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

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

email: cj@maths.bath.ac.uk

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

19 April 2004

SUMMARY

We consider the construction of efficient group sequential designs where the goal is to minimize the expected sample size averaged 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. All designs satisfy the same pre-specified Type I error and power requirements. We investigate four nested classes of group sequential designs: (A) Equal group sizes with critical boundary values determined by a monomial error spending function (the "ρ-family"); (B) As A but the initial group size can be different; (C) Arbitrary group sizes and arbitrary boundaries specified in advance; (D) Adaptive — as C but future group sizes and critical values are updated depending on the current value of the test statistic. By examining the optimal procedures within each class and their performance, we conclude that there is very little loss in efficiency in using the optimum class B procedure rather than the much more complex designs of classes C and D. We provide tables defining the optimal procedures of classes A and B so that they can be implemented. These tables also quantify the improvements in efficiency, generally slight, that can be obtained by considering the more complex class C and D procedures.

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 (θ) of interest 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), 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). However the lower sample size 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). Cui et al. (1999, Sec. 1) report on a placebo controlled myocardial infarction prevention trial with a sample size of 600 subjects per group, 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" or "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), Posch et al. (2003) among others. What is not so well recognized is that this adaptivity coupled with preservation of the Type I error rate can come at a steep price in inefficiency. This inefficiency comes from two sources — disobeying the sufficiency principle and placing too much confidence in a highly variable interim estimate of the treatment difference. Jennison and Turnbull (2003) demonstrate this inefficiency by examples in which conventional α-level group sequential tests have uniformly higher power and uniformly lower expected sample sizes than adaptive designs. The theoretical inadmissibility of flexible/adaptive tests is described by Tsiatis and Mehta (2003). In addition, results from a trial with a flexible/adaptive modification may be less credible for two reasons. First, otherwise homogeneous subjects' outcomes are weighted differently in the final statistic, depending on their order of entry into the trial. Second, there is liable to be suspicion that a trial has been manipulated when its design is modified at an interim stage based on unmasked primary endpoint results.

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 proposals support the view that a clinical trial should guarantee power at effect sizes θ of clinical or commercial interest. Smaller effect sizes 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 sampling cost needed to detect a particularly small effect size is prohibitive: then, power must be set at the smallest value $\theta = \Delta$ that resources permit. At the planning stage, the expected sample size (or average sample number, "ASN") should be considered over the whole range of θ values. In particular, the ASN should be considered at the null hypothesis and both the clinically/commercially minimum effect size Δ **and** the anticipated or hoped for effect size $\theta = L\Delta$, say, where $L > 1$.

Our objective in this paper is to compare different approaches to meeting a given power condition: this may be by specifying a conventional group sequential design at the outset or by following an adaptive strategy in which a small initial sample size is increased in response to interim results. For comparability, we shall study tests with a specific Type I error rate, α, and power $1-\beta$ at a given effect size $\theta = \Delta$. In the case of adaptive/flexible designs it is the overall power of the complete procedure that we consider, however this is achieved. The criterion for optimality will be minimizing the average ASN over parameter values $\theta = 0$, $\theta = \Delta$ and $\theta = L\Delta$, i.e.,

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

This choice 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 types of conventional group sequential tests, moving from simple to more complex designs. In optimizing over adaptive designs, we consider the extension of group sequential tests to "sequentially planned" designs proposed by Schmitz (1993) which allows future group sizes and decision boundaries to be adapted to previously observed responses in a pre-planned way, following the sufficiency principle. Implementation of Schmitz's designs is complicated and, unlike many other adaptive procedures, they are not intended to be used flexibly. The important point for us is that the class of sequentially planned designs includes adaptive tests, hence the performance of optimal Schmitz designs provides an upper bound on what can be achieved in *any* flexible/adaptive design.

Schäfer and Müller (2004) have addressed the issue of uncertainty about the effect size to be detected taking a somewhat different approach. For a K-group trial, they consider a sequence of K values of θ, $L_1\Delta < \ldots < L_K\Delta$. They show how to construct boundaries for a procedure with Type I error rate α such that, for each $k = 1, \ldots, K$, the probability of stopping to reject H_0 at or before the kth stage 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 $\theta = \Delta$. Minimizing the criterion (1), we obtain tests with low ASN at particular θ values greater than Δ, without giving up too much in ASN at $\theta = \Delta$ or at the null hypothesis $\theta = 0$.

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).

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

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

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

In a K-stage group sequential procedure, analyses are performed at a sequence of looks at the accumulating data. Suppose that at analysis k ($k = 1, \ldots, K$), the cumulative number of observations per treatment arm is n_k, where of course $n_1 < \ldots < n_K$. At each stage 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:

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

Here, $a_K = b_K$ ensures that the test terminates at analysis K. The ratio $R = n_K/n_f$ is termed the *inflation factor*.

There are many choices of pairs $\{a_k, b_k\}$, $k = 1, \ldots, K$, that will satisfy the α and β error probability requirements. One convenient way is through specification of a parametric error spending function — see 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 $f(0) = g(0) = 0$, $f(t) = \alpha$ for $t \geq 1$ and $g(t) = \beta$ for $t \geq 1$. The critical values $\{a_k, b_k\}$ 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 $k = K$. 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:

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

parametrized by the single parameter $\rho > 0$. This choice defines the so-called "ρ-family" of tests. By requiring termination by stage K with $a_K = b_K$, the value of ρ determines the inflation factor $R = n_K/n_f$ and vice versa. For details, see Jennison and Turnbull (2000, Chap. 7.3). For equally spaced looks and selected values of K, α and β, this relationship is illustrated in Table 7.6 of Jennison and Turnbull (2000).

We shall consider 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, number of looks K and inflation factor R. Thus, in each case, the total number of observations can be no more than $R\,n_f$.

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 equally spaced looks 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 looks 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 reduce the ASN in this case, reducing the criterion (1). Optimization is over the choice of n_1, after which ρ is determined by R. The numerical results in Tables 1 to 10 will show that further optimization over later group sizes leads only to minor improvements.

C. **Optimal non-adaptive GSTs.** Complete freedom is allowed to choose critical values $\{a_k, b_k\}$ **and** cumulative sample sizes $\{n_1, \ldots, n_K\}$ to optimize the criterion (1), subject to $n_K = R\,n_f$. Note that the choice of n_k and $\{a_k, b_k\}$, $k = 1, \ldots, K$, is set at the beginning and this cannot be updated as observations accrue.

D. **Schmitz's 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 $\{n_k\}$ 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 Catalog of tests and their implementation

Tables 1 to 10 give details of the procedures for the twenty combinations $\alpha = 0.025$, power $1 - \beta = 0.8$, 0.9, inflation factor $R = 1.05$, 1.1, 1.2, 1.3, 1.4, and anticipated effect size ratios $L = 2$, 4. Each table consists of two parts: the procedures in part (a) guarantee power

$1 - \beta = 0.8$ at $\theta = \Delta$, whereas those in part (b) are for $1 - \beta = 0.9$. Thus there are twenty sub-tables. All sample sizes in the tables are expressed as a percentage of that of the fixed sample test.

The first column of each table shows the maximum number of groups K, taking values 1 to 6. The fixed sample test ($K = 1$) serves as a benchmark for the others. There are two rows for each value of $K \geq 2$. The first, A, is the unique ρ-family error spending test with equal group sizes and inflation factor R. The second is for the optimal test in class B which allows freedom in choosing the first group size. The value of n_1 in column 3 is simply $100R/K$ for case A, while for B it is the optimal first group size. For each of these tests, specifying n_1, K and R determines ρ, the parameter of the error spending function, and this is shown in the fourth column.

The next three columns give the ASNs of each test at $\theta = 0$, Δ and $L\Delta$. The column headed 'Avg' is the average of these three ASNs, our criterion (1). The last two columns show the minimum average ASN over tests in class C, all non-adaptive GSTs, and over class D, all adaptive GSTs; these values are repeated in the pair of rows for a given K. These optimal values show the greatest improvements that are possible over the simpler error spending tests, A and B.

The reductions in average ASN as R increases from 1.05 to 1.4 represent the return on a larger maximum sample size. In fact, in classes A, B and C where n_K is set equal to $R n_f$, results improve initially as R increases but then start to rise again. In the examples of optimal GSTs with equal group sizes reported by Eales and Jennison (1992) and Barber and Jennison (2002), the overall minimum is typically between $R = 1.2$ and 1.4, depending on the number of groups, K, and the weighted combination of ASNs being minimized. This behaviour is evident in some of our examples for $K = 2$ and 3 where the test in class A does better with $R = 1.05$ than $R = 1.4$

The values of n_1 in the optimal test in class B are much lower for $L = 4$ than for $L = 2$. Thus, optimism about the size of treatment effect leads to a more aggressive strategy, in the hope of a "home run" finding under $\theta = L\Delta$. Despite the focus when $L = 4$ on a very low ASN under $\theta = 4\Delta$, ASNs at $\theta = 0$ and Δ are quite similar to those for the case $L = 2$.

For $K = 2$, the optimal tests in classes C and D show only a slight advantage over the simpler tests in class B. With only two analyses, after the first group size has been chosen optimally, there is little extra flexibility to make further improvements. For $K \geq 3$, the optimised test in class C improves on that in B by up to 2 or 3% of the fixed sample size and the optimal test in class D improves by a further 1 or 2%. As K increases, the situation eventually approaches continuous monitoring, where adaptivity offers no advantage.

However, the most remarkable lesson to come from these tables is that a member of the fairly simple proposed class of procedures, B, performs almost as well as the best possible non-adaptive and adaptive tests. Barber and Jennison (2002) found ρ-family tests to have good all round performance among GSTs with equal group sizes: adding the option of a smaller first group size gives good results under more extreme alternatives. We can also see there is little to be gained by extending the class B to allow optimization of later group sizes, since the optimal average ASN would have to lie between the optimal values for class B and class C.

Studying the last two columns of each table shows that, in general, adaptive tests offer only minor improvements over non-adaptive tests — and these results are for precisely defined optimal adaptive tests. Any deviation from optimality, such as the form of the adaptive sample size rule or use of a non-sufficient statistic, can lead to worse efficiency than the simple, conventional GSTs A or B. Indeed, as shown in the example of Jennison and Turnbull (2003), an ostensibly reasonable adaptation rule can lead to disastrous inefficiency and lengthening of the trial.

In further computations, not shown here, we have found qualitatively results using other weighted combinations of ASNs in place of the criterion (1).

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 $\alpha = 0.025$. We wish to ensure adequate power $1 - \beta = 0.9$ 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 $n_f = 2\sigma^2(z_\alpha + z_\beta)^2/\Delta^2 = 336$ subjects per treatment arm, a total sample size of 672. Suppose we choose $R = 1.05$, allowing an increase of 5% in the maximum sample size to 704 subjects, in order to run a group sequential test with $K = 4$ looks and the same Type I and Type II error rates. With equally spaced looks, the target is 176 subjects per group and we see from Table 3(b) that the error spending function parameter is $\rho = 2.74$ and average ASN over $\theta = 0$, Δ and 4Δ is $672 \times 54.5\% = 366$ subjects. If however we choose a first group size of $672 \times 0.153 = 103$ and $\rho = 2.60$, the average ASN is reduced to $672 \times 52.2\% = 351$. The optimal non-adaptive and adaptive tests would lead to average ASNs of 341 and 331, respectively, quite a meager return on the additional complexity of these two procedures.

(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 $\alpha = 0.025$. 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 $R = 1.40$ and $K = 3$ looks. Thus, with equal numbers of events between successive analyses, we would need a maximum of 390 events and the target for the study design would be 130 new events at each analysis. From Table 2(a), we see the error spending function parameter is $\rho = 0.50$ and this leads to an average, over the cases $\theta = 0$, Δ and 2Δ, of 172 observed events. If however we conduct the first analysis after $278 \times 0.332 = 92$ events and take $\rho = 0.52$, the average number of events is reduced to 165. The optimal non-adaptive and adaptive tests would lead to average numbers of events equal to 161 and 158, respectively. Again it appears that the savings are hardly worthwhile to justify the complexity of the class C or D procedures.

5 Conclusion

We do not dispute that adaptive/flexible designs can sometimes play a valuable role, especially in responding to news 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.

However there is no substitute for adequate planning before the trial begins. The "wait-and-see" or "start small, then ask for more" approach is a seductive one. Yet, if sample size readjustment is made based on the observed treatment effect (internal information), a large penalty in efficiency may be paid in order to maintain the Type I error rate. If those planning a study disagree over the likely effect size, it is important to distinguish between the smallest effect size at which high power is to be required and more optimistic effect sizes under which very early stopping is desired. When, after careful discussion, the desired power curve for a study is agreed, a conventional GST attaining this power should be chosen for the protocol. If there is hope that the effect size may be much higher than the value at which power has been specified, we recommend using one of the ρ-family designs with an early first analysis, as described in Section 2 and assessed in Section 3. If, because of unforeseen external circumstances, it becomes essential that the design be changed in mid-course, this can still be done while preserving Type I error using, for example, the adaptive methodology of Müller and Schäfer (2001).

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.

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).

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* **23**, 971–993.

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.

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.

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.

Table 1: Optimal tests and efficiencies for four classes of procedures ($L=2$, $R=1.05$)

				ASN for ρ-family test				Average ASN	
K	Design	First group	ρ	$\theta=0$	$\theta=\Delta$	$\theta=L\Delta$	Avg	[C] Optimal non-adv	[D] Optimal adaptive
(a) *Power* 0.80 *at* $\theta=\Delta$									
1	Fixed	100.0	–	100.0	100.0	100.0	100.0	100.0	100.0
2	[A] Equal gps	52.5	1.87	69.7	84.7	55.4	70.0	69.2	69.2
2	[B] Opt. 1st gp	46.3	1.75	69.8	86.3	52.0	69.4	69.2	69.2
3	[A] Equal gps	35.0	2.38	65.6	81.0	47.3	64.6	62.8	62.0
3	[B] Opt. 1st gp	38.2	2.42	65.3	80.6	47.6	64.5	62.8	62.0
4	[A] Equal gps	26.2	2.69	63.7	79.0	45.5	62.7	60.0	58.8
4	[B] Opt. 1st gp	33.4	2.78	63.3	78.4	45.2	62.3	60.0	58.8
5	[A] Equal gps	21.0	2.90	62.7	77.8	44.1	61.5	58.4	57.0
5	[B] Opt. 1st gp	30.0	3.00	62.3	77.3	43.7	61.1	58.4	57.0
6	[A] Equal gps	17.5	3.07	62.0	77.0	43.2	60.8	57.4	56.0
6	[B] Opt. 1st gp	27.6	3.17	61.7	76.6	42.8	60.4	57.4	56.0
(b) *Power* 0.90 *at* $\theta=\Delta$									
1	Fixed	100.0	–	100.0	100.0	100.0	100.0	100.0	100.0
2	[A] Equal gps	52.5	1.89	69.8	79.9	53.2	67.6	66.3	66.3
2	[B] Opt. 1st gp	43.3	1.69	70.8	82.5	45.9	66.4	66.3	66.3
3	[A] Equal gps	35.0	2.42	65.6	75.8	41.3	60.9	59.3	58.3
3	[B] Opt. 1st gp	35.2	2.42	65.6	75.8	41.4	60.9	59.3	58.3
4	[A] Equal gps	26.2	2.74	63.7	73.7	39.3	58.9	56.3	54.9
4	[B] Opt. 1st gp	31.0	2.81	63.3	73.2	39.4	58.6	56.3	54.9
5	[A] Equal gps	21.0	2.98	62.6	72.4	38.3	57.8	54.5	53.0
5	[B] Opt. 1st gp	28.0	3.06	62.2	72.0	38.0	57.4	54.5	53.0
6	[A] Equal gps	17.5	3.15	61.9	71.6	37.4	57.0	53.5	51.9
6	[B] Opt. 1st gp	25.7	3.24	61.6	71.3	37.1	56.6	53.5	51.9

Table 2: Optimal tests and efficiencies for four classes of procedures ($L = 2$, $R = 1.1$)

K	Design	First group	ρ	ASN for ρ-family test $\theta=0$	$\theta=\Delta$	$\theta=L\Delta$	Avg	Average ASN [C] Optimal non-adv	[D] Optimal adaptive
(a) *Power* 0.80 *at* $\theta = \Delta$									
1	Fixed	100.0	–	100.0	100.0	100.0	100.0	100.0	100.0
2	[A] Equal gps	55.0	1.27	67.9	83.1	56.8	69.3	67.3	67.3
2	[B] Opt. 1st gp	44.2	1.17	66.8	85.6	50.1	67.5	67.3	67.3
3	[A] Equal gps	36.7	1.64	61.5	78.7	45.2	61.8	60.4	59.6
3	[B] Opt. 1st gp	35.9	1.63	61.5	78.8	45.0	61.8	60.4	59.6
4	[A] Equal gps	27.5	1.84	59.2	76.5	42.1	59.3	57.3	56.2
4	[B] Opt. 1st gp	30.8	1.86	59.0	76.2	42.2	59.1	57.3	56.2
5	[A] Equal gps	22.0	1.98	57.9	75.2	40.6	57.9	55.7	54.3
5	[B] Opt. 1st gp	27.4	2.01	57.7	74.8	40.5	57.7	55.7	54.3
6	[A] Equal gps	18.3	2.08	57.1	74.3	39.6	57.0	54.6	53.2
6	[B] Opt. 1st gp	24.8	2.11	56.9	73.9	39.4	56.7	54.6	53.2
(b) *Power* 0.90 *at* $\theta = \Delta$									
1	Fixed	100.0	–	100.0	100.0	100.0	100.0	100.0	100.0
2	[A] Equal gps	55.0	1.27	68.2	78.2	55.3	67.3	64.6	64.5
2	[B] Opt. 1st gp	41.6	1.12	68.2	81.6	44.1	64.6	64.6	64.5
3	[A] Equal gps	36.7	1.64	61.9	73.2	40.3	58.5	57.0	56.1
3	[B] Opt. 1st gp	33.1	1.61	62.2	73.8	38.8	58.3	57.0	56.1
4	[A] Equal gps	27.5	1.86	59.6	70.9	36.2	55.6	53.7	52.5
4	[B] Opt. 1st gp	28.6	1.87	59.5	70.8	36.4	55.6	53.7	52.5
5	[A] Equal gps	22.0	2.00	58.2	69.5	34.7	54.1	51.9	50.5
5	[B] Opt. 1st gp	25.5	2.03	58.0	69.2	34.8	54.0	51.9	50.5
6	[A] Equal gps	18.3	2.11	57.4	68.6	33.7	53.2	50.8	49.3
6	[B] Opt. 1st gp	23.3	2.14	57.1	68.3	33.7	53.0	50.8	49.3

Table 3: Optimal tests and efficiencies for four classes of procedures ($L = 2$, $R = 1.2$)

K	Design	First group	ρ	ASN for ρ-family test $\theta=0$	$\theta=\Delta$	$\theta=L\Delta$	Avg	Average ASN [C] Optimal non-adv	[D] Optimal adaptive
(a) *Power* 0.80 *at* $\theta = \Delta$									
1	Fixed	100.0	–	100.0	100.0	100.0	100.0	100.0	100.0
2	[A] Equal gps	60.0	0.74	68.7	82.8	60.9	70.8	66.4	66.1
2	[B] Opt. 1st gp	43.0	0.69	64.6	86.2	48.9	66.6	66.4	66.1
3	[A] Equal gps	40.0	1.00	58.6	77.2	45.1	60.3	58.5	57.8
3	[B] Opt. 1st gp	33.8	0.99	58.1	78.0	43.0	59.7	58.5	57.8
4	[A] Equal gps	30.0	1.13	55.1	74.7	40.0	56.6	55.1	54.0
4	[B] Opt. 1st gp	28.7	1.13	55.1	74.8	39.8	56.5	55.1	54.0
5	[A] Equal gps	24.0	1.22	53.4	73.2	37.7	54.8	53.2	52.0
5	[B] Opt. 1st gp	25.2	1.22	53.4	73.1	37.8	54.7	53.2	52.0
6	[A] Equal gps	20.0	1.28	52.3	72.2	36.4	53.6	52.1	50.8
6	[B] Opt. 1st gp	22.9	1.28	52.3	72.0	36.4	53.6	52.1	50.8
(b) *Power* 0.90 *at* $\theta = \Delta$									
1	Fixed	100.0	–	100.0	100.0	100.0	100.0	100.0	100.0
2	[A] Equal gps	60.0	0.72	69.2	78.3	60.1	69.2	63.9	63.6
2	[B] Opt. 1st gp	40.8	0.65	66.5	82.0	43.1	63.9	63.9	63.6
3	[A] Equal gps	40.0	0.98	59.4	71.5	41.7	57.6	55.3	54.5
3	[B] Opt. 1st gp	31.4	0.96	59.4	72.8	36.8	56.3	55.3	54.5
4	[A] Equal gps	30.0	1.12	56.0	68.8	34.9	53.2	51.7	50.5
4	[B] Opt. 1st gp	26.8	1.11	56.1	69.2	33.9	53.1	51.7	50.5
5	[A] Equal gps	24.0	1.21	54.3	67.2	32.1	51.2	49.7	48.4
5	[B] Opt. 1st gp	23.7	1.21	54.3	67.2	32.0	51.2	49.7	48.4
6	[A] Equal gps	20.0	1.27	53.2	66.2	30.6	50.0	48.5	47.1
6	[B] Opt. 1st gp	21.3	1.27	53.2	66.1	30.7	50.0	48.5	47.1

Table 4: Optimal tests and efficiencies for four classes of procedures ($L = 2$, $R = 1.3$)

K	Design	First group	ρ	ASN for ρ-family test $\theta = 0$	$\theta = \Delta$	$\theta = L\Delta$	Avg	Average ASN [C] Optimal non-adv	[D] Optimal adaptive
(a) *Power* 0.80 *at* $\theta = \Delta$									
1	Fixed	100.0	–	100.0	100.0	100.0	100.0	100.0	100.0
2	[A] Equal gps	65.0	0.48	71.4	84.0	65.5	73.6	66.7	65.9
2	[B] Opt. 1st gp	43.0	0.48	64.1	87.7	48.7	66.9	66.7	65.9
3	[A] Equal gps	43.3	0.69	58.3	77.1	46.7	60.7	57.9	57.1
3	[B] Opt. 1st gp	33.2	0.69	56.7	78.4	42.3	59.1	57.9	57.1
4	[A] Equal gps	32.5	0.80	53.6	74.3	39.9	55.9	54.2	53.2
4	[B] Opt. 1st gp	28.0	0.80	53.3	74.8	38.8	55.6	54.2	53.2
5	[A] Equal gps	26.0	0.86	51.4	72.7	36.8	53.6	52.2	51.0
5	[B] Opt. 1st gp	24.6	0.86	51.3	72.8	36.6	53.6	52.2	51.0
6	[A] Equal gps	21.7	0.91	50.0	71.6	35.2	52.3	50.9	49.6
6	[B] Opt. 1st gp	22.0	0.90	50.1	71.6	35.2	52.3	50.9	49.6
(b) *Power* 0.90 *at* $\theta = \Delta$									
1	Fixed	100.0	–	100.0	100.0	100.0	100.0	100.0	100.0
2	[A] Equal gps	65.0	0.45	71.8	79.8	65.1	72.2	64.3	63.5
2	[B] Opt. 1st gp	41.0	0.44	66.4	83.4	43.1	64.3	64.3	63.5
3	[A] Equal gps	43.3	0.67	59.2	71.5	44.3	58.3	54.8	54.0
3	[B] Opt. 1st gp	30.9	0.66	58.4	73.2	36.0	55.9	54.8	54.0
4	[A] Equal gps	32.5	0.77	54.8	68.3	35.6	52.9	50.9	49.8
4	[B] Opt. 1st gp	25.4	0.77	54.8	69.2	32.7	52.2	50.9	49.8
5	[A] Equal gps	26.0	0.84	52.6	66.6	31.7	50.3	48.8	47.5
5	[B] Opt. 1st gp	22.9	0.84	52.7	66.9	30.9	50.1	48.8	47.5
6	[A] Equal gps	21.7	0.88	51.4	65.5	29.6	48.8	47.5	46.2
6	[B] Opt. 1st gp	20.6	0.88	51.4	65.6	29.4	48.8	47.5	46.2

Table 5: Optimal tests and efficiencies for four classes of procedures ($L = 2$, $R = 1.4$)

K	Design	First group	ρ	ASN for ρ-family test $\theta = 0$	$\theta = \Delta$	$\theta = L\Delta$	Avg	Average ASN [C] Optimal non-adv	[D] Optimal adaptive
(a) *Power* 0.80 *at* $\theta = \Delta$									
1	Fixed	100.0	–	100.0	100.0	100.0	100.0	100.0	100.0
2	[A] Equal gps	70.0	0.32	74.8	85.7	70.3	76.9	67.5	65.9
2	[B] Opt. 1st gp	43.5	0.35	64.2	89.6	49.0	67.6	67.5	65.9
3	[A] Equal gps	46.7	0.50	59.1	77.6	49.0	61.9	57.8	56.9
3	[B] Opt. 1st gp	33.2	0.52	56.1	79.3	42.1	59.2	57.8	56.9
4	[A] Equal gps	35.0	0.59	53.3	74.5	40.6	56.1	53.8	52.8
4	[B] Opt. 1st gp	27.6	0.60	52.3	75.3	38.4	55.4	53.8	52.8
5	[A] Equal gps	28.0	0.65	50.5	72.8	36.8	53.4	51.7	50.5
5	[B] Opt. 1st gp	24.2	0.65	50.2	73.2	36.1	53.2	51.7	50.5
6	[A] Equal gps	23.3	0.68	48.9	71.7	34.7	51.8	50.3	49.0
6	[B] Opt. 1st gp	21.7	0.69	48.8	71.9	34.5	51.8	50.3	49.0
(b) *Power* 0.90 *at* $\theta = \Delta$									
1	Fixed	100.0	–	100.0	100.0	100.0	100.0	100.0	100.0
2	[A] Equal gps	70.0	0.29	75.1	82.0	70.0	75.7	65.2	63.5
2	[B] Opt. 1st gp	41.8	0.31	66.8	85.1	43.6	65.2	65.2	63.5
3	[A] Equal gps	46.7	0.48	60.1	72.1	47.2	59.8	54.8	53.8
3	[B] Opt. 1st gp	31.0	0.48	58.2	74.0	35.8	56.0	54.8	53.8
4	[A] Equal gps	35.0	0.56	54.6	68.5	37.1	53.4	50.6	49.5
4	[B] Opt. 1st gp	26.0	0.57	54.2	69.5	32.5	52.0	50.6	49.5
5	[A] Equal gps	28.0	0.62	52.0	66.6	32.2	50.3	48.4	47.1
5	[B] Opt. 1st gp	22.7	0.62	52.0	67.2	30.3	49.8	48.4	47.1
6	[A] Equal gps	23.3	0.65	50.5	65.5	29.5	48.5	47.0	45.7
6	[B] Opt. 1st gp	20.4	0.66	50.6	65.8	28.8	48.4	47.0	45.7

Table 6: Optimal tests and efficiencies for four classes of procedures ($L = 4$, $R = 1.05$)

K	Design	First group	ρ	ASN for ρ-family test $\theta = 0$	$\theta = \Delta$	$\theta = L\Delta$	Avg	Average ASN [C] Optimal non-adv	[D] Optimal adaptive
(a) *Power* 0.80 *at* $\theta = \Delta$									
1	Fixed	100.0	–	100.0	100.0	100.0	100.0	100.0	100.0
2	[A] Equal gps	52.5	1.87	69.7	84.7	52.5	69.0	64.9	64.9
2	[B] Opt. 1st gp	30.9	1.44	75.4	92.0	30.9	66.1	64.9	64.9
3	[A] Equal gps	35.0	2.38	65.6	81.0	35.0	60.5	56.2	55.3
3	[B] Opt. 1st gp	18.9	2.14	68.5	83.1	21.0	57.6	56.2	55.3
4	[A] Equal gps	26.2	2.69	63.7	79.0	26.4	56.4	53.1	51.5
4	[B] Opt. 1st gp	18.1	2.58	64.5	79.6	20.9	55.0	53.1	51.5
5	[A] Equal gps	21.0	2.90	62.7	77.8	22.1	54.2	51.3	49.6
5	[B] Opt. 1st gp	17.4	2.86	62.9	78.0	20.6	53.8	51.3	49.6
6	[A] Equal gps	17.5	3.07	62.0	77.0	20.4	53.2	50.1	48.4
6	[B] Opt. 1st gp	16.7	3.06	62.1	77.1	20.3	53.1	50.1	48.4
(b) *Power* 0.90 *at* $\theta = \Delta$									
1	Fixed	100.0	–	100.0	100.0	100.0	100.0	100.0	100.0
2	[A] Equal gps	52.5	1.89	69.8	79.9	52.5	67.4	64.9	64.9
2	[B] Opt. 1st gp	36.4	1.55	73.4	85.5	36.4	65.1	64.9	64.9
3	[A] Equal gps	35.0	2.42	65.6	75.8	35.0	58.8	54.4	53.5
3	[B] Opt. 1st gp	15.5	2.11	69.3	78.8	17.2	55.1	54.4	53.5
4	[A] Equal gps	26.2	2.74	63.7	73.7	26.3	54.5	50.8	49.3
4	[B] Opt. 1st gp	15.3	2.60	64.7	74.6	17.4	52.2	50.8	49.3
5	[A] Equal gps	21.0	2.98	62.6	72.4	21.2	52.1	49.0	47.2
5	[B] Opt. 1st gp	14.9	2.91	62.9	72.8	17.3	51.0	49.0	47.2
6	[A] Equal gps	17.5	3.15	61.9	71.6	18.3	50.6	47.8	46.0
6	[B] Opt. 1st gp	14.6	3.12	62.0	71.8	17.1	50.3	47.8	46.0

Table 7: Optimal tests and efficiencies for four classes of procedures ($L = 4$, $R = 1.1$)

				ASN for ρ-family test				Average ASN	
K	Design	First group	ρ	$\theta = 0$	$\theta = \Delta$	$\theta = L\Delta$	Avg	[C] Optimal non-adv	[D] Optimal adaptive
(a) *Power* 0.80 *at* $\theta = \Delta$									
1	Fixed	100.0	–	100.0	100.0	100.0	100.0	100.0	100.0
2	[A] Equal gps	55.0	1.27	67.9	83.1	55.0	68.7	63.4	63.3
2	[B] Opt. 1st gp	32.0	1.03	70.5	90.6	32.0	64.4	63.4	63.3
3	[A] Equal gps	36.7	1.64	61.5	78.7	36.7	58.9	54.1	53.3
3	[B] Opt. 1st gp	18.0	1.47	65.1	81.7	19.5	55.4	54.1	53.3
4	[A] Equal gps	27.5	1.84	59.2	76.5	27.5	54.4	50.8	49.4
4	[B] Opt. 1st gp	16.4	1.76	60.4	77.6	18.9	52.3	50.8	49.4
5	[A] Equal gps	22.0	1.98	57.9	75.2	22.3	51.8	48.9	47.4
5	[B] Opt. 1st gp	15.7	1.94	58.4	75.6	18.5	50.8	48.9	47.4
6	[A] Equal gps	18.3	2.08	57.1	74.3	19.4	50.3	47.7	46.1
6	[B] Opt. 1st gp	14.8	2.06	57.3	74.5	18.0	49.9	47.7	46.1
(b) *Power* 0.90 *at* $\theta = \Delta$									
1	Fixed	100.0	–	100.0	100.0	100.0	100.0	100.0	100.0
2	[A] Equal gps	55.0	1.27	68.2	78.2	55.0	67.1	63.3	63.3
2	[B] Opt. 1st gp	36.0	1.05	70.1	84.3	36.0	63.5	63.3	63.3
3	[A] Equal gps	36.7	1.64	61.9	73.2	36.7	57.3	52.6	51.8
3	[B] Opt. 1st gp	14.6	1.43	66.9	77.4	15.8	53.4	52.6	51.8
4	[A] Equal gps	27.5	1.86	59.6	70.9	27.5	52.7	48.7	47.4
4	[B] Opt. 1st gp	13.6	1.75	61.3	72.4	15.5	49.7	48.7	47.4
5	[A] Equal gps	22.0	2.00	58.2	69.5	22.0	49.9	46.8	45.2
5	[B] Opt. 1st gp	13.2	1.95	58.9	70.1	15.3	48.1	46.8	45.2
6	[A] Equal gps	18.3	2.11	57.4	68.6	18.5	48.2	45.6	43.9
6	[B] Opt. 1st gp	12.7	2.08	57.7	68.9	15.0	47.2	45.6	43.9

Table 8: Optimal tests and efficiencies for four classes of procedures ($L = 4$, $R = 1.2$)

				ASN for ρ-family test				Average ASN	
K	Design	First group	ρ	$\theta=0$	$\theta=\Delta$	$\theta=L\Delta$	Avg	[C] Optimal non-adv	[D] Optimal adaptive
(a) *Power* 0.80 *at* $\theta = \Delta$									
1	Fixed	100.0	–	100.0	100.0	100.0	100.0	100.0	100.0
2	[A] Equal gps	60.0	0.74	68.7	82.8	60.0	70.5	62.8	62.5
2	[B] Opt. 1st gp	32.6	0.64	67.1	91.1	32.6	63.6	62.8	62.5
3	[A] Equal gps	40.0	1.00	58.6	77.2	40.0	58.6	52.4	51.8
3	[B] Opt. 1st gp	17.6	0.92	61.9	81.4	18.6	53.9	52.4	51.8
4	[A] Equal gps	30.0	1.13	55.1	74.7	30.0	53.3	48.8	47.7
4	[B] Opt. 1st gp	15.3	1.09	56.7	76.5	17.2	50.1	48.8	47.7
5	[A] Equal gps	24.0	1.22	53.4	73.2	24.1	50.2	46.8	45.5
5	[B] Opt. 1st gp	14.2	1.20	54.2	74.1	16.6	48.3	46.8	45.5
6	[A] Equal gps	20.0	1.28	52.3	72.2	20.3	48.3	45.5	44.2
6	[B] Opt. 1st gp	13.3	1.26	52.8	72.7	16.1	47.2	45.5	44.2
(b) *Power* 0.90 *at* $\theta = \Delta$									
1	Fixed	100.0	–	100.0	100.0	100.0	100.0	100.0	100.0
2	[A] Equal gps	60.0	0.72	69.2	78.3	60.0	69.1	62.8	62.4
2	[B] Opt. 1st gp	36.5	0.63	67.9	84.3	36.5	62.9	62.8	62.4
3	[A] Equal gps	40.0	0.98	59.4	71.5	40.0	57.0	51.3	50.6
3	[B] Opt. 1st gp	14.6	0.88	64.8	77.3	15.2	52.4	51.3	50.6
4	[A] Equal gps	30.0	1.12	56.0	68.8	30.0	51.6	47.1	45.9
4	[B] Opt. 1st gp	12.4	1.06	58.6	71.2	13.9	47.9	47.1	45.9
5	[A] Equal gps	24.0	1.21	54.3	67.2	24.0	48.5	45.0	43.6
5	[B] Opt. 1st gp	11.7	1.18	55.6	68.4	13.5	45.8	45.0	43.6
6	[A] Equal gps	20.0	1.27	53.2	66.2	20.0	46.5	43.7	42.1
6	[B] Opt. 1st gp	11.1	1.25	53.9	66.9	13.2	44.7	43.7	42.1

Table 9: Optimal tests and efficiencies for four classes of procedures ($L = 4$, $R = 1.3$)

				ASN for ρ-family test				Average ASN	
K	Design	First group	ρ	$\theta=0$	$\theta=\Delta$	$\theta=L\Delta$	Avg	[C] Optimal non-adv	[D] Optimal adaptive
(a) *Power* 0.80 *at* $\theta = \Delta$									
1	Fixed	100.0	–	100.0	100.0	100.0	100.0	100.0	100.0
2	[A] Equal gps	65.0	0.48	71.4	84.0	65.0	73.5	63.4	62.4
2	[B] Opt. 1st gp	33.8	0.45	65.8	92.5	33.8	64.0	63.4	62.4
3	[A] Equal gps	43.3	0.69	58.3	77.1	43.3	59.6	51.9	51.3
3	[B] Opt. 1st gp	17.8	0.66	60.3	82.2	18.5	53.7	51.9	51.3
4	[A] Equal gps	32.5	0.80	53.6	74.3	32.5	53.5	48.0	47.0
4	[B] Opt. 1st gp	14.9	0.78	55.0	76.7	16.5	49.4	48.0	47.0
5	[A] Equal gps	26.0	0.86	51.4	72.7	26.0	50.0	46.0	44.7
5	[B] Opt. 1st gp	13.5	0.85	52.3	74.0	15.7	47.3	46.0	44.7
6	[A] Equal gps	21.7	0.91	50.0	71.6	21.8	47.8	44.6	43.3
6	[B] Opt. 1st gp	12.7	0.90	50.6	72.4	15.2	46.1	44.6	43.3
(b) *Power* 0.90 *at* $\theta = \Delta$									
1	Fixed	100.0	–	100.0	100.0	100.0	100.0	100.0	100.0
2	[A] Equal gps	65.0	0.45	71.8	79.8	65.0	72.2	63.4	62.4
2	[B] Opt. 1st gp	37.4	0.43	67.4	85.5	37.4	63.4	63.4	62.4
3	[A] Equal gps	43.3	0.67	59.2	71.5	43.3	58.0	51.1	50.4
3	[B] Opt. 1st gp	15.8	0.62	63.7	78.0	16.1	52.6	51.1	50.4
4	[A] Equal gps	32.5	0.77	54.8	68.3	32.5	51.9	46.5	45.4
4	[B] Opt. 1st gp	12.3	0.75	57.6	71.5	13.4	47.5	46.5	45.4
5	[A] Equal gps	26.0	0.84	52.6	66.6	26.0	48.4	44.2	42.9
5	[B] Opt. 1st gp	11.2	0.82	54.3	68.4	12.7	45.1	44.2	42.9
6	[A] Equal gps	21.7	0.88	51.4	65.5	21.7	46.2	42.9	41.4
6	[B] Opt. 1st gp	10.6	0.87	52.4	66.6	12.4	43.8	42.9	41.4

Table 10: Optimal tests and efficiencies for four classes of procedures ($L = 4$, $R = 1.4$)

				ASN for ρ-family test				Average ASN	
K	Design	First group	ρ	$\theta = 0$	$\theta = \Delta$	$\theta = L\Delta$	Avg	[C] Optimal non-adv	[D] Optimal adaptive
(a) *Power* 0.80 *at* $\theta = \Delta$									
1	Fixed	100.0	–	100.0	100.0	100.0	100.0	100.0	100.0
2	[A] Equal gps	70.0	0.32	74.8	85.7	70.0	76.8	64.4	62.4
2	[B] Opt. 1st gp	34.7	0.34	65.5	94.5	34.7	64.9	64.4	62.4
3	[A] Equal gps	46.7	0.50	59.1	77.6	46.7	61.1	51.9	51.2
3	[B] Opt. 1st gp	18.0	0.50	59.6	83.5	18.5	53.9	51.9	51.2
4	[A] Equal gps	35.0	0.59	53.3	74.5	35.0	54.3	47.7	46.7
4	[B] Opt. 1st gp	14.8	0.60	54.2	77.5	16.2	49.3	47.7	46.7
5	[A] Equal gps	28.0	0.65	50.5	72.8	28.0	50.4	45.5	44.3
5	[B] Opt. 1st gp	13.4	0.65	51.2	74.6	15.3	47.0	45.5	44.3
6	[A] Equal gps	23.3	0.68	48.9	71.7	23.4	48.0	44.1	42.8
6	[B] Opt. 1st gp	12.4	0.69	49.5	72.9	14.7	45.7	44.1	42.8
(b) *Power* 0.90 *at* $\theta = \Delta$									
1	Fixed	100.0	–	100.0	100.0	100.0	100.0	100.0	100.0
2	[A] Equal gps	70.0	0.29	75.1	82.0	70.0	75.7	64.4	62.4
2	[B] Opt. 1st gp	38.6	0.31	67.6	87.0	38.6	64.4	64.4	62.4
3	[A] Equal gps	46.7	0.48	60.1	72.1	46.7	59.6	51.3	50.3
3	[B] Opt. 1st gp	17.8	0.47	62.8	78.7	17.8	53.1	51.3	50.3
4	[A] Equal gps	35.0	0.56	54.6	68.5	35.0	52.7	46.3	45.2
4	[B] Opt. 1st gp	12.2	0.56	57.4	72.5	13.1	47.7	46.3	45.2
5	[A] Equal gps	28.0	0.62	52.0	66.6	28.0	48.9	43.9	42.6
5	[B] Opt. 1st gp	11.1	0.62	53.8	69.0	12.4	45.0	43.9	42.6
6	[A] Equal gps	23.3	0.65	50.5	65.5	23.3	46.5	42.5	41.0
6	[B] Opt. 1st gp	10.2	0.65	51.8	67.0	11.9	43.6	42.5	41.0