Efficient group sequential designs when there are several effect sizes under consideration

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

and

Bruce W. Turnbull
School of Operations Research and Industrial Engineering,
Cornell University, USA

August 2, 2004

SUMMARY

We consider the construction of efficient group sequential designs where the goal is a low expected sample size not only at the null hypothesis and the alternative (taken to be the minimal clinically meaningful effect size), but also at more optimistic anticipated effect sizes. Pre-specified Type I error rate and power requirements can be achieved both by standard group sequential tests and by more recently proposed adaptive procedures. We investigate four nested classes of designs: (A) Group sequential tests with equal group sizes and stopping boundaries determined by a monomial error spending function (the "ρ-family"); (B) As A but the initial group size is allowed to be different from the others; (C) Group sequential tests with arbitrary group sizes and arbitrary boundaries, fixed in advance; (D) Adaptive tests — as C but at each analysis, future group sizes and critical values are updated depending on the current value of the test statistic. By examining the performance of optimal procedures within each class, we conclude that class B provides simple and efficient designs with efficiency close to that of the more complex designs of classes C and D. We provide tables and figures illustrating the performances of optimal designs within each class and defining the optimal procedures of classes A and B.

Key words: clinical trial; group sequential test; sample size re-estimation; adaptive design; flexible design; optimal design; error spending function.

1 Introduction

Along with practical considerations, the sample size for a clinical trial is determined by setting up null and alternate hypotheses concerning a primary parameter of interest, θ, and then specifying a Type I error rate α and power $1-\beta$ to be controlled at a given treatment effect size $\theta = \Delta$. Usually, traditional values of α and β are used (e.g., $\alpha = 0.025, 0.05$, $\beta = 0.05, 0.1, 0.2$); however, there can be much debate

over the choice of Δ. Some textbooks advocate that Δ should be chosen to represent the minimum "clinically relevant" or "commercially viable" effect size — see for example Senn (1997, p. 170) and Piantadosi (1997, p. 149). Others such as Shun et al. (2001) say that Δ can be taken to be the anticipated effect size — a value based on expectations from prior experimental, observational and theoretical evidence. Pocock (1983) suggests that either approach might be taken: on pages 125 and 132, Δ is to be a "realistic value", while in the example on page 128, it is to be a "clinically relevant" difference that is "important to detect". In Section 3.5 of the ICH Guidance E9 (Food and Drug Administration, 1998), it is also stated that Δ is to be based on a judgement concerning either the minimal clinically relevant effect size or the "anticipated" effect.

The choice of Δ is crucial because, for example, a halving in the chosen effect size will lead to a quadrupling in the sample size for a fixed sample test (and in the maximum sample size for a group sequential test). Using the lower sample size appropriate to a high treatment effect will leave the trial underpowered to detect a smaller but still important effect. Because of this, Shun et al. (2001) and others have proposed that the trial be designed using the higher effect size (and corresponding lower sample size), but that sample size be re-estimated at an interim analysis based on the emerging observed treatment difference. This has been termed the "start small then ask for more" strategy (Anderson and Liu, 2004).

There have been several accounts in the literature of studies in which sample size has been adapted in order to increase power at lower effect sizes. Cui et al. (1999, Sec. 1) report on a placebo controlled myocardial infarction prevention trial with a sample size of 600 subjects per treatment arm, this number being based on a planned effect size of a 50% reduction in incidence and 95% power. However midway through the trial, only about a 25% reduction in incidence was observed, a reduction which was still of clinical and commercial importance. Because of the low conditional power at this stage, the sponsor of the trial submitted a proposal to expand the sample size. In recent years, classes of procedures termed "flexible", "adaptive", "self-designing" or "variance spending" have been developed which enable such sample size re-estimation to be done while preserving the Type I error rate α. See Bauer (1989), Proschan and Hunsberger (1995), Fisher (1998), Cui et al. (1999), Wassmer (2000), Li et al. (2002) and Posch et al. (2003) among others.

Remarks by some authors, e.g., Shen and Fisher (1999) and Shun et al. (2001), suggest a desire to set a specific power, $1 - \beta$, at whatever is the true value of the effect size parameter. This aim may lead to adaptive designs with a power curve rising sharply from α at $\theta = 0$, then remaining almost flat at $1 - \beta$. In consequence, significant risk of a negative outcome remains even when the effect size is high and power close to one could easily have been attained.

All the above discussion supports the view that a clinical trial should guarantee power at effect sizes θ of clinical or commercial interest. Smaller effects are not pertinent since, as Shih (2001, p. 517) states "... trials need to consider sample size to detect a difference that is clinically meaningful, not merely to find a statistical significance." Limitations occur when the sample size needed to detect a particularly small effect is prohibitive: then, power must be specified at the smallest value of θ that resources permit.

Shun et al. (2001, p. 520) give an example of the dilemma investigators can face: It is agreed that the minimum clinically meaningful effect is $\theta = 5$ but the anticipated effect size if $\theta = 10$. Should the trial be planned with the large sample size necessary to deliver power under $\theta = 5$? Or, should one start with the lower sample size required for power under $\theta = 10$ then carry out a review when interim data are

available to estimate θ, increasing the sample size if this interim estimate is lower than 10? Shun et al. (2001) favour the second, adaptive approach "in order to possibly save resources". They do not, however, consider the option of a sequential test designed to achieve power under $\theta = 5$ which would be likely to stop early, after only a fraction of the planned maximum sample size, if the true value of θ were as high as 10. Our objective in this paper is to compare the non-adaptive and adaptive approaches to the design of such a study.

Our premise is that a common goal underlies both traditional, non-adaptive group sequential tests and more recent adaptive designs and this goal can be phrased in terms of the overall power curve, i.e., the power investigators wish their study to achieve as a function of the true value of θ. For example, in the situation described by Shun et al. (2001) it is clear that high power is required if θ is actually equal to 5, so the overall power curve of an acceptable design must take a suitably high value at $\theta = 5$. Note that this requirement holds independently of any additional evidence that θ is equal to 5: the point is that *if* $\theta = 5$, then the overall design package should provide a high probability of declaring a positive finding.

Many proposals for adaptive study designs base mid-course revisions of the remaining sample size on the *conditional* power under particular θ values. It is important not to confuse this conditional power with the overall power curve of the study design. Once the prescription for an adaptive trial design is set out as, for example, in the procedures defined by Proschan and Hunsberger (1995) or Li et al. (2002), the plan the trial will follow as it progresses is completely defined and one can compute the procedure's overall power curve.

An attraction of certain adaptive schemes is that they allow freedom to make changes in response to interim data that were not considered before the study started. Even then one can, at least conceptually, define a specific procedure by asking what changes would have been made under all possible sequences of observations. Then, the performance of this realisation of a "flexible monitoring scheme" is that of the pre-specified adaptive design defined by these rules. Thus, in evaluating the class of pre-specified adaptive designs, we shall also learn what flexible designs, which are not fully pre-specified, can offer.

The flexibility of adaptive designs can also be used to react to external information that affects a study's objectives. While this is an important property, it is quite distinct from the issue we shall address in our comparisons of sequential designs. In what follows, we confine our attention to the relatively simple situation where the external environment remains constant during the course of a study. We comment further on adaptation to external factors in Section 5.

In comparing designs, we shall study tests with a specific Type I error rate, α, and power $1 - \beta$ at a given effect size $\theta = \Delta$, noting it is the overall power of the complete procedure that we consider here, however this is achieved. We shall compare designs with respect to their expected sample size (or average sample number, "ASN") over the range of θ values of interest. Since Δ represents the minimum clinically or commercially meaningful effect, ASN should also be considered at the larger effect size that investigators are hoping for, $\theta = L\Delta$, say, where $L > 1$. Combining ASN at these two points with the ASN under the null hypothesis $\theta = 0$ gives our optimality criterion

$$\frac{1}{3}\{ASN(\theta = 0) + ASN(\theta = \Delta) + ASN(\theta = L\Delta)\}. \quad (1)$$

This criterion reflects equipoise and a balance between optimism and pessimism. We have also made investigations using other criteria, generalised to include additional θ values or varying the assigned weights, and obtained qualitatively similar conclusions.

We shall assess three classes, A, B and C, of conventional group sequential tests, moving from simple to more complex designs. There are a great many proposals from which to choose types of adaptive design and in our studies of these methods we have often found adaptive designs to have inferior performance to well chosen non-adaptive group sequential tests; see Jennison and Turnbull (2003) for one such case study. However, the class of adaptive tests contains non-adaptive tests as special cases so it is clear that efficient adaptive designs must exist! As our fourth class, D, we consider all possible adaptive designs, i.e., group sequential tests in which, at any stage, future group sizes and decision boundaries can be adapted to observed responses in a pre-planned way. The idea of seeking optimal tests within this class was proposed by Schmitz (1993) who referred to these as "sequentially planned sequential designs". Since this class contains all possible adaptive designs, the optimal designs we find for a given criterion will perform at least as well, by this criterion, as any adaptive or flexible test that one might propose.

The recent paper of Schäfer and Müller (2004) has also addressed the issue of uncertainty about the effect size to be detected, taking a somewhat different approach. For a K-group trial with Type I error rate α, these authors consider a sequence of K values of θ, $L_1\Delta < \ldots < L_K\Delta$, and show how to construct boundaries such that, for each $k = 1, \ldots, K$, the probability of stopping to reject H_0 at or before the kth analysis is equal to $1 - \beta$ when $\theta = L_k\Delta$. In contrast, our emphasis is on the ASN under different θ values, subject to the goal of protecting power at the most pessimistic effect size $\theta = \Delta$. Schäfer and Müller's (2004) procedures belong to our class C and, hence, results for the optimal test in class C provide an upper bound on how well these designs perform by our criteria.

2 Group sequential procedures

Consider a balanced two-sample comparison in which observations X_{Ai} on treatment A and X_{Bi} on treatment B, $i = 1, 2, \ldots$, are independent, normally distributed with common, known variance σ^2 and unknown means μ_A and μ_B, respectively. The parameter of interest is the difference in treatment means, $\theta = \mu_A - \mu_B$, and it is desired to test the null hypothesis H_0: $\theta = 0$ against the one-sided alternative $\theta > 0$ with Type I error probability α. Although this problem may seem unrealistically simple, it does in fact serve as a prototype for a wide variety of designs and response types: methods developed for this situation have wide applicability — see, for example, Jennison and Turnbull (2000, Chap. 3).

The assumption of known variance is significant here as it means we are not concerned with adapting sample size in response to new estimates of σ^2. There is a considerable literature on adaptive methods for "re-estimating" the sample size needed to meet a fixed power requirement as more is learnt about a nuisance parameter; see, for example, Wittes and Brittain (1990) and Gould and Shih (1992) or, for updating sample size in a group sequential test, Denne and Jennison (2000) and Mehta and Tsiatis (2001). However, we do not consider this question in the present paper. We refer readers to Schwartz and Denne (2003) for a careful explanation of how multi-stage designs can incorporate one or more of the objectives of: updating sample size in response to new variance estimates; adjusting sample size as the effect size to be detected is modified; stopping early according to a group sequential rule as soon as the observed data permit this.

For our stated problem, a fixed sample test attaining power $1 - \beta$ at the alternative $\theta = \Delta$ requires a

sample size

$$n_f = \frac{2\sigma^2(z_\alpha + z_\beta)^2}{\Delta^2} \tag{2}$$

per treatment arm where $z_p = \Phi^{-1}(1-p)$ denotes the upper p tail point of the standard normal distribution. The test statistic is given by

$$Z = \frac{1}{\sqrt{(2\sigma^2 n_f)}} \sum_{i=1}^{n_f} (X_{Ai} - X_{Bi}) \tag{3}$$

and H_0 is rejected if and only if $Z > z_\alpha$.

In a group sequential design, the accumulating data are analysed repeatedly. Suppose the maximum number of analyses is set at K and, for $k = 1, \ldots, K$, the cumulative number of observations per treatment arm at analysis k is n_k, where of course $n_1 < \ldots < n_K$. At analysis k, the statistic

$$Z_k = \sum_{i=1}^{n_k} (X_{Ai} - X_{Bi})/(\sigma\sqrt{2n_k})$$

is computed. A general one-sided group sequential test is defined by pairs of constants (a_k, b_k) with $a_k < b_k$ for $k = 1, \ldots, K-1$ and $a_K = b_K$. It takes the form:

$$\begin{array}{ll} \text{After group } k = 1, \ldots, K-1 & \\ \quad \text{if } Z_k \geq b_k & \text{stop, reject } H_0 \\ \quad \text{if } Z_k \leq a_k & \text{stop, accept } H_0 \\ \quad \text{otherwise} & \text{continue to group } k+1, \\ \text{after group } K & \\ \quad \text{if } Z_K \geq b_K & \text{stop, reject } H_0 \\ \quad \text{if } Z_K < a_K & \text{stop, accept } H_0. \end{array} \tag{4}$$

Here, $a_K = b_K$ ensures that the test terminates at analysis K.

In order for the group sequential test to achieve Type I error probability α and power $1-\beta$ at $\theta = \Delta$, its maximum sample, n_K, must be at least a little larger than the fixed sample size n_f. The ratio $R = n_K/n_f$ is termed the *inflation factor* of a group sequential test.

There are many choices of boundary $\{(a_k, b_k);\ k = 1, \ldots, K\}$, that will satisfy the α and β error probability requirements. One convenient definition is through the specification of a parametric error spending function. Following Jennison and Turnbull (2000, Chap. 7.3.2), we define two error spending functions $f(t)$ and $g(t)$, for Type I and Type II errors respectively, which are non-decreasing and satisfy $f(0) = g(0) = 0$, $f(1) = \alpha$ and $g(1) = \beta$. The critical values (a_k, b_k), $k = 1, \ldots, K$, are calculated by solving successively

$$Pr_{\theta=0}\{a_1 < Z_1 < b_1,\ \ldots,\ a_{k-1} < Z_{k-1} < b_{k-1},\ Z_k \geq b_k\} = f(n_k/n_K) - f(n_{k-1}/n_K) \tag{5}$$

and

$$Pr_{\theta=\delta}\{a_1 < Z_1 < b_1,\ \ldots,\ a_{k-1} < Z_{k-1} < b_{k-1},\ Z_k \leq a_k\} = g(n_k/n_K) - g(n_{k-1}/n_K), \tag{6}$$

starting with a_1 and b_1 at the first analysis, where we take $n_0 = 0$, and proceeding up to $k = K$. For given functions $f(t)$ and $g(t)$, the requirement $a_K = b_K$ at the final analysis determines the necessary maximum sample size n_K and, hence, the test's inflation factor $R = n_K/n_f$. Various software programs are available to aid this computation; for example, the commercial packages `East3` (Cytel, 2003) and `S+SeqTrial 2.0` (Insightful Corp, 2002), or those freely available at the website `http://www.medsch.wisc.edu/landemets/`.

We shall consider the simple error spending functions:

$$f(t) = \alpha t^{\rho} \quad \text{and} \quad g(t) = \beta t^{\rho}, \tag{7}$$

parametrized by the single parameter $\rho > 0$. This choice defines the so-called "ρ-family" of tests. The value of ρ determines the inflation factor $R = n_K/n_f$; this relationship is illustrated in Table 7.6 of Jennison and Turnbull (2000) for the case of equally sized groups and selected values of K, α and β.

A referee has commented on our assumption that the study will definitely stop with acceptance of H_0 if ever $Z_k \leq a_k$ and noted that if sampling were to continue in such circumstances, this would inflate the Type I error rate above α. If there is concern that the Data and Safety Monitoring Board may not adhere rigidly to the stopping boundaries specified in the study protocol, the Type I error rate can be protected by replacing (5) with

$$Pr_{\theta=0}\{Z_1 < b_1,\, \ldots, Z_{k-1} < b_{k-1},\, Z_k \geq b_k\} = f(n_k/n_K) - f(n_{k-1}/n_K).$$

Then, if the stopping rule is followed precisely, the Type I error rate will be achieved conservatively. Since this conservatism would complicate our comparison of different types of test, we shall continue with our original definition for the remainder of this paper.

We shall study four families of group sequential tests (GSTs) comprising two forms of error spending test, general non-adaptive GSTs, and adaptive GSTs. To ensure a fair comparison, tests in all families must meet the same design specifications given by α, β, Δ and K. It is well known that group sequential tests can yield greater reductions in ASN as the inflation factor R increases; see, for example, Eales and Jennison (1992). Practical constraints will often limit the maximum possible sample size and, hence, the values allowable for R. We have derived optimal tests for fixed values of R but we also report results without this constraint, obtained by optimising freely over R.

A. **Equal group sizes.** These tests are ρ-family error spending tests with equally spaced analyses: $n_k = (k/K)R\,n_f$, $k = 1, \ldots, K$. Since the value of ρ is determined by R, there is only one test for a given design specification. Planning with equal group sizes is the usual starting point when designing a GST.

B. **Proposed class of GSTs.** These are ρ-family error spending tests with one degree of freedom in setting group sizes. The first group size is n_1 and remaining analyses are equally spaced thereafter: $n_k = n_1 + (R\,n_f - n_1)(k-1)/(K-1)$, $k = 2, \ldots, K$. If the treatment difference is as high as $L\Delta$, where $L > 1$, there may be opportunity for early stopping well before the first analysis in procedure A: scheduling an early look helps lower the ASN in this case, reducing the overall criterion (1). Optimization is over the choice of n_1, after which ρ is determined by R.

C. **Optimal non-adaptive GSTs.** Complete freedom is allowed in choosing critical values (a_k, b_k) *and* cumulative sample sizes $n_1, \ldots, n_K$ to optimize the criterion (1), subject to $n_K \leq R\,n_f$. Note

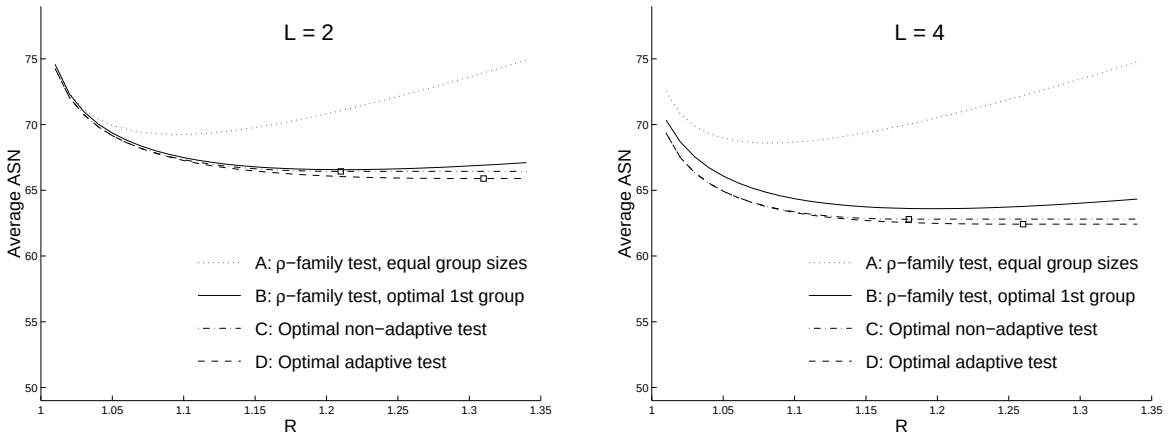

Figure 1: Optimized average ASN plotted against R for tests in classes A to D. Tests have Type I error rate $\alpha = 0.025$, power $1 - \beta = 0.8$ at $\theta = \Delta$, a maximum of $K = 2$ analyses, and minimize the average ASN criterion $\{ASN(\theta = 0) + ASN(\theta = \Delta) + ASN(\theta = L\Delta)\}/3$ for $L = 2$ and 4. The symbols □ on the curves for classes C and D indicate the values of R at which these curves become flat.

that the choice of n_k and (a_k, b_k), $k = 1, \ldots, K$, is set at the start of the study and cannot be updated as observations accrue.

D. **Optimal adaptive GSTs.** These are fully adaptive designs. At each analysis $k = 1, \ldots, K-1$, the next cumulative group size n_{k+1} and critical values (a_{k+1}, b_{k+1}) are chosen based on current data. The whole procedure is chosen to optimize criterion (1) subject to $n_K \leq R\, n_f$.

Optimal non-adaptive tests for a fixed sequence of group sizes are derived using methods described in Eales and Jennison (1992) for the symmetric case, $\alpha = \beta$, and by Barber and Jennison (2002) for the asymmetric case. A Bayes decision theory problem is set up and the solution found by a backward induction (dynamic programming) technique. A search over decision and sampling costs leads to a Bayes problem with solution equal to the optimal GST being sought. Further optimization over group sizes $n_1, \ldots, n_K$ described by Eales and Jennison (1992, Sec. 5) provides the optimal tests within class C. Construction of the optimal adaptive or "Schmitz" tests in class D proceeds by use of the dynamic programming technique in a generalization of the Bayes decision problem where group size is allowed to depend on the current value of the statistic Z_k.

3 Optimal tests and their properties

We present results for tests with Type I error rate $\alpha = 0.025$, power $1-\beta = 0.8$ at $\theta = \Delta$ and a maximum of K analyses where $K = 2$, 3, 4, 5 or 6. Tests are optimized for the average ASN criterion (1) with $L = 2$ and 4. All calculations are by numerical integration and results are accurate to at least the number of decimal places shown in the tables.

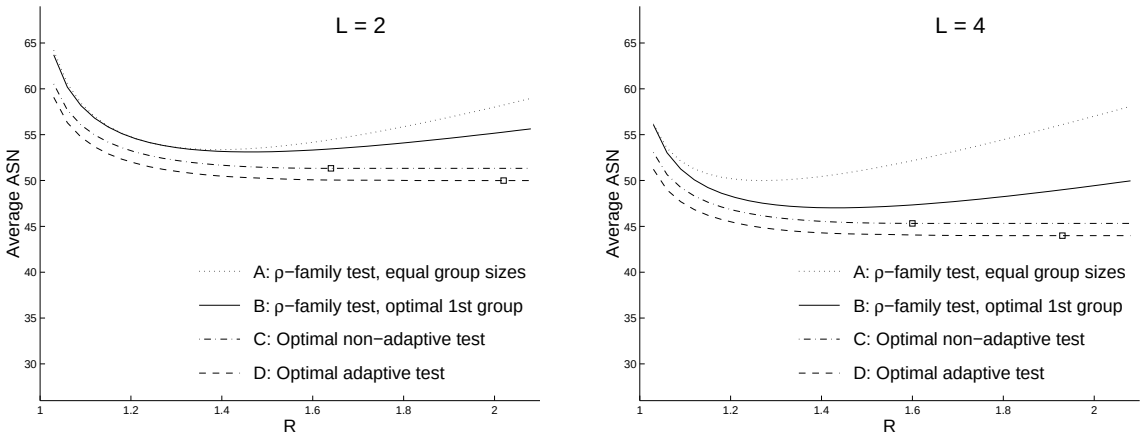

Figure 2: Optimized average ASN plotted against R for tests in classes A to D. Tests have Type I error rate $\alpha = 0.025$, power $1 - \beta = 0.8$ at $\theta = \Delta$, a maximum of $K = 5$ analyses, and minimize the average ASN criterion $\{ASN(\theta = 0) + ASN(\theta = \Delta) + ASN(\theta = L\Delta)\}/3$ for $L = 2$ and 4. The symbols □ on the curves for classes C and D indicate the values of R at which these curves become flat.

Figures 1 and 2 show how the optimized average ASN varies with R for tests with 2 and 5 analyses, respectively. For both classes A and B of ρ-family tests, average ASN decreases initially as R increases but eventually starts to increase again. The group sizes of class C tests are optimized subject to the constraint $n_K/n_f \leq R$: initially the optimal tests have $n_K/n_f = R$ but as R is increased a value, $\tilde{R}_C$ say, is reached such that the optimal test continues to take $n_K/n_f = \tilde{R}_C$ even though higher values are allowed. Similarly, for given L and K, there is a maximal inflation factor, $\tilde{R}_D$ say, for optimal tests in class D: even when larger values of n_K are permitted, all sample paths terminate with $n_K \leq \tilde{R}_D n_f$. The points at which the curves for classes C and D reach their plateaus are marked in the figures but it is clear that values close to the optimal average ASN are reached well before these points. Given the practical disadvantages of a high value of R, it is reasonable to choose an inflation factor considerably lower than these otherwise "optimal" values.

Tables 1 and 2 give further details of the procedures with optimal performance over R in Figures 1 and 2, respectively, taking $R = \tilde{R}_C$ for class C and $R = \tilde{R}_D$ for class D. In addition, these tables give results for tests with 3, 4 and 6 analyses. In view of the practical benefits of a low maximum sample size, we would recommend choosing values for the inflation factor R below the global optimum. Tables 3 and 4 give details of tests optimising criterion (1), with $L = 2$ and 4 respectively, with R fixed at 1.2; however, where the minimum average ASN occurs at a value of R below 1.2, this value is used instead. Further tables of results, similar to Tables 3 and 4 with $R = 1.05$, 1.1, 1.2, 1.3 and 1.4 and power 0.8 and 0.9 are available at our website `http://www.bath.ac.uk/~mascj/`.

The first column of each table shows the maximum number of analyses K, the fixed sample test with $K = 1$ serving as a benchmark for the others. For each $K \geq 2$, properties of the optimal tests in classes A to D are presented. The value of ρ is given for the ρ-family tests in classes A and B, then, for all four tests, we show the inflation factor R, the size of the first group of observations, the ASNs under $\theta = 0$, Δ and $L\Delta$, and the average of these three ASNs which forms the optimality criterion (1). All group sizes

Table 1: Properties of tests with Type I error rate $\alpha = 0.025$ and power $1 - \beta = 0.8$ at $\theta = \Delta$ which minimize criterion (1) with $L = 2$. The value of R is also chosen optimally to minimize this criterion.

K	Design	ρ	R	First group	ASN at $\theta=0$	ASN at $\theta=\Delta$	ASN at $\theta=2\Delta$	Average ASN
1	Fixed		1.00	100.0	100.0	100.0	100.0	100.0
2	A: ρ-family, equal groups	1.36	1.09	54.5	68.1	83.3	56.4	69.3
	B: ρ-family, opt. 1st group	0.67	1.21	43.0	64.5	86.3	48.9	66.6
	C: Optimal non-adaptive	–	1.21	43.0	63.2	86.8	49.3	66.4
	D: Optimal adaptive	–	1.31	41.9	63.9	85.7	48.1	65.9
3	A: ρ-family, equal groups	0.96	1.21	40.3	58.5	77.1	45.2	60.3
	B: ρ-family, opt. 1st group	0.61	1.34	33.1	56.4	78.7	42.1	59.1
	C: Optimal non-adaptive	–	1.38	30.5	53.5	79.9	40.0	57.8
	D: Optimal adaptive	–	1.57	28.3	53.7	78.6	38.2	56.8
4	A: ρ-family, equal groups	0.77	1.31	32.8	53.5	74.3	39.9	55.9
	B: ρ-family, opt. 1st group	0.59	1.41	27.6	52.3	75.4	38.4	55.4
	C: Optimal non-adaptive	–	1.53	24.4	48.9	76.2	36.0	53.7
	D: Optimal adaptive	–	1.80	21.6	48.8	74.8	33.9	52.5
5	A: ρ-family, equal groups	0.67	1.39	27.8	50.6	72.8	36.8	53.4
	B: ρ-family, opt. 1st group	0.56	1.46	23.9	49.8	73.6	35.9	53.1
	C: Optimal non-adaptive	–	1.64	20.9	46.4	73.9	33.7	51.3
	D: Optimal adaptive	–	2.02	20.2	45.9	72.0	32.1	50.0
6	A: ρ-family, equal groups	0.60	1.45	24.2	48.6	71.9	34.7	51.7
	B: ρ-family, opt. 1st group	0.55	1.49	21.8	48.2	72.4	34.3	51.6
	C: Optimal non-adaptive	–	1.77	18.5	44.7	72.3	32.2	49.8
	D: Optimal adaptive	–	2.21	17.7	44.1	70.4	30.7	48.4

and ASNs are expressed as a percentage of n_f, the sample size required by a fixed sample test.

Our key conclusion from these figures and tables is that ρ-family error spending tests offer simple and efficient designs for the objectives under consideration. The robust efficiency of ρ-family tests when compared to general non-adaptive GSTs with equal group sizes has been noted by Barber and Jennison (2002). It is evident from our results that restricting to equal group sizes hinders the reduction of ASN at the high effect sizes of $\theta = 2\Delta$ and $\theta = 4\Delta$. However, allowing choice of just the first group size in class B tests overcomes this difficulty, leading to noticeable improvements over the class A designs for $K = 2$ when $L = 2$ and for K up to 4 or 5 when $L = 4$. The first group sizes in the optimal class B test are much lower for $L = 4$ than for $L = 2$, showing how optimism about the size of treatment effect leads to a more aggressive strategy in the hope of a "home run" finding under $\theta = L\Delta$. But, despite the focus when $L = 4$ on a very low ASN under $\theta = 4\Delta$, ASNs at $\theta = 0$ and Δ are still quite similar to

Table 2: Properties of tests with Type I error rate $\alpha = 0.025$ and power $1 - \beta = 0.8$ at $\theta = \Delta$ which minimize criterion (1) with $L = 4$. The value of R is also chosen optimally to minimize this criterion.

K	Design	ρ	R	First group	ASN at $\theta=0$	ASN at $\theta=\Delta$	ASN at $\theta=4\Delta$	Average ASN
1	Fixed		1.00	100.0	100.0	100.0	100.0	100.0
2	A: ρ-family, equal groups	1.46	1.08	54.0	68.3	83.5	54.0	68.6
	B: ρ-family, opt. 1st group	0.64	1.20	32.6	67.1	91.1	32.6	63.6
	C: Optimal non-adaptive	–	1.18	30.0	64.1	94.3	30.0	62.8
	D: Optimal adaptive	–	1.26	30.2	64.4	92.6	30.2	62.4
3	A: ρ-family, equal groups	1.19	1.16	38.7	59.3	77.5	38.7	58.5
	B: ρ-family, opt. 1st group	0.68	1.29	17.8	60.4	82.1	18.5	53.7
	C: Optimal non-adaptive	–	1.35	17.6	54.3	82.6	18.8	51.9
	D: Optimal adaptive	–	1.50	18.0	53.6	81.3	18.7	51.2
4	A: ρ-family, equal groups	1.05	1.22	30.5	54.7	74.5	30.5	53.2
	B: ρ-family, opt. 1st group	0.63	1.38	14.8	54.3	77.3	16.2	49.3
	C: Optimal non-adaptive	–	1.51	14.9	49.1	77.3	16.6	47.7
	D: Optimal adaptive	–	1.72	13.8	48.5	75.9	15.2	46.5
5	A: ρ-family, equal groups	0.95	1.27	25.4	51.8	72.7	25.4	50.0
	B: ρ-family, opt. 1st group	0.61	1.43	13.3	51.0	74.8	15.2	47.0
	C: Optimal non-adaptive	–	1.60	12.6	46.4	74.8	14.8	45.3
	D: Optimal adaptive	–	1.93	11.6	45.6	72.9	13.5	44.0
6	A: ρ-family, equal groups	0.88	1.31	21.8	49.9	71.6	21.9	47.8
	B: ρ-family, opt. 1st group	0.60	1.46	12.2	49.1	73.3	14.4	45.6
	C: Optimal non-adaptive	–	1.69	10.5	44.8	73.1	13.2	43.7
	D: Optimal adaptive	–	2.11	8.4	44.1	71.4	11.8	42.4

those for the case $L = 2$. Comparisons with optimum non-adaptive tests in class C show that very little efficiency is lost by settling on the ρ-family tests of class B: adding a free choice of all group sizes and stopping boundary $\{(a_k, b_k);\ k = 1, \ldots, K\}$ reduces average ASN by at most 2% of n_f.

If we compare optimal class C and class D tests, either optimized over R or at a fixed value of R, the improvement in passing from optimal non-adaptive tests to optimal adaptive tests is a drop in average ASN of at most 1.4% of n_f. As we explained earlier, class D contains all varieties of adaptive test, thus our results provide an upper bound on how well specific proposals, for example, those of Proschan and Hunsberger (1995) or Shen and Fisher (1999), can perform when assessed in terms of their overall power and ASN functions. Comparing our results for tests in classes B and D, we deduce that for the cases considered, using criterion (1) with $L = 2$ or $L = 4$, ρ-family error spending tests with an optimized first group size have average ASN within 3.2% of n_f of the average ASN for any adaptive design meeting the

Table 3: Properties of tests with Type I error rate $\alpha = 0.025$ and power $1 - \beta = 0.8$ at $\theta = \Delta$ which minimize criterion (1) with $L = 2$. The value of R is fixed at 1.2 unless a lower value minimizes (1).

K	Design	ρ	R	First group	ASN at $\theta = 0$	ASN at $\theta = \Delta$	ASN at $\theta = 2\Delta$	Average ASN
1	Fixed		1.00	100.0	100.0	100.0	100.0	100.0
2	A: ρ-family, equal groups	1.36	1.09	54.5	68.1	83.3	56.4	69.3
	B: ρ-family, opt. 1st group	0.69	1.20	43.0	64.6	86.2	48.9	66.6
	C: Optimal non-adaptive	–	1.20	43.0	63.3	86.7	49.3	66.4
	D: Optimal adaptive	–	1.20	43.2	64.0	85.3	49.0	66.1
3	A: ρ-family, equal groups	1.00	1.20	40.0	58.6	77.2	45.1	60.3
	B: ρ-family, opt. 1st group	0.99	1.20	33.8	58.1	78.0	43.0	59.7
	C: Optimal non-adaptive	–	1.20	31.1	55.1	79.5	40.9	58.5
	D: Optimal adaptive	–	1.20	28.8	54.9	78.9	39.5	57.8
4	A: ρ-family, equal groups	1.13	1.20	30.0	55.1	74.7	40.0	56.6
	B: ρ-family, opt. 1st group	1.13	1.20	28.7	55.1	74.8	39.8	56.5
	C: Optimal non-adaptive	–	1.20	25.6	51.7	76.2	37.5	55.1
	D: Optimal adaptive	–	1.20	24.0	50.9	75.2	36.0	54.0
5	A: ρ-family, equal groups	1.22	1.20	24.0	53.4	73.2	37.7	54.8
	B: ρ-family, opt. 1st group	1.22	1.20	25.2	53.4	73.1	37.8	54.7
	C: Optimal non-adaptive	–	1.20	22.6	49.8	74.3	35.7	53.2
	D: Optimal adaptive	–	1.20	21.6	48.8	73.0	34.2	52.0
6	A: ρ-family, equal groups	1.28	1.20	20.0	52.3	72.2	36.4	53.6
	B: ρ-family, opt. 1st group	1.28	1.20	22.9	52.3	72.0	36.4	53.6
	C: Optimal non-adaptive	–	1.20	20.4	48.6	73.1	34.5	52.1
	D: Optimal adaptive	–	1.20	19.2	47.5	71.8	33.0	50.8

same conditions on Type I error and power.

Proponents of adaptive tests may turn this comment around and point out the opportunity to define adaptive designs attaining this extra efficiency. We believe it will be quite a challenge to find simply defined adaptive procedures with such robustly efficient performance! We have observed the sampling rules of optimal adaptive tests to be qualitatively different from rules based on conditional power commonly used in adaptive designs (see also Posch et al, 2003, p. 961). Some constructions of adaptive tests involve the use of non-sufficient statistics, again distinguishing them from the optimal adaptive designs (see Tsiatis and Mehta, 2003). Thus, some quite new types of adaptive procedure will be needed if this challenge is to be met.

We conclude this section by noting that we have made many further computations using other weighted combinations of ASNs in place of the criterion (1) and in all cases we have found qualitatively similar

Table 4: Properties of tests with Type I error rate $\alpha = 0.025$ and power $1 - \beta = 0.8$ at $\theta = \Delta$ which minimize criterion (1) with $L = 4$. The value of R is fixed at 1.2 unless a lower value minimizes (1).

K	Design	ρ	R	First group	ASN at $\theta = 0$	ASN at $\theta = \Delta$	ASN at $\theta = 4\Delta$	Average ASN
1	Fixed		1.00	100.0	100.0	100.0	100.0	100.0
2	A: ρ-family, equal groups	1.46	1.08	54.0	68.3	83.5	54.0	68.6
	B: ρ-family, opt. 1st group	0.64	1.20	32.6	67.1	91.1	32.6	63.6
	C: Optimal non-adaptive	–	1.18	30.0	64.1	94.3	30.0	62.8
	D: Optimal adaptive	–	1.20	28.8	65.1	93.5	28.8	62.5
3	A: ρ-family, equal groups	1.19	1.16	38.7	59.3	77.5	38.7	58.5
	B: ρ-family, opt. 1st group	0.92	1.20	17.6	61.9	81.4	18.6	53.9
	C: Optimal non-adaptive	–	1.20	18.0	55.7	82.3	19.2	52.4
	D: Optimal adaptive	–	1.20	16.8	55.2	82.1	18.0	51.8
4	A: ρ-family, equal groups	1.13	1.20	30.0	55.1	74.7	30.0	53.3
	B: ρ-family, opt. 1st group	1.09	1.20	15.3	56.7	76.5	17.2	50.1
	C: Optimal non-adaptive	–	1.20	15.1	51.6	77.8	17.1	48.8
	D: Optimal adaptive	–	1.20	12.0	51.0	77.5	14.5	47.7
5	A: ρ-family, equal groups	1.22	1.20	24.0	53.4	73.2	24.1	50.2
	B: ρ-family, opt. 1st group	1.20	1.20	14.2	54.2	74.1	16.6	48.3
	C: Optimal non-adaptive	–	1.20	12.5	49.4	76.1	15.0	46.8
	D: Optimal adaptive	–	1.20	9.6	48.6	75.0	12.9	45.5
6	A: ρ-family, equal groups	1.28	1.20	20.0	52.3	72.2	20.3	48.3
	B: ρ-family, opt. 1st group	1.26	1.20	13.3	52.8	72.7	16.1	47.2
	C: Optimal non-adaptive	–	1.20	11.0	48.2	74.5	13.9	45.5
	D: Optimal adaptive	–	1.20	9.6	46.8	73.1	12.6	44.2

results to those reported here.

4 Examples

(a) Normal data. In a randomized trial of cholesterol lowering drugs, patients are assigned to receive a new compound or an active control. The endpoint is the reduction in cholesterol level from week zero to week six. The parameter of interest $\theta = \mu_A - \mu_B$ is the difference in mean reductions. From historical data, the patient-to-patient standard deviation of cholesterol reductions is taken to be $\sigma = 60$ mg/dl. We test H_0: $\theta = 0$ versus H_1: $\theta > 0$ with Type I error rate $\alpha = 0.025$. We wish to ensure adequate power $1 - \beta = 0.8$ at a minimum clinically relevant effect size of $\theta = \Delta = 15$ mg/dl, but the improvement is expected to be about 60 mg/dl. Hence $L = 4$. From (2), a fixed sample plan would

require $n_f = 2\sigma^2(z_\alpha + z_\beta)^2/\Delta^2 = 252$ subjects per treatment arm, a total sample size of 504.

Suppose we choose $R = 1.2$, allowing an increase of 20% in the maximum sample size to 606 subjects, in order to run a group sequential test with $K = 3$ analyses and the same Type I and Type II error rates. With equally sized groups, the target is 101 subjects per group on each treatment arm and we see from Table 4 that the error spending function parameter is $\rho = 1.19$ and average ASN over $\theta = 0$, Δ and 4Δ is $504 \times 58.5\% = 295$ subjects. If however we choose a first group size of $504 \times 0.176 = 89$ and $\rho = 0.92$, the average ASN is reduced to $504 \times 53.9\% = 272$. The optimal non-adaptive and adaptive tests would lead to average ASNs of 264 and 261, respectively, quite a meager return on the additional complexity of these two procedures.

It may be helpful to relate this example to the problem described by Shun et al. (2001) to which we referred in Section 1, bearing in mind that the two effect sizes in the current example are $\theta = 15$ and 60, rather than the previous $\theta = 5$ and 10. The recommendation of Shun et al. (2001) was to start with a small sample size sufficient to detect the higher effect size, $\theta = 60$ in our example, and then to decide on a second stage sample size having seen this first set of observations. (In view of the higher ratio between the optimistic effect size of 60 and minimal clinically significant effect of 15, it may be advisable to take a somewhat higher initial sample size in this case.) Suppose such a sampling rule and terminal decision rule are defined and these produce a test with overall Type I error rate 0.025 and power 0.8 under $\theta = 15$, then we have a 2-group adaptive design in our class D and Figure 1 and Tables 2 and 4 tell us something about its properties. From Table 2, we see the optimal 2-group adaptive design for criterion (1) with $L = 4$ has an overall maximum sample size $R = 1.26$ times the fixed sample size needed to achieve power 0.8 at $\theta = 15$, i.e., $1.26 \times 504 = 635$ subjects, and this results in an average ASN of $504 \times 62.4\% = 314$. The test constructed following Shun et al's (2001) approach can do no better than this optimal design and so must have an average ASN of at least 314 subjects. Also from Table 2, we see that a 2-group ρ-family test with $R = 1.20$ and first group size 164 does almost as well, achieving average ASN of 321. It is noteworthy that the additional analysis in a 3-group, ρ-family test reduces average ASN to 272 giving a significant advantage over the optimal 2-stage adaptive test.

(b) Survival data. Suppose a one-sided logrank test is to be used to compare the survival distributions of subjects on treatment and control arms. The parameter of interest is $\theta = \log\lambda$ where λ is the relative risk (hazard ratio), which is assumed to be constant. We wish to test H_0: $\theta = 0$ versus H_1: $\theta > 0$ with Type I error rate $\alpha = 0.025$. We specify power $1 - \beta = 0.8$ at a minimum clinically relevant effect size of $\lambda = 1.4$, i.e., at $\theta = \Delta = 0.336$, but the relative risk is expected to be $\lambda = 2$. Hence $L = \log 2/\log 1.4 \approx 2$. A fixed sample plan requires, approximately, a combined total of $4(z_\alpha + z_\beta)^2/\Delta^2 = 278$ deaths (or more generally "events") to be observed in the two arms — see Schoenfeld and Richter (1982) or Jennison and Turnbull (2000, p. 79).

Let us choose an inflation factor $R = 1.2$ and $K = 5$ analyses. Thus, with equal numbers of events between successive analyses, we would need a maximum of 334 events and the target for the study design would be 67 new events at each analysis. From Table 3, we see the error spending function parameter is $\rho = 1.22$ and this leads to an average, over the cases $\theta = 0$, Δ and 2Δ, of 152 observed events. In this case, optimizing the number of events at the first analysis has a negligible effect and there is very little difference between the optimal class A and class B tests. The optimal non-adaptive and adaptive tests offer small improvements with average numbers of events equal to 148 and 145, respectively but once again it appears that the savings are hardly worthwhile to justify the complexity of the class C or D procedures.

5 Discussion

We have posed the problem of how to choose the effect size $\theta = \Delta$ at which to specify the power of a clinical trial when there is disagreement or uncertainty about the likely treatment effect. Our conclusion when there is a choice between a minimal clinically significant effect size and larger effect sizes that investigators hope to see is that *the power requirement should be set at the minimal clinically significant effect*. This decision may need to be moderated if the resulting sample size is prohibitive, bearing in mind that a good sequential design will reduce the observed sample size if the effect size is much larger than the minimal effect size.

In choosing between the available types of experimental design it is the *overall* power at $\theta = \Delta$ that matters. Proschan and Hunsberger (1995) propose "designed extension" procedures which start with an initial sample size appropriate to a detecting a higher effect size and continue with a second stage of sampling if first stage results are not decisive, increasing power to detect lower effect sizes in the process. However, they recognise the importance of overall power as they present power calculated at a series of possible effect sizes in their Table 2. Our simple observation is that, since the overall power curve determines the effectiveness of a procedure in detecting treatment effects of different sizes, one may just as well start by formulating the requirements of a statistical design in terms of its overall power.

Our findings in Section 3 show that, once the power requirement and the effect sizes under which low ASNs are most important have been specified, the ρ-family of error spending tests offers a class of very efficient group sequential designs. If the initial group size is chosen carefully, designs can be obtained which are close to optimal among all non-adaptive group sequential tests. Since they are defined as error spending tests, these procedures are equipped to deal with departures from the planned group sizes, while automatically preserving their Type I error rate and power; see Jennison and Turnbull (2000, Ch. 7) for more details. Moreover, if the sample size needed to provide a given power depends on a nuisance parameter, such as the variance of normal observations, Mehta and Tsiatis (2001) show how the "information monitoring" approach to implementing error spending tests provides a framework for combining group sequential testing with sample size re-estimation based on updated estimates of the nuisance parameter.

Adding the element of adaptivity to a group sequential design, so that future group sizes are chosen in the light of current data, can provide a little extra efficiency, but our numerical results for optimal adaptive tests show that at best this gain is slight. Moreover, we are not aware of any simply defined adaptive tests which come close to achieving such efficiency. Indeed, our studies of some of the specific proposals that have been made (see, for instance, Jennison and Turnbull, 2003) indicate that ostensibly reasonable adaptation rules can lead to very poor efficiency compared with well chosen group sequential designs.

A possible benefit of adaptive designs is their flexibility to adapt to information external to a trial. For example, news of a competing product's adverse side-effects could lead to reassessment of the minimum commercially relevant effect size for a drug currently under study. The flexibility of these tests can also be used to adapt a trial to a change in treatment definition (such as a new dosage or selection of one dose from an initial range of doses), or to the substitution of an alternate endpoint; see, for example, Bauer and Köhne (1994), Bauer and Röhmel (1995), Fisher (1998) and Lehmacher and Wassmer (1999). In another form of adaptation, Wang et al. (2001) use Cui et al's (1999) method to create a group sequential test which can switch adaptively between hypothesis tests of superiority and non-inferiority. Although conventional group sequential tests do not accommodate such unplanned adaptation quite as

naturally, Denne (2001) and Müller and Schäfer (2001, 2004) have shown this can be achieved through consideration of the conditional Type I error rate at an interim analysis.

It is important to distinguish such adaptation from the notion of "adapting to" an internal estimate of the effect size θ from data in the current trial. This is exactly the interim data used in defining the group sequential tests and adaptive tests considered in Section 3 and, by definition, there is no way this information can be used to improve upon the optimal tests reported there.

The example of a clinical trial for treatment of Parkinson's disease described by Müller and Schäfer (2004, Sec. 4) provides an interesting final case for discussion. Response was a score from a patient questionnaire and there was initial uncertainty about the response variance as well as variation of opinion on the minimum clinically relevant effect. The authors propose an initial group sequential design which is modified as new estimates of the response variance are obtained. We are not so convinced by their proposal for design modifications in the light of interim estimates of the effect size. The sample size adaptations are based on attaining a specific conditional power if the true effect is equal to the lower limit of a confidence interval for the effect size. This construction produces yet another rule in our class D — but we have seen that adding an adaptive element offers little scope for improving efficiency over that of a well chosen GST and we have warned that some adaptive schemes can be quite inefficient. More significantly, this adaptation does nothing to resolve the original difference of opinion concerning the minimum clinically relevant effect: interim estimates provide information about the true effect size, not about the range of values to be regarded as clinically significant.

Acknowledgement

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

References

Anderson, K. and Liu, Q. (2004). Optimal adaptive vs optimal group sequential design. Conference on *Adaptive Design for Clinical Trials* Philadelphia, March 4–5, 2004.

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

Bauer, P. (1989). Multistage testing with adaptive designs (with discussion). *Biometrie und Informatik in Medizin und Biologie* **20**, 130–148.

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

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

Cytel Software Corp. (2003). *EaSt v.3: Software for the Design, Simulation and Interim Monitoring of Flexible Clinical Trials.* Cambridge, Massachussets.

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

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

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

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

Food and Drug Administration (1998). E9: Statistical Principles for Clinical Trials. *Federal Register* **63**(179), 49583–49598, (16 September, 1998).

Gould, A. L. and Shih, W. J. (1992). Sample size re-estimation without unblinding for normally distributed outcomes with unknown variance. *Communications in Statistics (A) – Theory and methods* **21**, 2833-2853.

Insightful Corp. (2002). *S+SeqTrial 2.0*. Seattle, Washington.

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

Jennison, C. and Turnbull, B.W. (2003). Mid-course sample size modification in clinical trials based on the observed treatment effect. *Statistics in Medicine* **22**, 971–993.

Li, G., Shih, W. J., Xie, T. and Lu. J. (2002). A sample size adjustment procedure for clinical trials based on conditional power. *Biostatistics* **3**, 277–287.

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

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

Müller, H-H. and Schäfer, H. (2004). A general principle for changing a design any time during the course of a clinical trial. *Statistics in Medicine* **23**, 2497–2508.

Piantadosi, S. (1997) *Clinical Trials: A Methodologic Perspective.* Wiley, New York.

Pocock, S. (1983). *Clinical Trials: A Practical Approach*. Wiley, New York.

Posch, M., Bauer, P. and Brannath, W. (2003). Issues in designing flexible trials. *Statistics in Medicine* **22**, 953–969.

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

Schäfer, H. and Müller, H-H. (2004). Construction of group sequential designs in clinical trials on the basis of detectable treatment differences. *Statistics in Medicine* **23**, 1413–1424.

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

Schoenfeld, D.A. and Richter, J.R. (1982). Nomograms for calculating the number of patients needed for a clinical trial with survival as an endpoint. *Biometrics* **38**, 163–170.

Schwartz, T. A. and Denne, J. S. (2003). Common threads between sample size recalculation and group sequential procedures. *Pharmaceutical Statistics* **2**, 263–271.

Senn, S. (1997). *Statistical Issues in Drug Development.* Wiley, New York.

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

Shih, W.J. (2001). Sample size re-estimation — journey for a decade. *Statistics in Medicine* **20**, 515–518.

Shun, Z., Yuan, W., Brady, W.E. and Hsu, H. (2001). Type I error in sample size reestimations based on observed treatment difference (with commentary). *Statistics in Medicine* **20**, 497–513. Rejoinder. 519–520.

Tsiatis, A.A. and Mehta, C. (2003). On the inefficiency of the adaptive design for monitoring clinical trials. *Biometrika* **90**, 367–378.

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

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