, it will be necessary for $n_K = n_{max}$ to be greater than n_{Nf} and $n_{K_S} = n_{max,S}$ to be greater than n_{Sf}. We shall refer to the ratios

$$r_S = n_{max,S}/n_{Sf} \quad \text{and} \quad r_N = n_{max}/n_{Nf}$$

as "inflation factors" and use these to indicate how much the maximum sample size, for the first phase or the whole design, has been increased beyond the minimum requirement.

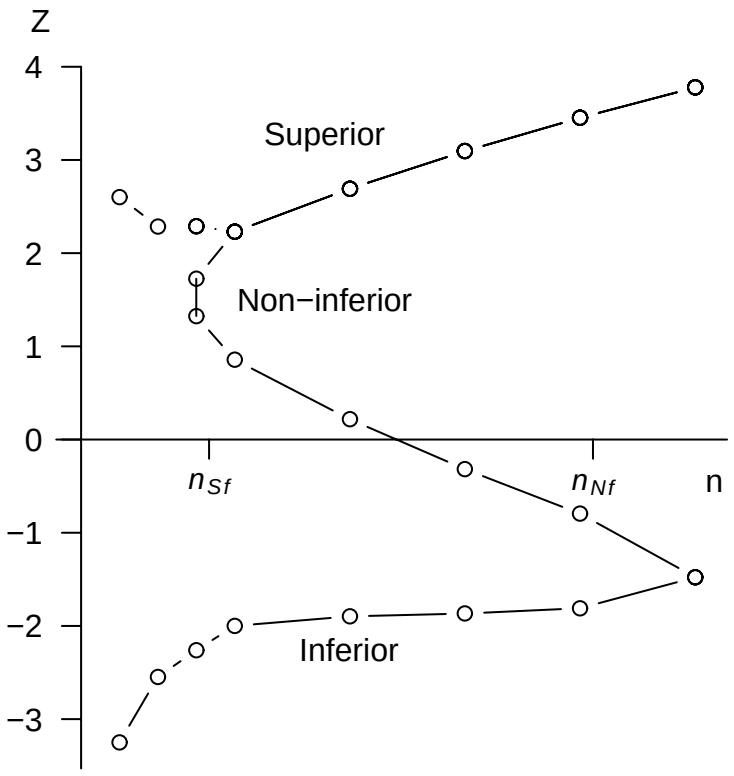

Figure 2: *Stopping boundaries for inferiority, non-inferiority and superiority*

A typical rule is illustrated in Figure 2. In this example, $b_k = c_k$ for $k = 1$ and 2 so early stopping for non-inferiority is not possible at the first two analyses and the two sections of continuation region merge into one. Since $c_k = d_k$ for $k = 4$ to 8, there is no upper continuation region at these analyses. However, we still allow the possibility to stop with a conclusion of superiority if the last group of observations results in a sufficiently high value of Z_5, Z_6, Z_7 or Z_8. The boundaries in Figure 2 are those of an optimal design which we shall describe in Section 2.2. They are broadly similar to the two-sided tests with an inner wedge described by Jennison and Turnbull [9, Chapter 6], however, they lack the symmetry of those designs around $\theta = 0$ and the roles of two of the error probabilities, α_N and β_N, are reversed.

The boundary points $a_1, b_1, c_1, d_1, \ldots, a_K, b_K, c_K, d_K$ must be chosen to satisfy the error constraints (1) to (4). A fixed sample size trial can only meet all four constraints simultaneously if $n_{Nf} = n_{Sf}$. In contrast, the additional degrees of freedom of a group sequential design allow suitable boundaries to be found as long as $n_{max} \geq \max(n_{Nf}, n_{Sf})$. Moreover, we can exploit the

remaining degrees of freedom to find a trial design with low expected sample size under specified values of θ. We have developed methods to find group sequential designs that minimise criteria of the form $\sum_{i=1}^{m} w_i E_{\theta_i}(N)$, where N denotes the sample size per treatment on termination. We have studied designs for a variety of optimality criteria, but in this paper we shall focus on

$$F^* = \{E_{-\delta_N}(N) + E_{-\delta_N/2}(N) + E_0(N) + E_{\delta_S/2}(N) + E_{\delta_S}(N)\}/5 \qquad (5)$$

which combines performance across the range of effect sizes of interest. We shall illustrate these procedures with an example for particular design parameters in Section 2.3 and make an efficiency comparison with the design of Lai et al. [6] in Section 2.4.

2.2 Derivation of Optimal Designs

Our methods enable us to find an optimal design with a specified number of analyses K and cumulative sample sizes $n_1, \ldots, n_K$. Comparing these optimal designs for different sequences $n_1, \ldots, n_K$ can inform the choice of suitable group sizes. Increasing K will decrease expected sample size but at the cost of more interim analyses, so comparing optimal designs for different values of K helps assess their costs and benefits.

Our derivation of optimal designs extends the methods of Eales and Jennison [10], Eales and Jennison [11], and Barber and Jennison [12] to the asymmetric three-decision problem. Given K and $n_1, \ldots, n_K$, we seek the stopping boundary minimising $\sum_{i=1}^{m} w_i E_{\theta_i}(N)$ subject to error probability requirements (1) to (4). We follow a Lagrangian approach and introduce the unconstrained problem of minimising

$$\sum_{i=1}^{m} w_i E_{\theta_i}(N) + \lambda_1 P_1 + \lambda_2 P_2 + \lambda_3 P_3 + \lambda_4 P_4, \qquad (6)$$

where λ_1 to λ_4 are positive and P_1 to P_4 denote the left hand sides of equations (1) to (4). The design minimising (6) must have the minimum value of $\sum_{i=1}^{m} w_i E_{\theta_i}(N)$ among all designs with the same P_1 to P_4. Hence, choosing Lagrange multipliers λ_1 to λ_4 so that the solution has $P_1 = \alpha_N$, $P_2 = \alpha_S$, $P_3 = \beta_N$ and $P_4 = \beta_S$ solves the original constrained problem.

For given λ_1 to λ_4, the method of dynamic programming can be used to minimise (6) quickly and accurately. This minimisation problem also has an interpretation as a Bayes sequential decision problem with a certain combination of prior on θ, costs for incorrect decisions, and sampling costs under the θ_is appearing in $\sum_{i=1}^{m} w_i E_{\theta_i}(N)$. This Bayesian interpretation provides insight into the dynamic programming solution where it is seen that decisions at each stage are based on expected future costs under the current posterior distribution for θ. Further details of the derivation of optimal designs are provided in Appendix II.

2.3 Numerical example

Suppose a non-inferiority margin of $\delta_N = 0.1$ has been established, power for the test of superiority is set at $\delta_S = 0.2$, and error probabilities are $\alpha_N = 0.025$, $\alpha_S = 0.025$, $\beta_N = 0.1$, and $\beta_S = 0.1$. If the response variance is $\sigma^2 = 0.5$, the fixed sample size per treatment for the test of superiority alone is $n_{Sf} = 263$, while the test for non-inferiority needs $n_{Nf} = 1051$.

We first consider a design with $K = 8$ analyses. The maximum sample size, n_{max}, must be at least a little greater than the larger of n_{Sf} and n_{Nf}, so we choose $n_{max} = 1.2\, n_{Nf} = 1261$. We set $n_4 = n_{max,S} = 1.2\, n_{Sf} = 316$ so a conclusion about the superiority objective can be reached in the first four analyses, leaving analyses 5 to 8 to concentrate on testing between non-inferiority and inferiority. Taking equal group sizes either side of analysis 4, we have $n_k = (k/4)n_4$ for $k = 1, \ldots, 3$, and $n_k = n_4 + ((k-4)/4)(n_{max} - n_4)$ for $k = 5, \ldots, 8$.

Optimising the design for F^* yields the boundary values a_1, b_1, c_1, $d_1, \ldots, a_8$, b_8, c_8, d_8 shown earlier in Figure 2. The absence of an inner wedge at the first two analyses indicates it is not possible to stop early for non-inferiority in this optimal design. The form of the upper part of the stopping boundary at later analyses merits some comment. Since optimisation has produced $c_k = d_k$ for $k = 4$ to 8, there is no upper continuation region after analysis 4; this is in line with our intent to deal with the issue of superiority by this analysis. The presence of values for d_5 to d_8 shows it is still possible to decide in favour of superiority at a later analysis if the last group of observations produces a large increase in the Z-statistic and the study ends with $Z_k > d_k$. In fact, such a sequence of Z_ks is unlikely under any value of θ and the dramatic change in observed treatment effect in the final group needed to achieve this might well raise questions about heterogeneity of the treatment effect over time. Let K_S denote the index of the first analysis at which $c_k = d_k$ in an optimal design. We have found that setting $c_k = d_k = \infty$ for $k > K_S$, so only the test between non-inferiority and inferiority is pursued at analyses $k = K_S + 1, \ldots, K$, has a negligible impact on error probabilities and, hence, on efficiency. One may, therefore, choose to remove the option of a decision in favour of superiority after the analysis at which values of c_k and d_k first converge. This will be the case in our definition of error spending designs in Section 2.5. However, unless otherwise stated we shall retain the option of stopping for superiority, and finite values for the d_k, in the optimal designs we report.

The optimal design's expected sample size is shown as a function of θ by the solid line in Figure 3. The two horizontal lines at n_{Nf} and n_{Sf} aid comparison with the sample sizes of the individual, fixed sample tests for non-inferiority and superiority. The sequential design is clearly effective in reducing expected sample size below n_{Nf}. Since the fixed sample size, n_{Sf}, in the individual test for superiority is insufficient to provide the desired power for the test of non-inferiority, it is to be expected that the sequential design has expected sample size greater than n_{Sf} at low values of θ. However, at high values of θ, where the main task is to distinguish between superiority and non-inferiority, $E_\theta(N)$

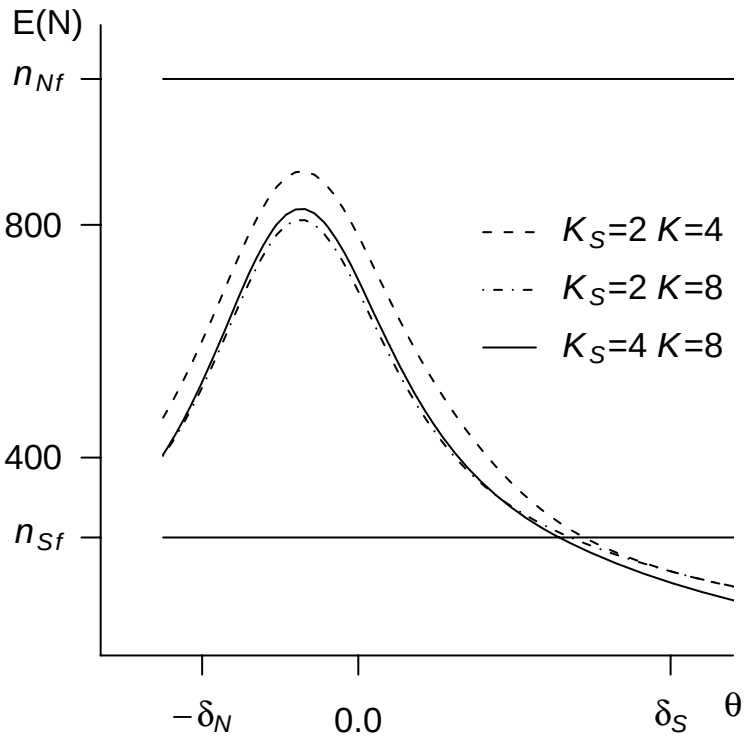

Figure 3: *Expected sample size functions for optimal designs with 4 and 8 analyses*

does fall below n_{Sf}. Additional curves in Figure 3 show the expected sample size function for two more optimal designs, one with $K = 4$ analyses and one with $K = 8$ analyses. For both these designs, $K_S = 2$, $n_2 = 1.1\, n_{Sf} = 289$, and the remaining $K - 2$ analyses are equally spaced up to $n_K = 1261$. With a total of 8 analyses, reducing K_S from 4 to 2 reduces F^* by about 1%. However, the design with $K_S = 4$ performs better for large values of θ. The efficiency gained by increasing the total number of analyses from 4 to 8 is close to 10%, a larger improvement than is typically seen in one-sided group-sequential tests of $\theta \leq 0$ against $\theta > 0$. This can be attributed to the fact that some analyses are well placed for one testing objective but poorly placed for the other, so the "effective" number of analyses for testing each individual hypothesis is less than K. We conclude that when the ratio n_{Nf}/n_{Sf} is as high as 4, designs with only a small number of groups may not achieve the full benefits of sequential monitoring.

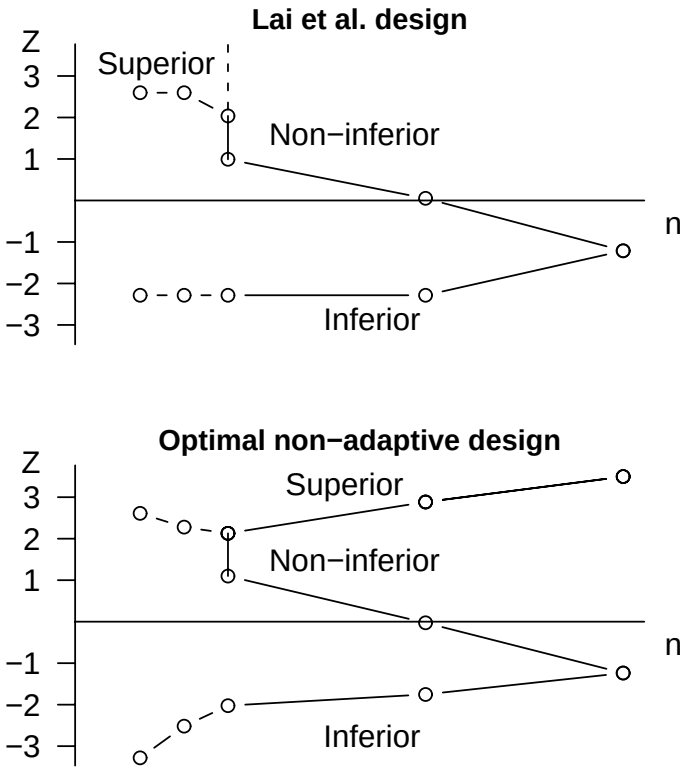

Figure 4: *Critical values for a Lai et al. 5-group design and an optimal 5-group design*

2.4 Comparison with the method of Lai et al.

Lai et al. [6] propose a group sequential procedure with K analyses that switches objective at a certain analysis. For $n_{Nf} > n_{Sf}$, the procedure allows early stopping for superiority at analyses $1, \ldots, K_S$ and for non-inferiority at analyses $K_S, \ldots, K$; the test can stop for the negative decision of inferiority at any point. The authors present their method in terms of generalized likelihood ratio statistics but we shall define critical values on the Z scale. Figure 4 displays a 5-group Lai et al. procedure with $K_S = 3$. Note that decisions of non-inferiority (solid line) and superiority (dashed line) are both possible at analysis 3. The authors define a parameter ϵ governing the amount of early stopping and recommend using $\epsilon = 1/3$. The outer boundaries have constant critical values on the Z scale so, in our notation, $d_1 = \ldots = d_{K_S-1} = \tilde{d}$ and $a_1 = \ldots = a_{K-1} = \tilde{a}$. For the non-inferiority boundary, $b_k = \tilde{b} - \delta_N \sqrt{\{n_k/(2\,\sigma^2)\}}$ for

$k = K_S, \ldots, K-1$. The values of $\tilde{a}$, a_K, $\tilde{b}$, $\tilde{d}$ and d_{K_S} are chosen so that: the probability under $\theta = 0$ of stopping to declare superiority by analysis $K_S - 1$ is $\epsilon\,\alpha_S$ and by analysis K_S is α_S; the probability under $\theta = -\delta_N$ of stopping to declare non-inferiority or superiority by analysis $K-1$ is $\epsilon\,\alpha_N$ and by analysis K is α_N; the probability under $\theta = 0$ of stopping to conclude inferiority by analysis $K-1$ is $\epsilon\,\beta_N$.

This construction does not produce specific type II error probabilities, instead these are determined by n_K, n_{K_S} and ϵ. There is a demarcation at analysis K_S, with stopping for superiority only possible at analyses 1 to K_S, and stopping for non-inferiority only at analyses K_S to K. The framework of Section 2 imposes no such constraints and we allow an inner wedge for non-inferiority before K_S and a continuing superiority/non-inferiority boundary thereafter.

We have applied the method of Lai et al. to the numerical example of Section 2.3 in which $\delta_N = 0.1$, $\delta_S = 0.2$, $\sigma^2 = 0.5$ and $\alpha_N = \alpha_S = 0.025$. We set $K = 5$, $K_S = 3$ and $\epsilon = 1/3$, with $n_{max,S} = 263$ and $n_{max} = 1051$, the values that would give 90% power if the two testing objectives were addressed in fixed sample trials. This implies $n_1 = 88$, $n_2 = 175$, $n_3 = n_{max,S} = 263$, $n_4 = 657$, and $n_5 = n_{max} = 1051$. It is the resulting boundaries that are shown in the upper panel of Figure 4. This design has type II error probabilities $\beta_N = 0.125$ and $\beta_S = 0.11$. Using these numbers to define fixed sample sizes n_{Sf} and n_{Nf}, we find the inflation factors for the Lai et al. design are $r_S = 1.035$ and $r_N = 1.086$.

We compared the Lai et al. design with a 5-group sequential design with the same group sizes and attained error probabilities, optimised for F^*. This optimised design with $K_S = 3$, $r_S = 1.035$ and $r_N = 1.086$ is depicted in the lower panel of Figure 4 and its expected sample size function is shown in Figure 5: the value of F^* is about 4% lower than that of the Lai et al. design. However, the inflation factor $r_S = 1.035$ is rather low and there is no obvious need to restrict $n_{max,S}$, given that higher sample sizes are allowed if the study continues to later analyses. Keeping $K_S = 3$ and increasing r_S from 1.035 to 1.2 while maintaining the same overall maximum sample size gives cumulative group sizes $n_1 = 102$, $n_2 = 204$, $n_3 = 306$, $n_4 = 678$, and $n_{max} = 1051$. The boundary optimising F^* for these group sizes has an inner wedge at the second interim analysis. It is evident from Figure 5 that this increase in r_S reduces the expected sample size function a little. In the example of Section 2.3, we found some advantage in scheduling fewer analyses for the superiority objective, leaving more to test between non-inferiority and inferiority. Here, we have considered $K_S = 2$ and $r_S = 1.2$, so $n_1 = 153$, $n_2 = 306$, $n_3 = 554$, $n_4 = 803$ and $n_5 = 1051$. The lowest curve in Figure 5 is for the test minimising F^* with these group sizes and we see this design has the smallest expected sample size at all but the highest effect sizes.

Overall, we recognise that Lai et al's proposal gives designs with quite good efficiency, but our wider class allows a design to be tailored to particular objectives and the "inner wedge", not considered by Lai et al, can be instrumental in reducing expected sample size.

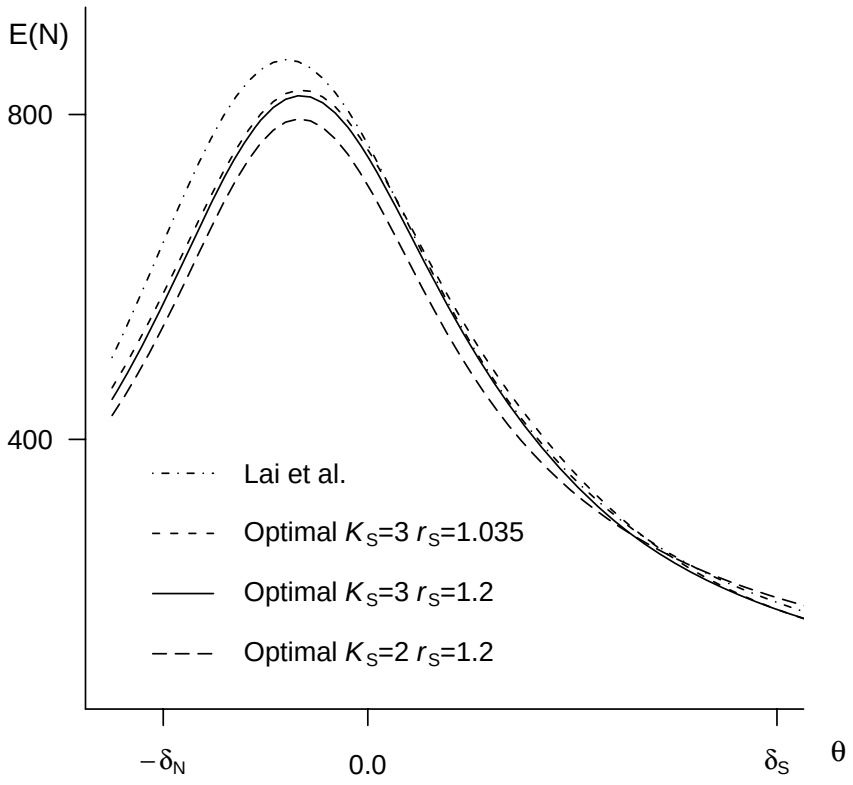

Figure 5: *Expected sample size functions for the Lai et al. design and three optimal* 5*-group designs*

2.5 Error spending designs

To be of real practical value, a group sequential method should be able to deal with variation in group sizes about their planned values. Error spending designs offer this flexibility and, we shall show, can do so with high efficiency in terms of expected sample size. In introducing these designs we broaden consideration to general response distributions, still with the parameter θ representing the treatment effect under investigation. Jennison and Turnbull [13] show that for normal linear models, and asymptotically for general parametric models, the sequence of estimates $\hat{\theta}_1, \ldots, \hat{\theta}_K$ based on accumulating data at K analyses is multivariate normal with

$$\hat{\theta}_k \sim N(\theta, \mathcal{I}_k^{-1}), \quad k = 1, \ldots, K,$$

and

$$\mathrm{Cov}(\hat{\theta}_{k_1}, \hat{\theta}_{k_2}) = \mathrm{Var}(\hat{\theta}_{k_2}) = \mathcal{I}_{k_2}^{-1} \quad \text{for } k_1 < k_2,$$

where $\mathcal{I}_k$ represents the Fisher information for θ at analysis k.

In error spending designs, the cumulative type I and type II error probabilities are specified as functions of the observed information at each analysis; boundaries for standardised statistics $Z_k = \hat{\theta}_k\sqrt{\mathcal{I}_k}$ are found that satisfy these conditions using the distribution theory stated above. This approach relaxes the requirement of achieving pre-planned information levels at each analysis implicit in the tests of Section 2. Lan and DeMets [14] introduced error spending to handle unpredictable group sizes in two-sided tests of a null hypothesis. We now extend this approach to our three-decision problem with its four type I and type II error probabilities.

The information levels required by individual tests of superiority and non-inferiority are

$$\mathcal{I}_{Sf} = \{\Phi^{-1}(1-\alpha_S) + \Phi^{-1}(1-\beta_S)\}^2/\delta_S^2$$

and

$$\mathcal{I}_{Nf} = \{\Phi^{-1}(1-\alpha_N) + \Phi^{-1}(1-\beta_N)\}^2/\delta_N^2.$$

Multiplying these by inflation factors r_S and r_N gives target information levels for an error spending design. Assuming $\mathcal{I}_{Nf} > \mathcal{I}_{Sf}$, the overall maximum information level which an error spending design may require is $\mathcal{I}_{max} = r_N\,\mathcal{I}_{Nf}$. We also specify a target information level by which testing for superiority should terminate, $\mathcal{I}_{max,S} = r_S\,\mathcal{I}_{Sf}$.

Type I and II error probabilities α_S and β_S for the test of superiority are spent according to functions $f_S(\mathcal{I})$ and $g_S(\mathcal{I})$ as $\mathcal{I}$ increases from zero to $\mathcal{I}_{max,S}$. Similarly, spending of error probabilities α_N and β_N for the test of non-inferiority follows functions $f_N(\mathcal{I})$ and $g_N(\mathcal{I})$ as $\mathcal{I}$ increases from zero to $\mathcal{I}_{max}$. At the design stage, we make a working assumption that a specific sequence of information levels will be observed and specify a combination of f_S, g_S, f_N, g_N, $\mathcal{I}_{max,S}$ and $\mathcal{I}_{max}$ for which boundaries converge to spend all four error probabilities exactly by the end of the study. We plan for K interim analyses at information levels

$$\mathcal{I}_k = k\,\mathcal{I}_{max,S}/K_S \quad \text{for } k = 1, \ldots, K_S \tag{7}$$

and

$$\mathcal{I}_k = \mathcal{I}_{max,S} + (k-K_S)(\mathcal{I}_{max} - \mathcal{I}_{max,S})/(K-K_S) \quad \text{for } k = K_S+1, \ldots, K. \tag{8}$$

In practice, the test will adapt to observed information levels, maintaining type I error probabilities precisely but with small perturbations to the type II error rates.

Our choice of error spending functions is motivated by the cumulative error rates seen in optimal designs. Since these designs do not allow very early decisions of non-inferiority, we also delay spending α_N and β_S until information reaches a minimum threshold $\gamma\,\mathcal{I}_{max,S}$, where $0 \leq \gamma \leq 1$. This is a sensible feature since, with only a small amount of data, one cannot be confident that θ is both above $-\delta_N$ and below δ_S. We propose a family of designs with spending

functions indexed by the parameter $\rho > 0$, similar in form to those for the two-decision problem used by Jennison and Turnbull [15]. The four error spending functions are:

$$f_N(\mathcal{I}) = \begin{cases} 0 & \text{if } \mathcal{I} < \gamma\,\mathcal{I}_{max,S} \\ \alpha_N\,(\mathcal{I}/\mathcal{I}_{max})^\rho & \text{if } \gamma\,\mathcal{I}_{max,S} \le \mathcal{I} < \mathcal{I}_{max} \\ \alpha_N & \text{if } \mathcal{I} \ge \mathcal{I}_{max} \end{cases}$$

$$f_S(\mathcal{I}) = \begin{cases} \alpha_S\,(\mathcal{I}/\mathcal{I}_{max,S})^\rho & \text{if } \mathcal{I} < \mathcal{I}_{max,S} \\ \alpha_S & \text{if } \mathcal{I} \ge \mathcal{I}_{max,S} \end{cases}$$

$$g_N(\mathcal{I}) = \begin{cases} \beta_N\,(\mathcal{I}/\mathcal{I}_{max})^\rho & \text{if } \mathcal{I} < \mathcal{I}_{max} \\ \beta_N & \text{if } \mathcal{I} \ge \mathcal{I}_{max} \end{cases}$$

$$g_S(\mathcal{I}) = \begin{cases} 0 & \text{if } \mathcal{I} < \gamma\,\mathcal{I}_{max,S} \\ \beta_S\,(\mathcal{I}/\mathcal{I}_{max,S})^\rho & \text{if } \gamma\,\mathcal{I}_{max,S} \le \mathcal{I} < \mathcal{I}_{max,S} \\ \beta_S & \text{if } \mathcal{I} \ge \mathcal{I}_{max,S}. \end{cases}$$

where $\gamma > 0$. Figure 6 shows these functions for the case $\rho = 1$ and $\gamma = 0.5$. When $\mathcal{I}_{Sf} < \mathcal{I}_{Nf}$, Brannath et al. [16] comment on the desirability of spending the type I error probability α_S for the superiority objective more rapidly than that for the test of non-inferiority, α_N. This feature is built into our definitions of spending functions but there would be no difficulty in taking such considerations further and varying the values of ρ in the four spending functions.

Application of this error spending design with an observed sequence of information levels, $\mathcal{I}_1, \mathcal{I}_2, \ldots$, follows the general framework described by Jennison and Turnbull [9, Chapter 7] for other types of error spending test. At the first few analyses with $\mathcal{I}_k < \gamma\,\mathcal{I}_{max,S}$ only the outer boundary values d_k and a_k are required. These are calculated to satisfy

$$P_{\theta=0}(\text{Declare "Superiority" by analysis } k) = f_S(\mathcal{I}_k) \tag{9}$$

and

$$P_{\theta=0}(\text{Declare "Inferiority" by analysis } k) = g_N(\mathcal{I}_k). \tag{10}$$

We do allow stopping to declare superiority when $f_N(\mathcal{I}_k) = 0$, even though this represents a type I error for the test of non-inferiority under $\theta = -\delta_N$. Similarly, we permit a decision of inferiority when $g_S(\mathcal{I}_k) = 0$, even though this is a type II error for the test of superiority under $\theta = \delta_S$. The probabilities of these outcomes are computed so they can be accounted for at later analyses when $f_N(\mathcal{I}_k)$ and $g_S(\mathcal{I}_k)$ become positive.

For $\gamma\,\mathcal{I}_{max,S} \le \mathcal{I}_k \le \mathcal{I}_{max,S}$, we compute d_k and a_k to satisfy (9) and (10) and perform a two-dimensional search to find values b_k and c_k satisfying

$$P_{\theta=-\delta_N}(\text{Declare "Non-inferiority" or "Superiority" by analysis } k) = f_N(\mathcal{I}_k) \tag{11}$$

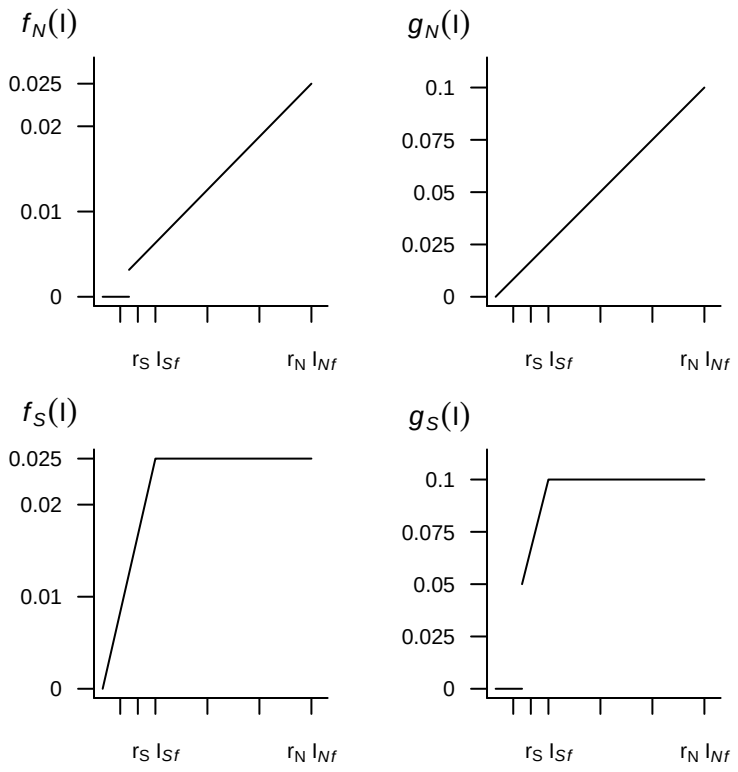

Figure 6: *Error spending functions with* $\rho = 1$ *and* $\gamma = 0.5$

and

$$P_{\theta=\delta_S}(\text{Declare "Inferiority" or "Non-inferiority" by analysis } k) = g_S(\mathcal{I}_k). \tag{12}$$

Further details of how b_k and c_k can be found are given in Appendix III.

At the first analysis $\tilde{K}_S$ where $\mathcal{I}_{\tilde{K}_S} \geq \mathcal{I}_{max,S}$ we calculate $d_{\tilde{K}_S}$ so that

$$P_{\theta=0}(\text{Declare "Superiority" by analysis } \tilde{K}_S) = \alpha_S$$

and set $c_{\tilde{K}_S} = d_{\tilde{K}_S}$. We find $a_{\tilde{K}_S}$ and $b_{\tilde{K}_S}$ satisfying (10) and (11) with $k = \tilde{K}_S$ and $\mathcal{I}_k = \mathcal{I}_{\tilde{K}_S}$. At subsequent analyses with $\mathcal{I}_k < \mathcal{I}_{max}$ we set $c_k = d_k = \infty$ and calculate a_k and b_k to satisfy (10) and (11). Finally, at the first analysis $\tilde{K}$ with $\mathcal{I}_{\tilde{K}} \geq \mathcal{I}_{max}$ we find $b_{\tilde{K}}$ satisfying

$$P_{\theta=-\delta_N}(\text{Declare "Non-inferiority" or "Superiority" by analysis } \tilde{K}) = \alpha_N$$

and set $a_{\tilde{K}} = b_{\tilde{K}}$.

By construction, this error spending procedure attains the type I error probabilities α_N and α_S exactly. The type II error probabilities may differ slightly from β_N and β_S but they will be close to these targets if the observed information levels are similar to the sequence defined by (7) and (8) which was assumed for planning purposes.

Suppose α_S, α_N, β_N and β_S are specified, so $\mathcal{I}_{Sf}$ and $\mathcal{I}_{Nf}$ are fixed multiples of δ_S^{-2} and δ_N^{-2}, respectively. For given δ_N/δ_S, ρ, γ, K_S and K, and assuming analyses are scheduled according to (7) and (8), a two-dimensional search can be conducted to find the inflation factors r_S and r_N which give power $1-\beta_N$ at $\theta = 0$ and $1-\beta_S$ at $\theta = \delta_S$. Within the ρ family, increasing ρ reduces the rate at which error is spent, leading to smaller inflation factors. Thus, for given α_S, α_N, β_N, β_S, γ, K_S, K and $\delta_N/\delta_S < 1$, say, there is a one-to-one correspondence between ρ and r_N. While inflation factors do increase gradually with K, broadly speaking, setting $\rho = 3$ gives an inflation factor around $r_N = 1.05$ and wide outer boundaries similar to an O'Brien and Fleming [17] test, whereas $\rho = 1$ yields an inflation factor around 1.2 or 1.25 and narrower boundaries, as in a Pocock [18] test

Comparing the ρ family error spending tests with designs optimised for F^*, we have found the ρ family tests to be highly efficient and achieve values of F^* within a few percent of the minimum possible for a given inflation factor r_N. We conclude that error spending designs in the ρ family are both efficient and sufficiently flexible to handle unpredictable group sizes or information levels. These findings are in keeping with those of Barber and Jennison [12] for one-sided error spending tests with similar spending functions.

As an illustration of the preceding remarks, Figure 7 shows the expected sample size function for the design with $\alpha_N = \alpha_S = 0.025$, $\beta_N = \beta_S = 0.1$, $\delta_N = 0.1$, $\delta_S = 0.2$, $\rho = 1$, $\gamma = 0.5$, $K_S = 3$ and $K = 6$, as well as that for the optimal design minimising F^* for the same problem and group sizes. It is evident that the error spending design is highly efficient across the range of θ values and, overall, it achieves a value of F^* within 2% of the optimum.

If we consider the same example but vary ρ from 0.5 to 3, we obtain designs with inflation factors r_N ranging from around 1.5 to 1.05. Figure 8 shows values of F^* for these error spending designs plotted against the inflation factor r_N for each design. The slightly lower curve gives the value of F^* achieved by optimal designs for this criterion with the same group sizes, which is around 2 to 4 per cent smaller than that of the error spending design. The levelling off of F^* as ρ decreases and r_N increases indicates there is no advantage in taking r_N greater than around 1.2, which is attained by ρ of about 0.8.

We recommend 0.5 as a simple default value for γ. In a detailed assessment of a particular case one can go further and compare values of γ with respect to expected sample size functions under different sequences of information levels, paying particular attention to the effect of $\mathcal{I}_1$.

Another advantage of error spending tests is that they support use of the method of information monitoring, as proposed by Mehta and Tsiatis [19]. This approach can be used to manage a trial when the sample size needed for specific power depends on nuisance parameters which are only estimated once the trial

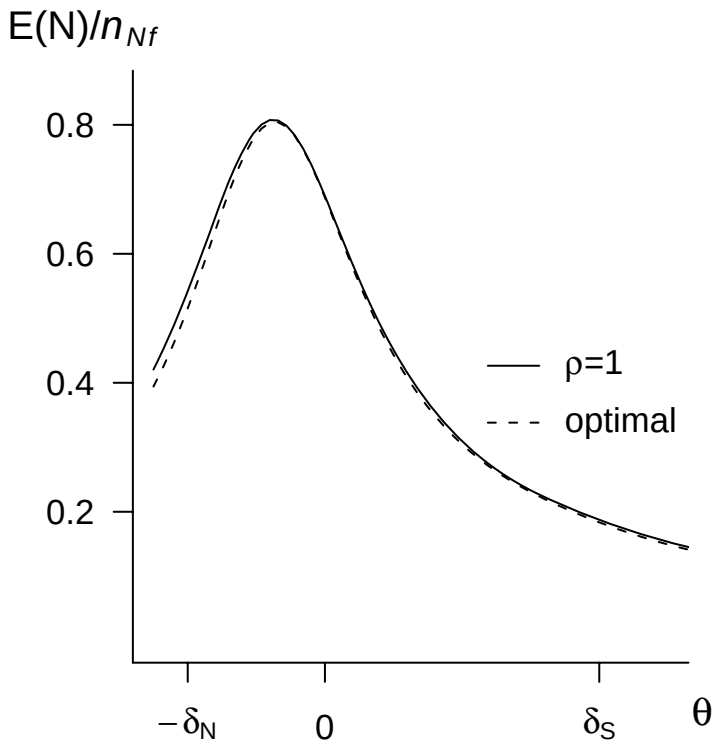

Figure 7: *$E(N)/n_{Nf}$ for a 6-group error spending design with ρ=1 and γ = 0.5 and for the optimal 6-group design with analyses performed at the same information levels*

is under way. One example of such a parameter is the variance of a normal response: thus, error spending and information monitoring provide a way to deal with unknown variance in the normal response problem introduced in Section 2.1.

3 Optimal adaptive designs

3.1 Framework

The need for different sample sizes to test superiority and non-inferiority has prompted proposals for designs in which future group sizes are based on previously observed data. Such procedures are examples of adaptive group sequential designs, as proposed for one-sided tests by Schmitz [20]. These methods extend those of Section 2 by allowing each new group size to depend

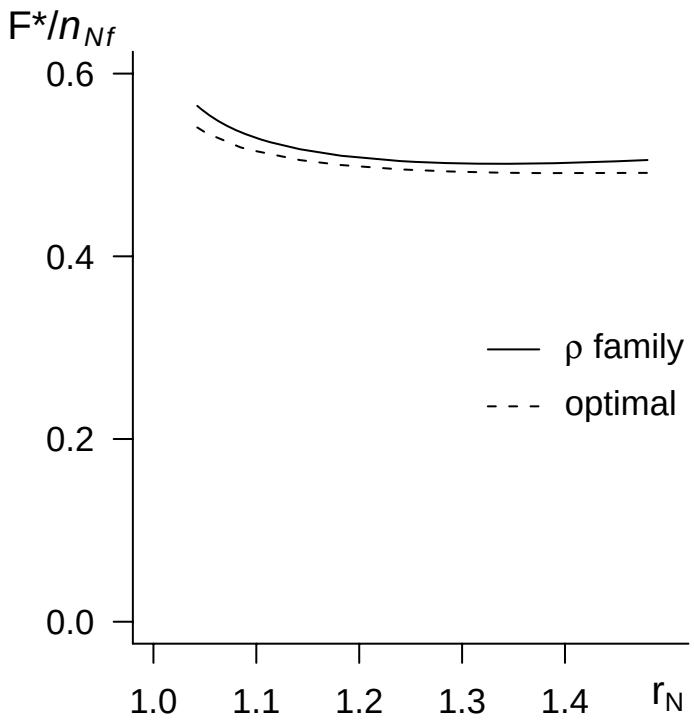

Figure 8: *F^*/n_{Nf} for ρ family error spending designs with ρ in the range 0.5 to 3 and for optimal designs minimising F^* with the same sequences of information levels*

on previous data. Since this wider class includes non-adaptive group sequential designs as special cases, optimising over it yields a lower value of an objective function, such as F^*, than that of the best non-adaptive design. We shall now explore the benefits of adaptation in reducing expected sample size in the three-decision problem and consider whether they justify the extra complexity of this approach. The definition of adaptive group sequential designs applies for general K but we shall focus on $K = 2$ for computational simplicity.

In the three-decision problem, adaptive designs may terminate at each analysis with a decision of inferiority, non-inferiority or superiority. However, if sampling continues at analysis k, the next group size is allowed to depend on Z_k. We seek the sequential decision rule that minimises F^* subject to the error constraints (1) to (4). The optimal K-group adaptive design can be derived by the Lagrangian approach used in Section 2 for the non-adaptive case. The unconstrained problem is a Bayes decision problem, solvable by dynamic programming.

In solving the unconstrained problem for the case $K = 2$, suppose the first analyses takes place after n_1 observations. We approximate the continuous range of values for n_2 by giving M possible cumulative group sizes, $n_{2,1}, \ldots, n_{2,M}$, for the final analysis. We find the optimal critical values $b_{2,1}, \ldots, b_{2,M}$ and $d_{2,1}, \ldots, d_{2,M}$ to decide between inferiority, non-inferiority and superiority in each case. The task at the first analysis is to determine which of the following actions is optimal: stop for inferiority, stop for non-inferiority, stop for superiority, or continue to the final analysis with cumulative group size $n_{2,m}$ for a value of $m \in \{1, \ldots, M\}$. This gives the optimal design for a given value of n_1 and a further search over n_1 gives the two-stage adaptive design with the overall minimum of F^*.

In numerical calculations we have used $M = 100$ and $n_{2,1}, \ldots, n_{2,M}$ equally spaced between n_1 and an upper limit $R\, n_{Nf}$ where $R > 1$. We performed sensitivity analyses to check there is no significant change to the design if M or R is increased. For most examples we have found that when using $R = 1.35$, the optimal choice of n_2 is lower than $R\, n_{Nf}$ for all values of Z_1. We give further details of implementing the dynamic programming algorithm in Appendix II.

3.2 Efficiency gains through adaptation

We illustrate with an example the possible efficiency gains from adaptation in two-stage designs. The results in Table 1 are for non-adaptive and adaptive two-stage designs which minimise F^* subject to error probabilities $\alpha_N = \alpha_S = 0.025$ and $\beta_N = \beta_S = 0.1$ for values of $\delta_{Sf}/\delta_{Nf} = (n_{Nf}/n_{Sf})^{1/2}$ ranging from 1 to $\sqrt{3}$. In the adaptive design the initial group size, n_1, is chosen optimally and the second group size, $n_2 - n_1$, is selected to be optimal for the observed Z_1. For the non-adaptive designs, n_1 and n_2 are fixed at optimal values for the objective function F^*. The maximum value of n_2 in the adaptive designs ranges from $1.20\, n_{Nf}$ to $1.26\, n_{Nf}$ in the six cases of Table 1 whereas, in the non-adaptive designs, n_2 takes lower values between $1.12\, n_{Nf}$ and $1.17\, n_{Nf}$. In fact, for the case $n_{Nf}/n_{Sf} = 3$, values of n_1 greater than n_{Sf} help to minimise F^* in both the non-adaptive and adaptive designs, but lead to power for the test of superiority greater than the stipulated $1 - \beta_S = 0.9$. Reformulating this requirement as an inequality that power should be at least 0.9, leads to the designs reported here which have both higher power for the test of superiority (around 0.93) and lower F^*.

The results in Table 1 show only minor benefits from adaptation. These benefits are greatest for intermediate values of the ratio n_{Nf}/n_{Sf} and in these cases there are areas of the adaptive design's continuation region at the first analysis where each of the three final decisions is plausible. This leads to substantial variation of the optimal values for n_2 with Z_1, as displayed in Figure 9 for the case $n_{Nf}/n_{Sf} = 1.5$. In view of the variation in the optimal n_2, it is not surprising that the best non-adaptive design, with only a single value of n_2, is less efficient. On the other hand, Figure 9 suggests it might be sufficient to choose between just two sample sizes, $n_{N,2}$ and $n_{S,2}$ say, in the lower and upper continuation regions, either side of the "inner wedge". We refer to such

$n_{Nf}/n_{Sf} =$	1.0	1.25	1.5	1.75	2.0	3.0
Optimal non-adaptive designs	81.8	75.3	71.7	69.1	67.0	64.6
Optimal adaptive designs	81.7	74.3	70.1	67.7	66.1	64.2

Table 1: Values of $100\,F^*/n_{Nf}$ for optimal two-stage designs with error probabilities at most $\alpha_N = \alpha_S = 0.025$ and $\beta_N = \beta_S = 0.1$ for selected values of $n_{Nf}/n_{Sf} = (\delta_S/\delta_N)^2$.

a procedure as a restricted adaptive design.

We computed a two-stage restricted design minimising F^* for the case $n_{Nf}/n_{Sf} = 1.5$, with n_1 set at the value chosen for the unrestricted adaptive design. The dashed lines in Figure 9 show the continuation intervals, which differ slightly from the unrestricted design, and values of $n_{N,2}$ and $n_{S,2}$. In Figure 10, the expected sample size function for the restricted adaptive design lies very close to that of the optimal unrestricted adaptive design, demonstrating that the key improvement in the adaptive design comes from choosing a sample size appropriate to the most relevant decision choice, superiority vs non-inferiority or non-inferiority vs inferiority, and not from any further fine-tuning. Values of $100F^*/n_{Nf}$ are 71.7 for the two-group non-adaptive test, 70.4 for the restricted adaptive test, and 70.1 for the two-group adaptive test.

Figure 10 also shows the expected sample size function for the optimal non-adaptive three-group design with $n_{Nf}/n_{Sf} = 1.5$ and cumulative sample sizes equal to the values of n_1, $n_{S,2}$ and $n_{N,2}$ in the restricted adaptive design. Since the third analysis is only used to distinguish between inferiority and non-inferiority, this is an example from our class of non-adaptive designs with $K = 3$ and $K_S = 2$ (and $c_3 = d_3 = \infty$). This three-group non-adaptive design has lower expected sample size than the optimal adaptive two-stage design across the range of θ values and it is significantly more efficient at low values of θ; its value of $100F^*/n_{Nf}$ is 66.8, compared to 70.1 for the optimal adaptive two-stage design. Our conclusions here concur with those of Jennison and Turnbull [7] about the two-decision problem: while adaptivity can lead to a small increase in efficiency, similar or larger improvements can be achieved with one additional interim analyses in a non-adaptive group sequential design. In view of the minor benefits accruing from adaptive choice of group size in the case $K = 2$, we have not carried out computation of optimal adaptive designs for higher values of K.

3.3 Competing adaptive methods

The sample size function for the optimal adaptive design in Figure 9 is qualitatively different from that arising from a conditional power rule, where sample size rises as Z_1 decreases, at least within each continuation interval. While optimal adaptive designs offer modest gains over their non-adaptive

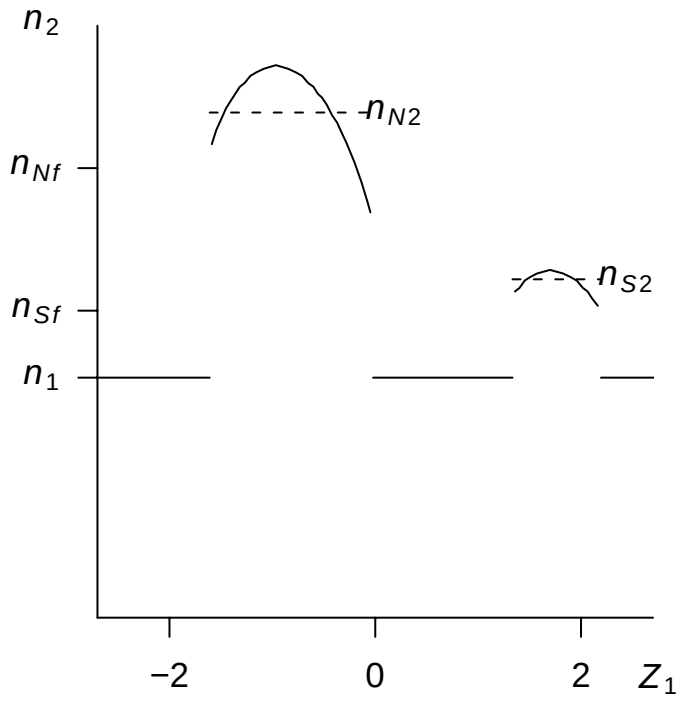

Figure 9: *Final sample size, n_2, as a function of Z_1 for the optimal adaptive design for $n_{Nf}/n_{Sf} = 1.5$ (solid lines) and sample sizes $n_{N,2}$ and $n_{S,2}$ for the optimal restricted adaptive design (dashed lines)*

counterparts, the following comparisons show published adaptive methods with sub-optimal sampling rules can be less efficient than simpler non-adaptive designs.

Koyama et al. [5] propose an adaptive two-stage procedure for simultaneous testing of superiority and non-inferiority. After the first stage, stopping is possible for inferiority, non-inferiority or superiority. If the trial continues, the second stage sample size is set as a function of the first stage test statistic, Z_1. The sample size function and terminal decision rules are chosen to achieve specified overall error probabilities α_N, α_S, β_N, and β_S. Koyama et al. [5] provide an example with $\delta_N = 1.0$, $\delta_S = 0.5$, $\sigma = 4$, α_N=$\alpha_S = 0.025$, $\beta_N = 0.1$ and $\beta_S = 0.2$. While we have focused on designs with $\delta_S > \delta_N$, our framework also applies to the case $\delta_S < \delta_N$ studied by [5], [6] and [16]. We have compared Koyama et al's adaptive procedure with a two-stage non-adaptive design with the same error probabilities optimised for F^*. In this design, the first analysis

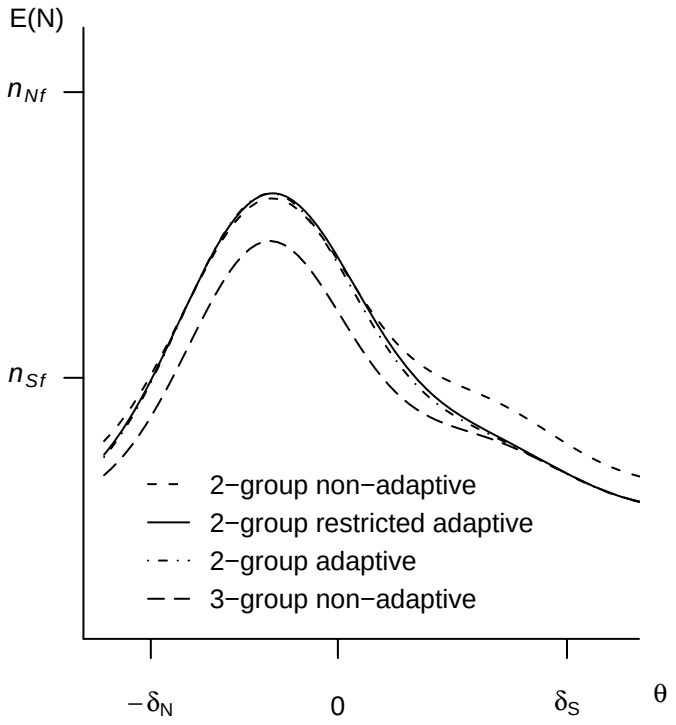

Figure 10: *Expected sample size functions for optimal non-adaptive, restricted adaptive and adaptive 2-group designs and the optimal non-adaptive 3-group design*

is scheduled after 337 observations per treatment and the final analysis after 1200 observations. Expected sample sizes per treatment are shown in Table 2. Not only is the non-adaptive procedure more efficient, but its maximal sample size per treatment is only 1200, compared to more than 1500 for the adaptive design.

The method proposed by Shih et al. [4] falls within the framework for 2-stage adaptive procedures defined in Section 3.1. Early stopping for futility, non-inferiority or superiority is possible at the first analysis and critical values at both analyses are chosen to control the overall type I error probabilities at specified values α_N and α_S. The procedure does not aim to achieve a particular overall power, rather the second stage sample size is chosen with reference to conditional power given the first stage data. We have simulated the design presented for the survival data example in Section 3 of Shih et al. [4] and found the overall power curves and expected sample size function of this design: with a

θ	$E_\theta(N)$ for Koyama et al's design	$E_\theta(N)$ for a 2-group non-adaptive design
$-\delta_N$	337	337
0	643	625
δ_S	996	937

Table 2: Comparison of Koyama et al's [5] adaptive design and an optimal non-adaptive design

survival endpoint, "sample size" should be interpreted as the number of events observed at termination. We constructed a non-adaptive 2-group sequential design with the same type I error probability and overall power curves at least as high over the range of effect sizes. This non-adaptive group sequential design had lower expected sample size by between 3% and 11% at values of θ in the range $-\delta_N$ to $2\,\delta_N$. We attribute the lower efficiency of the adaptive procedure to the choice of sample size function: for the optimal adaptive rule, values of n_2 are highest in the middle of each arm of the continuation region and lower nearer the boundary points, whereas the conditional power construction implies n_2 increases monotonically as $\hat{\theta}$ decreases.

Wang et al. [2] propose an adaptive group sequential closed (AGSC) procedure which starts out as a group sequential design but can shift adaptively between superiority and non-inferiority objectives. When $\delta_N < \delta_S$ and $n_{Nf} > n_{Sf}$, the initial design has n_{Sf} observations and K analyses. At each interim analysis, conditional power calculations determine whether to shift to the non-inferiority objective. If so, group sizes are increased to lead to a final sample size of n_{Nf} at analysis K with down-weighting as in the method of Cui et al. [3] to maintain the type I error rate. Type II error rates are not controlled directly but are governed by n_{Nf}, n_{Sf}, the group sequential stopping boundary and the adaptive decision rule.

We evaluated the AGSC method by simulation with one million replicates. We assumed normal responses with $\sigma^2 = 9$, $\alpha_N = \alpha_S = 0.025$, $\delta_N = 0.4$ and $\delta_S = 0.8$. The initial design had five equally spaced analyses and a total of $n_{Sf} = 221$ observations per treatment arm, increasing to $n_{Nf} = 883$ under adaptation. Figure 11 compares the AGSC method and a non-adaptive 5-group sequential design with $K_S = 2$, $n_{K_S} = 221$, $K = 5$, and $n_5 = 883$. The non-adaptive design has higher power and a substantially lower expected sample size function. Since the AGSC method has no lower boundary to allow stopping for inferiority, its high expected sample size under low values of θ is to be expected. At higher effect sizes, using non-sufficient statistics as a result of down-weighting later observations is a source of inefficiency. More important, we believe, is the reliance on uncertain estimates of $\hat{\theta}$ at the interim analyses in making the decision to increase sample size four-fold. While we have found

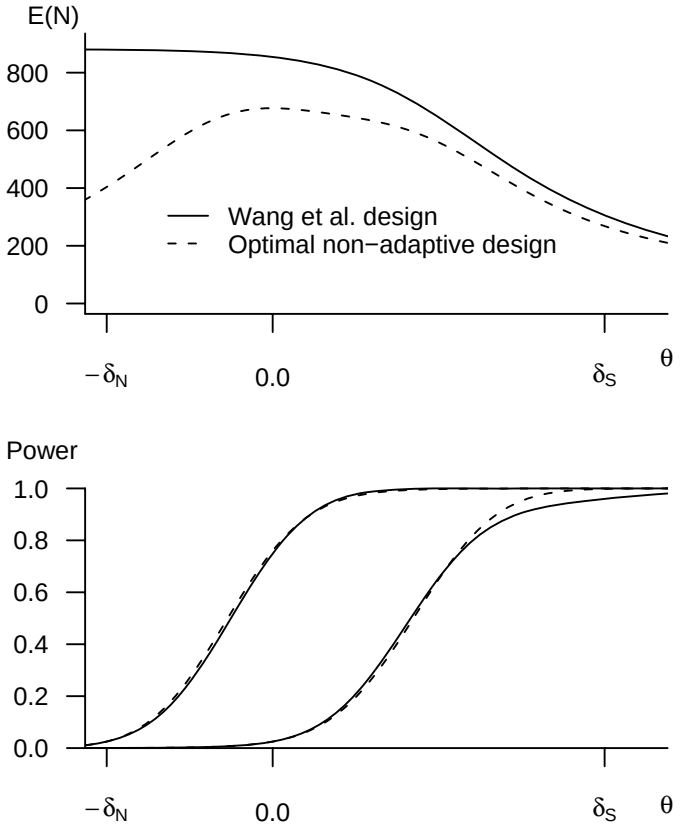

Figure 11: *Expected sample size functions and power curves for Wang et al's AGSC design and an optimal non-adaptive 5-group sequential design*

the addition of a lower futility boundary improves performance for low values of θ, the method still fails to match the performance of the non-adaptive group sequential test at higher effect sizes.

4 An example in type 2 diabetes

The EMEA guidance [21] recommends decrease from baseline HbA1c, a measure of blood glucose control, as the primary endpoint for studies of type 2 diabetes. In the trial reported by Home et al. [22], response was the percentage decrease in HbA1c, the non-inferiority margin was $\delta_N = 0.4$ and a standard deviation of 1.4 was used in the sample size calculation. Göke et al. [23] report a clinical

trial with power to detect an improvement of $\delta_S = 0.5$ under the new treatment and, again, a standard deviation of 1.4 was used to determine sample size.

Consider designing a trial to compare a new treatment for type 2 diabetes against a standard, testing for both superiority and non-inferiority. Suppose responses are normally distributed with $X_{Ai} \sim N(\mu_A, \sigma^2)$ on the new treatment and $X_{Bi} \sim N(\mu_B, \sigma^2)$ on the standard. Denoting the treatment effect by θ=$\mu_A - \mu_B$, we wish to test simultaneously the null hypothesis $H_{N,0}$: $\theta \leq -0.4$ against $\theta > -0.4$ with power specified at $\theta = 0$ and the null hypothesis $H_{S,0} : \theta \leq 0$ against $\theta > 0$ with power at $\theta = 0.5$. Thus, we set $\delta_N = 0.4$, $\delta_S = 0.5$, $\alpha_N = 0.025$, $\alpha_S = 0.025$, $\beta_N = 0.1$, and $\beta_S = 0.1$ in our general framework. With $\sigma = 1.4$, fixed sample sizes per treatment are $n_{Sf} = 165$ and $n_{Nf} = 258$ for the two individual hypothesis tests.

For n_{Nf}/n_{Sf} around 1.5, the results of Section 3.2 indicate adaptation may be helpful if only two analyses are possible. We computed an adaptive two-group design optimised for F^* with $n_1 = 99$ and no upper limit for the second group size. We also derived a "restricted adaptive" design, as introduced in Section 3.2, where $n_1 = 99$ and, if sampling continues, the choice of the final sample size is either 198 or 309. The two upper curves in Figure 12 are the expected sample size functions for these restricted adaptive and adaptive designs. We see that restricting the second group size to just two values has a negligible effect on efficiency.

Comparison of expected sample size functions allows an informed choice of a suitable design. In making this choice, investigators may also consider the logistical challenges of setting up a trial with data-driven choice of the second group size. Information leakage should be considered since, in both the adaptive and restricted adaptive designs, knowledge of the second stage sample size provides an indication of the first stage results. In this joint testing problem, leakage can also be an issue for a non-adaptive group sequential design: in a 3-group procedure with $K_S = 2$, continuation past the second analysis implies the new treatment has not been found to be superior and the decision will be either "non-inferior" or "inferior".

The error spending method of Section 2.5 can be used to give a design with close to optimal efficiency as well as the flexibility to deal with unpredictable group sizes. If we use ρ family error spending functions with $\rho = 1$ and design for $K = 3$ analyses with $K_S = 2$ and $\gamma = 0.4$, the inflation factors are $r_S = 1.167$ and $r_N = 1.195$, so $n_{max,S} = 1.167 n_{Sf} = 193$ and the maximum sample size is $n_{max} = 1.195\, n_{Nf} = 308$. If observed sample sizes follow the design pattern of $n_1 = 97$, $n_2 = 193$ and $n_3 = 308$, power of 0.9 is attained exactly in both hypothesis tests. The expected sample size function for this design shown in Figure 12 is almost identical to that obtained by a 3-group sequential design with $r_S = r_N = 1.2$ optimised for F^*.

Suppose patient accrual is lower than expected and only $\tilde{n}_1 = 71$ responses are observed at the first analysis. Since $\tilde{n}_1 < \gamma n_{max,S}$, there is no inner wedge at the first analysis. If accrual remains slow throughout the trial, a fourth analysis will be needed to reach n_{max} but the error spending design adjusts easily to the new sequence of sample sizes. Suppose we observe

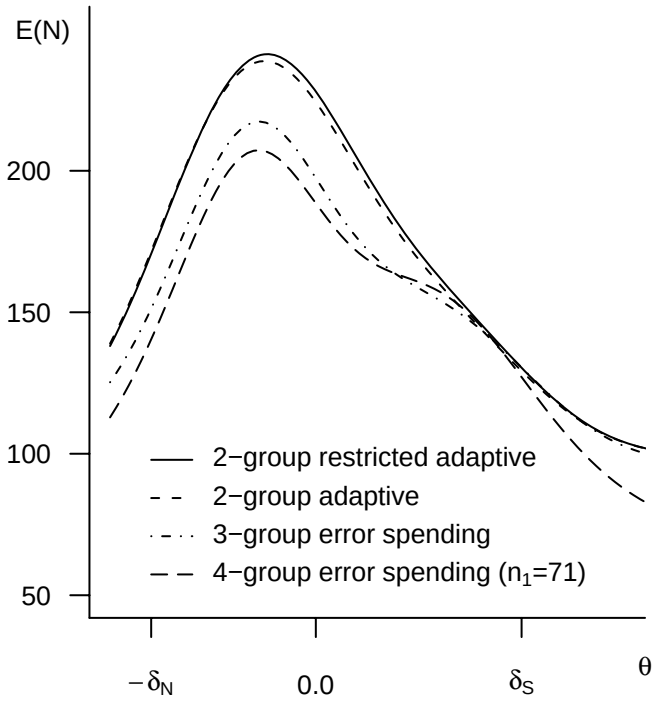

Figure 12: *Expected sample size functions for designs in the type 2 diabetes example*

$\tilde{n}_2 = 144$, $\tilde{n}_3 = 220$ and $\tilde{n}_4 = 308$. The critical values for what is now a 4-group design are computed following the prescription in Section 2.5: the resulting boundaries are shown in Figure 13. The type I error probabilities are automatically controlled at $\alpha_N = 0.025$ and $\alpha_S = 0.025$ and the attained type II error probabilities are $P_{\theta=0}$(Conclude "Inferiority") $= 0.102$ and $P_{\theta=\delta_S}$(Conclude "Inferiority or Non-inferiority") $= 0.088$, both close to their intended values of $\beta_N = \beta_S = 0.1$. The inner wedge plays an important role, allowing stopping for any of the three possible outcomes, superiority, non-inferiority and inferiority, at analyses two and three. The expected sample size function in this case is the lowest curve in Figure 12. So, not only does the error spending design deal well with the observed pattern of group sizes, but results for this four group design show it gains efficiency by adapting to a higher number of smaller group sizes when these arise in practice.

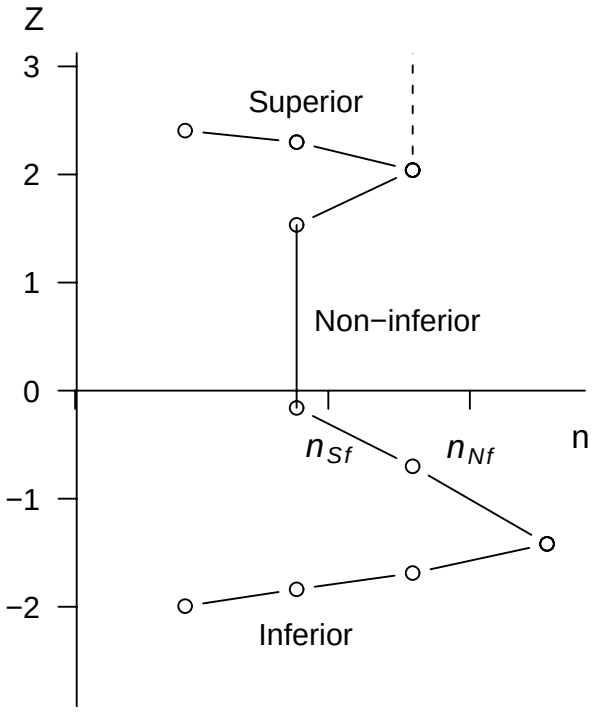

Figure 13: *Critical values for 4-group error spending design*

5 Discussion

We have introduced a framework to define group sequential designs which test simultaneously for superiority and non-inferiority and allow early stopping for any one of the three conclusions of superiority, non-inferiority and inferiority. We can compute the design of this type which minimises a weighted combination of expected sample sizes at several effect sizes. We have also defined error spending versions of these designs which can handle unpredictable group sizes while retaining almost optimal efficiency. Expressing these error spending designs in terms of information for the effect size parameter shows they are applicable to a wide variety of response types and can deal with nuisance parameters governing the sample size needed for a specific power through the information monitoring approach of Mehta and Tsiatis [19].

We have followed other authors in addressing the situation where the non-inferiority margin δ_N is smaller than the effect size δ_S at which power is set in the test for superiority. Here, much larger sample sizes are required when the study focuses on distinguishing between non-inferiority and inferiority. If such large sample sizes are known to be available if needed, one might expect investigators

to consider increasing the power of the test for superiority to detect smaller effect sizes than δ_S. If this occurs, n_{Sf} will increase and the ratio n_{Nf}/n_{Sf} will be closer to one. Our framework will still be appropriate and the inner wedge, which allows early stopping to declare non-inferiority, will play an important role in such cases.

In the two-decision problem, Barber and Jennison [12] found the greatest benefit of an additional interim analyses to arise when moving from a fixed sample test to a two-group test. For the three-decision problem, two analyses are required simply to meet the different error constraints for the pair of hypothesis tests. When sample sizes for the two testing objectives are very different, a total of four analyses is needed to allow two suitably placed analyses for each hypothesis test. Thus, it is a feature of the group sequential designs for the three decision problem that a larger number of interim analyses is likely to be worthwhile than for group sequential designs for the two-decision problem.

In the "adaptive" designs considered in Section 3, future group sizes are based on current data, in particular the observed effect $\hat{\theta}$. Remember, though, that our "non-adaptive" group sequential designs also respond to the observed data: the stopping rule provides a very definite response and, when $n_{Nf} > n_{Sf}$, the absence of an upper continuation region at the last few analyses shows a shift of focus to the test between non-inferiority and inferiority.

In exploring the adaptive choice of group sizes, we have found only minor benefits of adaptation in two-stage designs, the case most often considered in the literature. In fact, we saw in Section 3.3 that non-adaptive group sequential designs can out-perform some proposed adaptive methods with the same number of analyses. The greatest advantage we have found of an adaptive over an optimal non-adaptive 2-group design is around 3% of the fixed sample size. This may be a significant benefit in a clinical trial with thousands of patients — but then there is reason to pursue the even greater benefits of a non-adaptive 3-group sequential design. We have not invested effort in deriving optimal adaptive designs with three or more analyses as we do not anticipate substantively different results from the two-group case.

APPENDIX

I Monotonicity of type I and type II error probabilities

It seems intuitive that the probability of rejecting a null hypothesis such as $H_{S,0}$: $\theta \leq 0$ should increase with θ in any sensible experimental design. A coupling argument can provide a proof for some group sequential designs (see, for example, Jennison and Turnbull [9, Page 183]), but this approach does not extend to group sequential designs with an inner wedge. Adaptive designs pose further problems, indeed Jennison and Turnbull [24, Section 4.2]) present an adaptive design with a non-monotone power function. However, Shih et al. [4]

are able to prove monotonicity of the type I error probability within the null hypothesis for two-stage adaptive designs. We now generalise their result to K-group designs.

Consider first the non-adaptive case and a K-group design, as defined in Section 2.1. Let $f_k(z_k;\theta)$ denote the density of Z_k at analysis k under treatment effect θ in the absence of any prior early stopping. Define $p_k(z_k)$ to be the conditional probability that $Z_1,\ldots,Z_{k-1}$ lie in the continuation regions $(a_1,b_1)\cup(c_1,d_1),\ldots,(a_{k-1},b_{k-1})\cup(c_{k-1},d_{k-1})$ given that $Z_k=z_k$. Since Z_k is sufficient for θ, this probability does not depend on θ. We can write

$$P_\theta(\text{Declare "Superiority"}) = \sum_{k=1}^{K}\int_{d_k}^{\infty} f_k(z_k;\theta)\,p_k(z_k)\,dz_k. \tag{13}$$

Now, marginally, $Z_k \sim N(\theta\sqrt{\mathcal{I}_k},\,1)$. Thus, if $d_k \geq 0$, $f_k(z_k;\theta)$ is an increasing function of θ for all $\theta \leq 0$ and $z_k > d_k$. It follows that all the integrands in the right hand side of (13) are increasing in θ for $\theta \leq 0$. Hence, as long as $d_k \geq 0$ for each $k = 1,\ldots,K$, P_θ(Declare "Superiority") increases monotonely with θ for $\theta \leq 0$ and the maximum type I error probability over $\theta \leq 0$ occurs at $\theta = 0$. The condition $d_k \geq 0$ implies that stopping to reject $H_{S,0}$: $\theta \leq 0$ is only possible when $\hat{\theta}_k \geq 0$, which is to be expected in any sensible design.

A similar argument shows P_θ(Declare "Non-inferiority or Superiority") is monotone over $\theta \leq -\delta_N$. In this case, the integrals in (13) have range $(b_k,c_k)\cup(d_k,\infty)$ and the condition for integrands to be increasing for $\theta \leq -\delta_N$ becomes $b_k \geq -\delta_N\sqrt{\mathcal{I}_k}$, so $H_{N,0}$: $\theta \leq -\delta_N$ is only rejected when $\hat{\theta}_k \geq -\delta_N$. The same approach can be used to establish monotonicity results for type II error probabilities: P_θ(Conclude "Inferiority") decreases with θ for $\theta \geq 0$ as long as this decision only occurs when $\hat{\theta}_k \leq 0$, and P_θ(Conclude "Inferiority" or "Non-inferiority") decreases with θ for $\theta \geq \delta_S$ as long as this decision requires $\hat{\theta}_k \leq \delta_S$.

We can obtain results for adaptive group sequential designs by essentially the same argument. Since the sample size at each analysis now depends on previous responses, the sum over k in (13) becomes a double sum over k and the set of possible sequences $\{\mathcal{I}_1,\ldots,\mathcal{I}_k\}$. In some designs the critical value d_k may depend on the whole sequence $\{\mathcal{I}_1,\ldots,\mathcal{I}_k\}$. It is useful, conceptually, to define the sequence of Z-statistics at all information levels that might arise, noting the joint distribution theory stated at the start of Section 2.5 applies to this whole sequence. We let $f_k(z_k,\mathcal{I}_k;\theta)$ denote the $N(\theta\sqrt{\mathcal{I}_k},\,1)$ density of Z_k for treatment effect θ and information level $\mathcal{I}_k$ in the absence of any prior early stopping. We define $p_k(z_k,\mathcal{I}_1,\ldots,\mathcal{I}_k)$ to be the conditional probability of following the sequence of information levels $\mathcal{I}_1,\ldots,\mathcal{I}_k$ to reach analysis k with information $\mathcal{I}_k$ and $Z_k = z_k$, given that Z_k takes this value when information is equal to $\mathcal{I}_k$. Again, this conditional probability does not dependent on θ. In place of (13) we now have

$$P_\theta(\text{"Superiority"}) = \sum_{K=1}^{k}\sum_{\{\mathcal{I}_1,\ldots,\mathcal{I}_k\}}\int_{d_k(\mathcal{I}_1,\ldots,\mathcal{I}_k)}^{\infty} f_k(z_k,\mathcal{I}_k;\theta)\,p_k(z_k,\mathcal{I}_1,\ldots,\mathcal{I}_k)\,dz_k.$$

As before, all the integrands in this equation are monotone increasing in θ, as long as each critical value $d_k(\mathcal{I}_1, \ldots, \mathcal{I}_k)$ is positive and, hence, the maximum type I error rate over $\theta \leq 0$ occurs at $\theta = 0$. Results for other error probabilities follow as before with the same conditions on critical values when these are expressed in terms of the final $\hat{\theta}_k$.

II Deriving optimal group sequential designs by solving Bayes decision problems

We illustrate our methods in the derivation of a design minimising F^* subject to error constraints (1) to (4). In this case, we place a five point prior distribution on θ with probability 1/5 at $-\delta_N$, $-\delta_N/2$, 0, $\delta_S/2$ and δ_S. We define a loss function associated with decisions on termination D_I: declare inferiority, D_N: declare-inferiority, and D_S: declare inferiority,

$$L(D, \theta) = \begin{cases} k_1 & \text{for } D = D_N \text{ or } D_S \text{ and } \theta = -\delta_N \\ k_2 & \text{for } D = D_S \text{ and } \theta = 0 \\ k_3 & \text{for } D = D_I \text{ and } \theta = 0 \\ k_4 & \text{for } D = D_I \text{ or } D_N \text{ and } \theta = \delta_S \\ 0 & \text{otherwise.} \end{cases} \tag{14}$$

With a cost $c(\theta) = 1$ per observation at each value of θ, the total expected cost is

$$F^* + \{k_1 P_{\theta=-\delta_N}(D_N \cup D_S) + k_2 P_{\theta=0}(D_S) + k_3 P_{\theta=0}(D_I) + k_4 P_{\theta=\delta_S}(D_I \cup D_N)\}/5.$$

We use dynamic programming, as described below, to solve this unconstrained Bayes decision problem for either non-adaptive or adaptive designs. It then remains to perform a numerical search for values of k_1, k_2, k_3 and k_4 which give a solution satisfying the error probability constraints (1) to (4). The standard Lagrangian argument implies that this Bayes sequential decision rule minimises F^* among all designs satisfying (1) to (4).

Consider first the non-adaptive case where n_k, $k = 1, \ldots, K$, are pre-specified. Let $p^{(k)}(\theta|z_k)$ denote the posterior distribution of θ given $Z_k = z_k$. If sampling continues until the final analysis K, a decision D_I, D_N, or D_S must be chosen. The critical values at this analysis are obtained by solving $k_3\, p^{(K)}(0|z_K) = k_1\, p^{(K)}(-\delta_N|z_K)$ to find b_K, and $k_4\, p^{(K)}(\delta_S|z_K) = k_2\, p^{(K)}(0|z_K)$ to find d_K; the monotone likelihood ratio property of the normal distribution implies each of these equations has a unique solution. The dynamic programming algorithm works backwards from this point to find the optimal decision rule at earlier analyses.

If $Z_k = z_k$, the expected loss on stopping to make the Bayes optimal decision at stage k is

$$\gamma^{(k)}(z_k) = \min\{k_3\, p^{(k)}(0|z_k) + k_4\, p^{(k)}(\delta_S|z_k),\ k_1\, p^{(k)}(-\delta_N|z_k) + k_4\, p^{(k)}(\delta_S|z_k),$$

$$k_1\, p^{(k)}(-\delta_N|z_k) + k_2\, p^{(k)}(0|z_k)\},$$

the minimum of the expected costs of stopping for inferiority, non-inferiority or superiority.

We denote by $F^{k+1}(z_{k+1}|z_k)$ the conditional cumulative distribution function of Z_{k+1} given $Z_k = z_k$. For $k = 1, \ldots, K-2$, the additional expected cost for proceeding from stage k to stage $k+1$ and acting optimally thereafter is

$$\begin{aligned}\beta^{(k)}(z_k) &= (n_{k+1} - n_k)\sum_{i=1}^{5} c(\theta_i)p^{(k)}(\theta_i|z_k) \\ &+ \int \min\{\beta^{(k+1)}(z_{k+1}), \gamma^{(k+1)}(z_{k+1})\} dF^{(k+1)}(z_{k+1}|z_k) \quad (15)\end{aligned}$$

while at stage $K-1$, we have

$$\begin{aligned}\beta^{(K-1)}(z_{K-1}) &= (n_K - n_{K-1})\sum_{i=1}^{5} c(\theta_i)\, p^{(K-1)}(\theta_i|z_{K-1}) \\ &+ \int \gamma^K(z_K) dF^{(K)}(z_K|z_{K-1}). \quad (16)\end{aligned}$$

The functions $\beta^{(k)}(z_k)$ can be calculated recursively, working backwards from analysis $K-1$: using the stage $k+1$ stopping boundary and values of $\beta^{(k+1)}$ and $\gamma^{(k+1)}$ previously calculated on a grid of z_{k+1} values, the integral in (15) can be found by numerical integration using, say, Simpson's rule. At each analysis k, the roots of $\beta^{(k)}(z_k) = \gamma^{(k)}(z_k)$ are found by a numerical search and these define the stopping boundaries.

The above method can be extended to find optimal adaptive designs using the approach followed by [7] for the two-decision problem. Consider the case $K = 2$, with n_1 fixed and n_2 allowed to take values in the set $\{n_{2,1}, \ldots, n_{2,M}\}$. We first find the M pairs of critical values $b_{2,m}$ and $d_{2,m}$ defining the optimal decisions when the second analysis takes place at cumulative sample size $n_{2,m}$, $m = 1, \ldots, M$. We then divide the range of values of Z_1 into intervals over which each of the following actions is found to be optimal: stop and declare inferiority, stop and declare non-inferiority, stop and declare superiority, continue to analysis 2 with cumulative group size $n_{2,m}$, $m = 1, \ldots, M$. As before, a numerical search is performed to find the set of costs k_1, k_2, k_3, and k_4 for which the solution satisfies the error probability constraints (1) to (4) and this gives the solution to the original constrained problem. This process is then nested within a search over n_1 to optimise both group sizes.

III Calculation of critical values for error spending designs

Consider an analysis k with $\gamma\,\mathcal{I}_{max,S} \leq \mathcal{I}_k \leq \mathcal{I}_{max,S}$, the case where all four critical values, a_k, b_k, c_k and d_k, are required. We assume boundary values for

analyses 1 to $k-1$ have already been calculated. Define the increments in error probabilities under $\theta = 0$ for analysis k

$$\Delta f_S^k = f_S(\mathcal{I}_k) - f_S(\mathcal{I}_{k-1}) \quad \text{and} \quad \Delta g_N^k = g_N(\mathcal{I}_k) - g_N(\mathcal{I}_{k-1}).$$

For the other two error probabilities, under $\theta = -\delta_N$ and δ_S, we set increments

$$\Delta f_N^k = f_N(\mathcal{I}_k) - f_N(\mathcal{I}_{k-1}) \quad \text{and} \quad \Delta g_S^k = g_S(\mathcal{I}_k) - g_S(\mathcal{I}_{k-1})$$

unless this is the first analysis with $\mathcal{I}_k \geq \gamma \mathcal{I}_{max,S}$, in which case we take

$$\Delta f_N^k = f_N(\mathcal{I}_k) - P_{\theta=-\delta_N}(\text{Stop to declare superiority by analysis } k-1)$$

and

$$\Delta g_S^k = g_S(\mathcal{I}_k) - P_{\theta=\delta_S}(\text{Stop to declare inferiority by analysis } k-1)$$

to account for the error probability incurred at analyses where $f_N(\mathcal{I})$ and $g_S(\mathcal{I})$ are zero.

We denote the continuation region at analysis i by $\mathcal{C}_i = [a_i, b_i] \cup [c_i, d_i]$. Two one-dimensional searches can be used to find a_k and d_k satisfying

$$P_{\theta=0}(Z_1 \in \mathcal{C}_1, \ldots, Z_{k-1} \in \mathcal{C}_{k-1},\, Z_k < a_k) = \Delta g_N^k$$

and

$$P_{\theta=0}(Z_1 \in \mathcal{C}_1, \ldots, Z_{k-1} \in \mathcal{C}_{k-1},\, Z_k > d_k) = \Delta f_S^k.$$

Let

$$\Delta f_N^{k1} = P_{\theta=-\delta_N}(Z_1 \in \mathcal{C}_1, \ldots, Z_{k-1} \in \mathcal{C}_{k-1},\, Z_k > d_k)$$

and

$$\Delta g_S^{k1} = P_{\theta=\delta_S}(Z_1 \in \mathcal{C}_1, \ldots, Z_{k-1} \in \mathcal{C}_{k-1},\, Z_k < a_k).$$

We now want to find b_k and c_k satisfying

$$P_{\theta=-\delta_N}(Z_1 \in \mathcal{C}_1, \ldots, Z_{k-1} \in \mathcal{C}_{k-1},\, Z_k \in [b_k, c_k]) = \Delta f_N^k - \Delta f_N^{k1} \tag{17}$$

and

$$P_{\theta=\delta_S}(Z_1 \in \mathcal{C}_1, \ldots, Z_{k-1} \in \mathcal{C}_{k-1},\, Z_k \in [b_k, c_k]) = \Delta g_S^k - \Delta g_S^{k1}. \tag{18}$$

Since b_k and c_k must lie in the interval $[a_k, d_k]$, an upper bound b_k^u for b_k is found by solving

$$P_{\theta=-\delta_N}(Z_1 \in \mathcal{C}_1, \ldots, Z_{k-1} \in \mathcal{C}_{k-1},\, Z_k \in [b_k^u, d_k]) = \Delta f_N^k - \Delta f_N^{k1} \tag{19}$$

and a lower bound c_k^l for c_k by solving

$$P_{\theta=\delta_S}(Z_1 \in \mathcal{C}_1, \ldots, Z_{k-1} \in \mathcal{C}_{k-1},\, Z_k \in [a_k, c_k^l]) = \Delta g_S^k - \Delta g_S^{k1}. \tag{20}$$

Using these values of b_k^u and c_k^l, we can now find a lower bound b_k^l for b_k as the solution to

$$P_{\theta=-\delta_N}(Z_1 \in \mathcal{C}_1, \ldots, Z_{k-1} \in \mathcal{C}_{k-1},\, Z_k \in [b_k^l, c_k^l]) = \Delta f_N^k - \Delta f_N^{k1}$$

and an upper bound c_k^u for c_k as the solution to

$$P_{\theta=\delta_S}(Z_1 \in \mathcal{C}_1, \ldots, Z_{k-1} \in \mathcal{C}_{k-1},\ Z_k \in [b_k^u, c_k^u]) = \Delta g_S^k - \Delta g_S^{k1}.$$

We have now reduced the original interval $[a_k, d_k]$ to $[b_k^l, c_k^u]$ and can repeat the same steps with c_k^u in place of d_k in (19) and b_k^l in place of a_k in (20). We have found repeated iterations of these steps to give an efficient method for finding b_k and c_k satisfying (17) and (18).

References

[1] Morikawa T, Yoshida M. A useful testing strategy in phase III trials: combined test of superiority and test of equivalence. *Journal of Biopharmaceutical Statistics* 1995; **5**:297–306.

[2] Wang SJ, Hung HMJ, Tsong Y, Cui L. Group sequential test strategies for superiority and non-inferiority hypotheses in active controlled clinical trials. *Statistics in Medicine* 2001; **20**:1903–1912.

[3] Cui L, Hung H, Wang S. Modification of sample size in group sequential trials. *Biometrics* 1999; **55**:853–857.

[4] Shih WJ, Quan H, Li G. Two-stage adaptive strategy for superiority and non-inferiority hypotheses in active controlled clinical trials. *Statistics in Medicine* 2004; **23**:2781–2798.

[5] Koyama T, Sampson AR, Gleser LJ. A framework for two-stage adaptive procedures to simultaneously test non-inferiority and superiority. *Statistics in Medicine* 2005; **24**:2439–2456.

[6] Lai TL, Shih MC, Zhu G. Modified Haybittle-Peto group sequential designs for testing superiority and non-inferiority hypotheses in clinical trials. *Statistics in Medicine* 2006; **25**:1149–1167.

[7] Jennison C, Turnbull BW. Adaptive and nonadaptive group sequential tests. *Biometrika* 2006; **93**:1–21.

[8] Gould AL. 3-decision rules useful for assessing superiority or equivalence. *Enar Biometrics Society*, Memphis, Tennessee, 25 March 1997.

[9] Jennison C, Turnbull BW. *Group Sequential Methods with Applications to Clinical Trials.* Chapman & Hall/CRC: London, 2000.

[10] Eales JD, Jennison C. An improved method for deriving optimal one-sided group sequential tests. *Biometrika* 1992; **79**:13–24.

[11] Eales JD, Jennison C. Optimal two-sided group sequential tests. *Sequential Analysis* 1995; **14**:273–286.

[12] Barber S, Jennison C. Optimal asymmetric one-sided group sequential tests. *Biometrika* 2002; **89**:49–60.

[13] Jennison C, Turnbull BW. Group sequential analysis incorporating covariate information. *Journal of the American Statistical Association* 1997; **92**:1330–1341.

[14] Lan KKG, DeMets DL. Discrete sequential boundaries for clinical trials. *Biometrika* 1983; **70**:659–663.

[15] Jennison C, Turnbull BW. Efficient group sequential designs when there are several effect sizes under consideration. *Statistics in Medicine* 2006; **25**:917–932.

[16] Brannath W, Bauer P, Maurer W, Posch M. Sequential tests for noninferiority and superiority. *Biometrics* 2003; **59**:106–114.

[17] O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. *Biometrics* 1979; **35**:549–556.

[18] Pocock SJ. Group sequential methods in the design and analysis of clinical trials. *Biometrika* 1977; **64**:191–199.

[19] Mehta CR, Tsiatis AA. Flexible sample size considerations using information-based interim monitoring. *Drug Information Journal* 2001; **35**:1095–1112.

[20] Schmitz N. *Optimal Sequentially Planned Decision Procedures.* Lecture Notes in Statistics; **79**, Springer-Verlag: New York, 1993.

[21] EMEA note for guidance on clinical investigation of medicinal products of diabetes mellitus.
http://www.emea.europa.eu/pdfs/human/ewp/108000en.pdf.
[14 Sep 2009].

[22] Home PD, Jones NP, Pocock SJ, Beck-Nielsen H, Gomis R, Hanefeld M, Komajda M, Curtis P. Rosiglitazone record study: glucose control outcomes at 18 months. *Diabetic Medicine* 2007; **24**:626–634.

[23] Göke B, Gausse-Nilsson I, Persson A. The effect of tesaglitazar as add-on treatment to metformin in patients with poorly controlled type 2 diabetes. *Diabetes and Vascular Disease Research* 2007; **4**:204–213.

[24] Jennison C, Turnbull BW. Mid-course sample size modification in clinical trials based on the observed treatment effect. *Statistics in Medicine* 2003; **22**:971–993.