Comparing Efficiency for Adaptive and Non-Adaptive Group Sequential Designs

Turnbull, Bruce W
Cornell University, Department of Operations Research and Information Engineering
227 Rhodes Hall
Ithaca New York 14853, USA
E-mail: bwt2@cornell.edu

Jennison, Chris
University of Bath
Bath BA2 7AY, UK
E-mail: cj@maths.bath.ac.uk

Ordinarily, in a clinical trial one specifies at the outset: the patient population; the treatments; the randomized allocation rule; the primary endpoint; the hypothesis to be tested; the sample size, or equivalently, the power at a specific effect size. Adaptive designs allow these elements to be reviewed during the trial. This is desirable because there may be limited information to guide these choices initially, but more knowledge will accrue as the study progresses. For statisticians the term “adaptive designs” refers to a body of statistical methodology that borrows ideas from both group sequential and multiple comparison procedures along with elements of meta-analysis.

Here we shall only discuss that aspect of adaptive design that concerns the issue of modifying the power or conditional power at an interim stage by modifying the target sample size. The cause for such a modification may be unexpected: for example, external reasons may make a smaller effect size suddenly commercially viable so it becomes worth detecting. But more usually it is because the study was originally underpowered due to over-optimism or an initial reluctance to commit sufficient resources. Later however, a moderate interim estimate of the true effect size shows that the trial is headed for a negative conclusion even though that effect is positive and still worth detecting. Indeed, this was the motivation for one of the seminal papers in this area — see Cui *et al.* (1999).

Suppose θ is a parameter of primary interest. Consider a group sequential study with up to K analyses which yields the sequence of standardized statistics $\{Z_1, \ldots, Z_K\}$. We say that these statistics have the *canonical joint distribution* with information levels $\{\mathcal{I}_1, \ldots, \mathcal{I}_K\}$ for the parameter θ if:

$$\begin{aligned}
&\text{(i)} \quad (Z_1, \ldots, Z_K) \text{ is multivariate normal,} \\
(1) \qquad &\text{(ii)} \quad E(Z_k) = \theta\sqrt{\mathcal{I}_k}, \quad k = 1, \ldots, K, \quad \text{and} \\
&\text{(iii)} \quad Cov(Z_{k_1}, Z_{k_2}) = \sqrt{(\mathcal{I}_{k_1}/\mathcal{I}_{k_2})}, \quad 1 \le k_1 \le k_2 \le K.
\end{aligned}$$

This canonical joint distribution arises in a great many situations — in two sample normal problems; for normal responses with covariates; in parallel and crossover designs, etc. It also arises approximately with binary and survival responses. For details, see Jennison and Turnbull (2000). For example, in the one sample problem where independent observations $X_1, X_2, \ldots$ are normally distributed with unknown mean θ and known variance σ^2, then $Z_k = \sum_{i=1}^{n_k} X_i/(\sigma\sqrt{n_k})$ and $\mathcal{I}_k = n_k/\sigma^2$ is proportional to the cumulative sample size n_k, $k = 1, \ldots, K$.

A one-sided group sequential test of the null hypothesis H_0: $\theta \le 0$ against $\theta > 0$ takes the form:

$$\begin{array}{lll} \text{After group } k=1,\dots,K-1 & & \\ & \text{if } Z_k \ge b_k & \text{stop, reject } H_0 \\ & \text{if } Z_k \le a_k & \text{stop, accept } H_0 \\ & \text{otherwise} & \text{continue to group } k+1, \\ \text{After group } K & & \\ & \text{if } Z_K \ge b_K & \text{stop, reject } H_0 \\ & \text{if } Z_K < a_K & \text{stop, accept } H_0, \end{array} \tag{2}$$

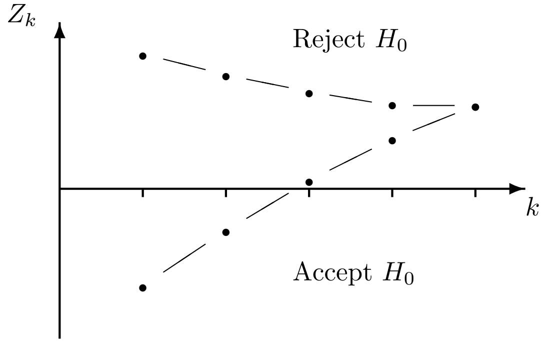

Figure 1: Stopping boundary for K=5

where $a_K = b_K$ to ensure termination at analysis K — see Figure 1. Typically, tests are designed with analyses at equally-spaced information levels (or "group sizes") $\Delta_1 = \ldots = \Delta_K$ where $\Delta_k = \mathcal{I}_k - \mathcal{I}_{k-1}; (k = 1, \ldots, K)$ with $\mathcal{I}_0 = 0$. Then, for given K, the maximum information $\mathcal{I}_K$ and boundary values (a_k, b_k), $k = 1, \ldots, K$, can be chosen to attain type I error probability α under $\theta = 0$ and power $1 - \beta$ at an alternative $\theta = \delta$. This computation uses the distribution (1) for the $\{Z_k\}$ and implicitly uses the fact that the information levels $\{\mathcal{I}_k\}$ are not influenced by responses $\{Z_k\}$.

Suppose a test of the above form is under way. However, based on data observed at analysis j, it is desired to increase the size of the planned succeeding information levels. Now it is no longer true that future information levels are independent of previous responses. Indeed, if we continued to use the boundary values $\{a_k,\ b_k\}$ given in (2), the Type I error rate is no longer guaranteed at α and is typically inflated.

Suppose, however, we go ahead with the adaptation so that now the cumulative information levels are $\tilde{\mathcal{I}}^{(1)}, \ldots, \tilde{\mathcal{I}}^{(K)}$ instead of the originally planned $\{\mathcal{I}_1, \ldots, \mathcal{I}_K\}$. (Of course $\tilde{\mathcal{I}}^{(k)} = \mathcal{I}_k$ for $k \le j$.) We *can* still maintain the Type I error with the same boundary if we proceed as follows. Let $\tilde{Z}^{(k)}$ be the usual Z-statistic *formed from data in stage k alone* and $\tilde{\Delta}_k = \tilde{\mathcal{I}}^{(k)} - \tilde{\mathcal{I}}^{(k-1)}$ the associated increment in information. Note that, even though the information increment $\tilde{\Delta}^{(k)}$ is an ingredient of the statistic $\tilde{Z}^{(k)}$ and can depend on knowledge of the past $\tilde{Z}^{(1)}, \ldots, \tilde{Z}^{(k-1)}$, under $\theta = 0$, each $\tilde{Z}^{(k)}$ still has a standard normal N(0,1) distribution conditionally — and hence unconditionally. Thus, under H_0, we may treat $\tilde{Z}^{(1)}, \tilde{Z}^{(2)}, \ldots$ as independent. Of course this is not true under the alternative $\theta > 0$, which is, of course, the reason why adapting the information levels can lead to increased power. Therefore we may use the same boundary values given in (2) and maintain the specified Type I error rate α, provided, instead of the $\{Z_k\}$, we monitor statistics:

$$\tilde{Z}(k) \;=\; (w_1 \tilde{Z}^{(1)} + \ldots + w_k \tilde{Z}^{(k)}) \,/\, (w_1^2 + \ldots + w_k^2)^{1/2}. \tag{3}$$

for $k = 1, \ldots, K$, where weights $w_k = \sqrt{\Delta_k};\ k = 1, \ldots, K$ are the square roots of the originally planned information increments. This follows because, under H_0, it is easy to see that the $\tilde{Z}(k)$ follow the same canonical joint distribution (1) — see Lehmacher and Wassmer (1999). Use of a procedure based on (3) is an example of a *combination* test — Bauer and Köhne, K. (1994).

To summarize so far, we have seen how, by using (3), the investigator has the freedom to decide how to modify the study in light of accruing data, yet maintain the Type I error. But what is the cost, if any, of this flexibility? To examine this question, we need to consider specific strategies for adaptive design. Jennison and Turnbull (2006a) consider the example of a group sequential test (GST) with $K = 5$ analyses testing H_0: $\theta \le 0$ against $\theta > 0$ with type I error probability $\alpha = 0.025$ and power $1 - \beta = 0.9$ at $\theta = \delta$. A fixed sample size test for this problem requires information for θ

$$\mathcal{I}_f = (z_\alpha + z_\beta)^2/\delta^2, \tag{4}$$

where z_p denotes the $1 - p$ quantile of the standard normal distribution. Suppose the study is designed as a one-sided test from the ρ-family of error-spending tests, as described in Jennison and Turnbull (2000, Sec.7.3), for example. We choose index $\rho = 3$. The boundary values $a_1, \ldots, a_5$ and $b_1, \ldots, b_5$ are chosen to satisfy

$$\Pr{}_{\theta}\{Z_1 > b_1 \text{ or } \ldots \text{ or } Z_1 \in (a_1, b_1), \ldots, Z_{k-1} \in (a_{k-1}, b_{k-1}), Z_k > b_k\} = (\mathcal{I}_k/\mathcal{I}_{max})^{\rho}\, \alpha,$$

$$\Pr{}_{\theta}\{Z_1 < a_1 \text{ or } \ldots \text{ or } Z_1 \in (a_1, b_1), \ldots, Z_{k-1} \in (a_{k-1}, b_{k-1}), Z_k < a_k\} = (\mathcal{I}_k/\mathcal{I}_{max})^{\rho}\, \beta$$

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

Suppose external information becomes available at the second analysis, leading the investigators to seek conditional power of 0.9 at $\theta = \delta/2$ rather than $\theta = \delta$. Since this decision is independent of data observed in the study, one might argue that modification could be made without prejudicing the type I error rate. However, it would be difficult to prove that the data revealed at interim analyses had played no part in the decision to re-design. Following the general method described in [1], it is decided to change the information increments in the third, fourth and fifth stages to $\tilde{\Delta}_k = \gamma\Delta_k$ for $k = 3, 4, 5$. The factor γ depends on the data available at stage 2 and is chosen so that the conditional power under $\theta = \delta/2$, given the observed value of Z_2, is equal to $1 - \beta = 0.9$. However γ is truncated to lie in the range 1 to 6, so that sample size is never reduced and the maximum total information is increased by at most a factor of 4. Figure 2(a) shows that the power curve of the adaptive test lies well above that of the original group sequential design. The power 0.78 attained at $\theta = 0.5\,\delta$ falls short of the target of 0.9 because of the impossibility of increasing conditional power when the test has already terminated to accept H_0 and the truncation of γ for values of Z_2 just above a_2.

It is of interest to assess the cost of the delay in learning the ultimate objective of the study. Our comparison is with a ρ-family error-spending test with $\rho = 0.75$, power 0.9 at $0.59\,\delta$ and the first four analyses at fractions 0.1, 0.2, 0.45 and 0.7 of the final information level $\mathcal{I}_5 = \mathcal{I}_{max} = 3.78\,\mathcal{I}_f$. This choice ensures that the power of the non-adaptive test is everywhere as high as that of the adaptive test, as seen in Fig. 2(a), and the expected information curves of the two tests are of a similar shape. Fig. 2(b) shows the expected information on termination as a function of θ/δ for these two tests; the vertical axis is in units of $\mathcal{I}_f$. Together, Figures 2(a) and 2(b) show that the non-adaptive test dominates the adaptive test in terms of both power and expected information over the range of θ values. Also, the non-adaptive test's maximum information level of $3.78\,\mathcal{I}_f$ is 10% lower than the adaptive test's $4.20\,\mathcal{I}_f$.

It is useful to have a single summary of relative efficiency when two tests differ in both power and expected information. If test A with type I error rate α at $\theta = 0$ has power function $1 - b_A(\theta)$ and expected information $E_{A,\theta}(\mathcal{I})$ under a particular $\theta > 0$, we define its efficiency index at θ to be

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

the ratio of the information needed to achieve power $1 - b_A(\theta)$ in a fixed sample test to $E_{A,\theta}(\mathcal{I})$. In comparing tests A and B, we take the ratio of their efficiency indices to obtain the efficiency ratio

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

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

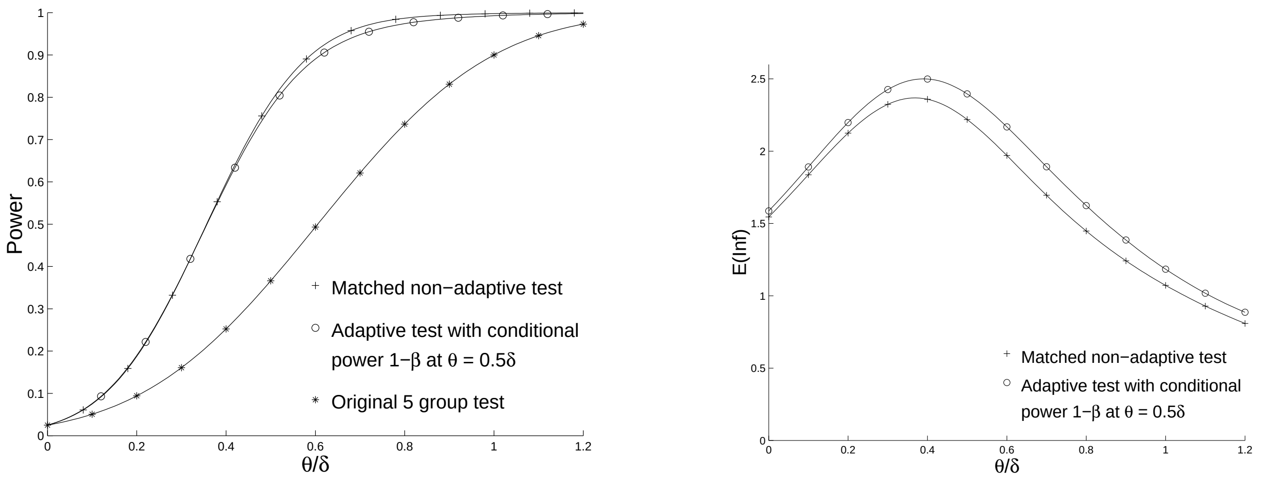

Figure 2: (a)Left panel: Power of the original test; of example adaptive design with sample size revised at look 2 to attain conditional power 0.9 at $\theta = 0.5\,\delta$; and of matched non-adaptive test (b)Right panel: $E_\theta(\mathcal{I})$ of the example adaptive design and the matched non-adaptive design, expressed in units of $\mathcal{I}_f$.

The plot in Fig. 3 shows the adaptive design is considerably less efficient that the simple group sequential test, especially for $\theta > \delta/2$. We have studied a variety of proposed adaptive designs and found similar inefficiencies to the above example. These include methods of Bauer and Köhne (1994), Proschan and Hunsberger (1995), Shen and Fisher (1999), Li *et al.* (2002). We have also found similar inefficiencies in adaptive designs which increase sample size in direct response to low interim estimates of the treatment effect. See the second example in Jennison and Turnbull (2006a) and further discussion in that paper. When adaptation makes smaller increases in sample size, the increase in power is smaller but efficiency loss is still present.

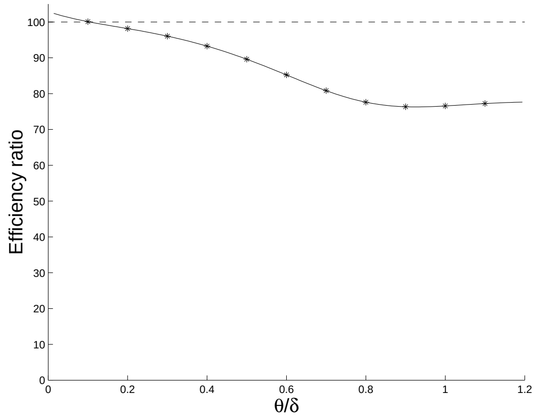

Figure 3: Efficiency ratio between example adaptive design and the matched non-adaptive design.

This leads to a seeming paradox: Since adaptive designs include adaptive designs as a sub-class, should they not be able to be more efficient? The resolution of the paradox is to realize that in order to have optimal efficiency properties, a sequential procedure should be based on the sufficient statistics $Z_1, \ldots, Z_K$ and, in particular, not the weighted combination statistics $\tilde{Z}(1), \ldots, \tilde{Z}(K)$ given in (3); also, the sample size modification rule should have the correct form for an efficient design and this is quite different from that given by the conditional power rules that are often proposed — see Jennison and Turnbull (2006a).

We now describe a formulation for examining optimality within a given class of group sequential tests (GSTs). An 'optimal" sequential test may be used directly or it can serve as a benchmarks for other tests proposed for their convenience or intuitive appeal. We consider again the one-sided testing problem of $H_0 : \theta \leq 0$ versus $H_A : \theta > 0$. We set the Type I error rate α and specify power $1 - \beta$

at $\theta = \delta$. The fixed sample test needs information $\mathcal{I}_f$ as given by (4). We specify the maximum number K of looks allowed and the maximum information allowed as $\mathcal{I}_{max} = R\mathcal{I}_f$. Here R is termed the "inflation factor". As special cases, K or R could be set to ∞ if we do not wish to place an upper bound on them. With these constraints we must look within the specified family of GSTs for that one that minimizes the average information (*e.g.* sample size) at one θ-value or averaged over several θ-values. To find this optimum procedure we must search for that sequence of information levels $\{\mathcal{I}_k\}$ and stopping boundary values $\{(a_k, b_k)\}$ that maximize the average expected sample size criterion subject to the α and β constraints. This involves searching in a high dimensional space. Rather than search this space directly, we create a sequential Bayes decision problem with a prior on θ, sampling costs, and costs for a wrong decision. The solution is found by a backward induction (dynamic programming) technique. Then a search (in two dimensions) over cost parameters leads to a Bayes problem with solution equal to the optimal GST with error rates equal to the specified α and β being sought. This is essentially a Lagrangian method for solving a constrained optimization problem. See Barber and Jennison (2002), Jennison and Turnbull (2006a) for more details.

We will consider optimal procedures from within each of the following three nested families of K-stage GST's which are of interest:

A. **Equal group sizes.** These tests have equally spaced analyses: $\mathcal{I}_k = (k/K)R\mathcal{I}_f$, $k = 1, \ldots, K$. Optimization is over the choice of boundary values $\{(a_k, b_k)\}$. Planning with equal group sizes is the usual starting point when designing a GST.

B. **Non-adaptive GSTs.** Complete freedom is allowed in choosing critical values (a_k, b_k) *and* cumulative information levels $\mathcal{I}_1, \ldots, \mathcal{I}_K$ to optimize the efficiency criterion subject to $\mathcal{I}_K \leq R\mathcal{I}_f$. In particular, the initial information level $\mathcal{I}_1$ is allowed to be small which may be advantageous if it is important to stop very early when there is a large treatment benefit — the "home run" treatment. Importantly, note that the choice of the $\{\mathcal{I}_k\}$ and $\{(a_k, b_k)\}$, $k = 1, \ldots, K$, is set at the start of the study and cannot be updated as observations accrue.

C. **Adaptive GSTs.** These are fully adaptive designs. At each analysis $k = 1, \ldots, K-1$, the next cumulative information level $\mathcal{I}_{k+1}$ and critical values (a_{k+1}, b_{k+1}) are chosen based on current data. The whole procedure is chosen to optimize the efficiency criterion subject to $\mathcal{I}_K \leq R\mathcal{I}_f$.

The class of designs (C) was first considered by Schmitz (1993). Note that while the Schmitz designs are adaptive in the sense that future increments in information levels are allowed to depend on past and current values of the Z-statistic, these designs are not "flexible". The way in which future information levels can depend on the past and current values of the Z-values is specified the start of the study. They cannot be chosen arbitrarily, unlike the procedures based on combination test statistics (3).

As an example of comparing families of tests, Jennison and Turnbull (2006b) consider the situation of testing H_0: $\theta = 0$ versus H_1: $\theta > 0$ with $\alpha = 0.025$ and power $1 - \beta = 0.9$ at $\theta = \delta$. They take as their efficiency criterion low values of $\int E_\theta(\mathcal{I})\, f(\theta)\, d\theta$, where $\mathcal{I}$ is the information attained when the stopping boundary is crossed and $f(\theta)$ is the density of a $N(\delta,\, \delta^2/4)$ distribution. In this case we impose the constraint on maximum information $\mathcal{I}_K \leq R \times \mathcal{I}_f$ with $R = 1.2$. Table 1 shows the optimal values of the efficiency criterion for classes (A),(B),(C) above as a percentage of the fixed sample information for values of K=1–6, 8, 10. We see that the advantage of varying group sizes *adaptively* is small — but it is present. On the other hand, such a procedure is much more complex than its non-adaptive counterparts.

Table 1: Optimal Average $E(\mathcal{I})$ as a percentage of the fixed sample information

K	*Optimal non-adaptive equal group sizes* (A)	*Optimal non-adaptive optimised group sizes* (B)	*Optimal adaptive design (Schmitz)*(C)
1	100.0	100.0	100.0
2	74.8	73.2	72.5
3	66.1	65.6	64.8
4	62.7	62.4	61.2
5	60.9	60.5	59.2
6	59.8	59.4	58.0
8	58.3	58.0	56.6
10	57.5	57.2	55.9

In the same setup, Jennison and Turnbull (2006b) also consider the efficiency criterion:

$$\{E_{\theta=0}(\mathcal{I}) + E_{\theta=\delta}(\mathcal{I}) + E_{\theta=L\delta}(\mathcal{I})\}/3,$$

with $L = 2$. This criteria might be appropriate if very early stopping is important then the treatment is very effective ($\theta = L\,\delta$ with $L > 1$). However now we fix K=2 and consider varying R.

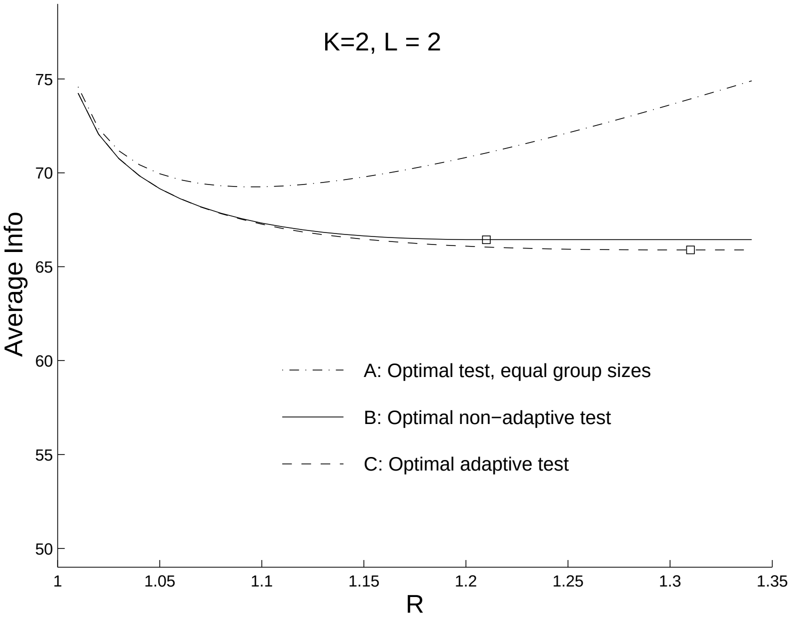

Figure 4: Optimized average information $\mathcal{I}$ plotted against R for tests in classes A, B, C. The small box symbols on the curves for classes B and C indicate the values of R at which these curves become flat.

Figure 4 shows how the optimized average information $E(\mathcal{I})$ varies with R. For class A tests, average information decreases initially as R increases but eventually starts to increase again. The group sizes of class B tests are optimized subject to the constraint $\mathcal{I}_K/\mathcal{I}_f \leq R$: initially the optimal tests have $\mathcal{I}_K/\mathcal{I}_f = R$ but as R is increased a value, $\tilde{R}_B$ say, is reached such that the optimal test continues to take $\mathcal{I}_K/\mathcal{I}_f = \tilde{R}_B$ even though higher values are allowed. Similarly, for given L and K, there is a maximal inflation factor, $\tilde{R}_C$ say, for optimal tests in class C: even when larger values of $\mathcal{I}_K$ are permitted, all sample paths terminate with $\mathcal{I}_K \leq \tilde{R}_C\mathcal{I}_f$. The points at which the curves for classes B and C reach their plateaus are marked in the figures but it is clear that values close to the optimal average $\mathcal{I}$ are reached well before these points. Given the practical disadvantages of a high value of R, it is reasonable to choose an inflation factor R considerably lower than these otherwise "optimal" values – between 1.1 and 1.5, say. Since the curves B and C are close, we see that there is an advantage of varying group sizes *adaptively* but this is slight, and most likely not worth the cost of the extra complexity. This is in agreement with the conclusions from Table 1. The same features occur when other values of K and L are considered — see Figures 1 and 2 and Tables I–IV of Jennison and Turnbull (2006b).

We have seen that the pre-planned adaptive designs of Schmitz (1993) can be slightly more efficient than conventional group sequential tests. However the more commonly used adaptive tests, namely those based on combination statistics such as (3), are typically 10-25% less efficient. Why is this? There are three contributing reasons:

1. *Use of non-sufficient statistics.* In Jennison and Turnbull (2006a) it is proved all admissible designs (adaptive or non-adaptive) are Bayes procedures. Hence, their decision rules and sample size rules must be functions of sufficient statistics. Unequal weighting of observations in adaptive designs means these are not based on sufficient statistics. Thus, they cannot be optimal designs for any criteria. The potential benefits of adaptivity are slight and any departure from optimality can leave room for an efficient non-adaptive design, with the same number of analyses, to do better. Note that this is stronger conclusion than that of Tsiatis and Mehta (2003) who allow the comparator non-adaptive design to have additional analyses.

2. *Sub-optimal sample size modification rule.* Rules based on conditional power differ qualitatively from those found for optimal adaptive designs. Conditional power rules invest a lot of resource in unpromising situations with a low interim estimate of the treatment effect. The optimal rule shows greater symmetry, taking higher sample sizes when the current test statistic is in the middle of the continuation region, away from both boundaries. See Figure 5.

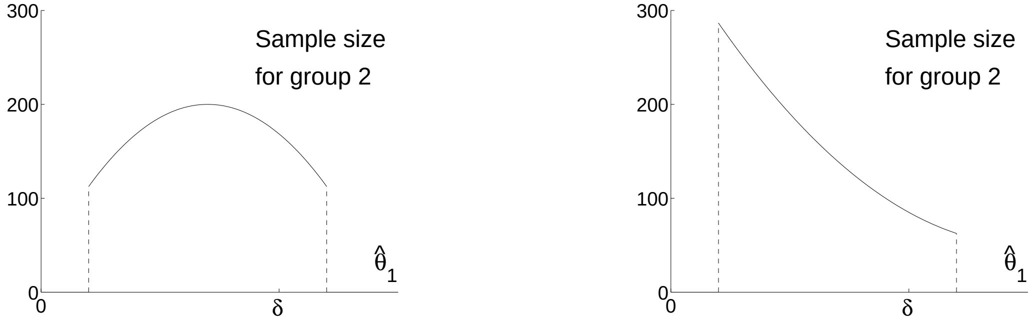

Figure 5: (a) Left panel: Typical shape of sample size function for an optimal adaptive test; (b) Right panel: Typical shape of sample size function for a conditional power adaptive design.

3. *Over-reliance on a highly variable interim estimator of θ.* The sample size modification rules of many adaptive designs involve the current interim estimator of effect size which is highly variable. This introduces extra noise and results in random variation in sample size that is in itself inefficient; see Jennison and Turnbull (2003) for further discussion of this point in the context of a two-stage design.

In conclusion, non-adaptive group sequential tests are well studied and optimal tests have been derived for a variety of criteria. Incorporating adaptivity in pre-planned group sequential designs, as proposed by Schmitz (1993) produces a small benefit. Using adaptive methods in an unplanned manner offers flexibility to the organisers of a study but, since the sufficiency principle is contravened, there is an efficiency cost. (It can also lead to some pathological results — see Burman and Sonesson (2006).) One argument for flexible adaptive designs is that they allow investigators to choose a study's power curve in response to early estimates of the effect size, θ. This may be appealing when there is uncertainty about the likely effect size and optimistic estimates are considerably larger than the minimum clinically or commercially significant effect. Schäfer & Müller (2004) consider tests for a range of detectable treatment effects and propose a design in which attention shifts to smaller effect sizes at successive analyses. An alternative solution is simply to specify high power at the small but clinically significant effect size and choose a group sequential test that achieves this while giving low expected sample size under larger effects — see Jennison and Turnbull (2006b).

A key role that remains for flexible adaptive methods is to help investigators respond to unexpected external events. As several authors have pointed out, it is good practice to design a study as efficiently as possible given initial assumptions, so the benefits of this design are obtained in the usual circumstances where no mid-course change is required. However, if the unexpected occurs, adaptive methods can be applied using the approach of maintaining conditional type I error probability — Denne(2001), Müller & Schäfer (2001). Finally, the use of flexible adaptive methods to rescue an under-powered study should not be overlooked. While it is easy to be critical of a poor initial choice of sample size, it would be naive to think that such problems will cease to arise.

REFERENCES (RÉFERENCES)

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

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

Burman, C-F. and Sonesson, C. (2006). Are flexible designs sound? *Biometrics* **62**, 664–669.

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

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

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

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

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

Jennison, C. and Turnbull, B.W. (2006b). Efficient group sequential designs when there are several effect sizes under consideration. *Statistics in Medicine* **35**(6), 917-932.

Lehmacher, W. and Wassmer, G. (1999). Adaptive sample size calculation in group sequential trials. *Biometrics* **55**, 1286–1290.

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

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.

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. New York: Springer-Verlag.

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

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