From Group Sequential to Adaptive Designs

Christopher Jennison
Department of Mathematical Sciences,
University of Bath, Bath BA2 7AY, U. K.
email: cj@maths.bath.ac.uk
and
Bruce W. Turnbull
Department of Statistical Science,
227 Rhodes Hall, Cornell University, Ithaca, New York 14853-3801, U. S. A.
*email:*bwt2@cornell.edu

December 16, 2009

1 Introduction

In long-term experiments, it is natural to wish to examine the data as they accumulate instead of waiting until the conclusion. However it is clear that, with frequent looks at the data, there is an increased probability of seeing spurious results and making a premature and erroneous decision. To overcome this danger of over-interpretation of interim results, special statistical analysis methods are required. To address this need, the first classic books on sequential analysis were published by Wald (1947), motivated primarily by quality control applications, and by Armitage (1960) for medical trials. In this chapter, we shall be concerned with the latter application. The benefits of monitoring data in clinical trials are obvious:

Administrative. One can check on accrual, eligibility and compliance, and generally ensure the trial is being carried out as per protocol.

Economic. Savings in time and money can result if the answers to the research questions become evident early — before the planned conclusion of the trial.

Ethical. In a trial comparing a new treatment with a control, it may be unethical to continue subjects on the control (or placebo) arm once it is clear that the new treatment is effective. Likewise if it becomes apparent that the treatment is ineffective, inferior or unsafe, then the trial should not continue.

It is now standard practice for larger Phase III clinical trials to have a Data Monitoring Committee (DMC) to oversee the study and consider the option of early termination. Note that many of the same considerations apply to animal and epidemiologic studies as well.

It was soon recognized by researchers that fully sequential procedures, with continuous monitoring of the accumulating data, were often impractical and, besides that, much of the economic savings could be achieved by procedures that examined the data on a limited number of occasions throughout the trial — at six month intervals, for example, in a multi-year trial. The corresponding body of statistical analysis and design techniques has become known as *group sequential methodology* because the accumulated data are examined after observing each successive *group* of new observations. There is a large body of literature in the biostatistical and medical journals and there have been several comprehensive books published. These include Whitehead (1997), Jennison and Turnbull (2000) and Proschan, Lan and Wittes (2006). Of related interest are books on the practical considerations for the operation of DMCs by Ellenberg, Fleming

and DeMets (2003) and Herson (2009) and a collection of case studies by DeMets, Furberg and Friedman (2006).

In this chapter, we shall survey some of the major ideas of group sequential methodology. For more details, the readers should refer to the aforementioned books. In particular, we shall cite most often the book by Jennison and Turnbull (2000) — hereafter referred to as "JT", because clearly it is the most familiar to us! In the spirit of this current volume, we shall also show how much flexibility and adaptability is already afforded by "classical" group sequential procedures. Then, we shall show how these methods can naturally be embodied in the more recently proposed adaptive procedures, and *vice versa*, and consider the relative merits of the two types of procedures. We conclude with some discussion and also provide a list of sources of computer software to implement the methods we describe.

2 The Canonical Joint Distribution of Test Statistics

The statistical properties of a group sequential procedure (GSP) will depend on the joint distribution of the accumulating test statistics being monitored and the decision rules that have been specified in the protocol. We start with the joint distribution. For motivation, consider the simple "prototype" example of a balanced two-sample normal problem. Here we sequentially observe responses X_{A1}, X_{A2}, ... from Treatment A and X_{B1}, X_{B2}, ... from Treatment B. We assume that the $\{X_{Ai}\}$ and $\{X_{Bi}\}$ are independent and normally distributed with common variance σ^2 and unknown means μ_A and μ_B, respectively. Here $\theta = \mu_A - \mu_B$ is the parameter of primary interest.

At interim analysis (or "look" or "stage") k ($k = 1, 2, \ldots$), we have cumulatively observed the first n_k responses from each treatment arm with $n_1 < n_2 < \ldots$. Then the standardized test statistic based on all the responses so far is $Z_k = \sum_{i=1}^{n_k} (X_{Ai} - X_{Bi})/(\sigma\sqrt{2n_k})$. It is easy to verify that the joint distribution of $Z_1, \ldots, Z_K$ has the defining properties:

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

where $\mathcal{I}_k = n_k/(2\sigma^2)$ is termed the *information* or *information level* at stage k.

If a GSP with up to K analyses yields the sequence of statistics $Z_1, \ldots, Z_K$ for the parameter of interest θ, and their joint distribution satisfies (1), we say that these statistics have the *canonical joint distribution* with information levels $\{\mathcal{I}_1, \ldots, \mathcal{I}_K\}$ for θ. In fact, this canonical joint distribution

arises in a great many situations, not just the balanced two sample normal problem defined above; see Jennison and Turnbull (1997). The list includes unbalanced two-sample comparisons; comparison of normal responses adjusted for baseline covariates; longitudinal data; parallel and crossover designs, etc. Calculation of the $\{Z_k\}$ defined above requires σ^2 to be known, but if the variance is unknown the theory applies approximately to the sequence of t-statistics. The same canonical joint distribution also holds approximately for binary and survival data. The specific details on how to construct the appropriate $\{Z_k\}$ and $\{\mathcal{I}_k\}$ sequences in each application are described in Chapter 3 of JT. Typically, Z_k is the Wald statistic for testing $\theta = 0$ and $\mathcal{I}_k$ is the reciprocal of the variance of the maximum likelihood (or other efficient) estimator of θ, each based on the accumulated data at stage k. The key conclusion is that statistical properties based on particular decision boundaries can be computed from (1) and the results will be applicable to a very wide variety of situations, enabling a unified theory.

3 Hypothesis Testing Problems and Decision Boundaries with Equally Spaced Looks

A decision boundary provides critical values for Z_k at each stage k, which determine whether to stop or continue the trial. If the decision is to stop, the action to be taken is also specified. Various shapes for boundaries have been proposed and these shapes depend on the hypotheses about θ to be tested. Initially, we assume a maximum number of looks K is specified and these are to be taken at equal increments of information — that is $\mathcal{I}_k = (k/K)\,\mathcal{I}_K$, for $k = 1, \ldots, K$. Later we shall relax this assumption, but it will still be convenient to use the equal information increment assumption initially for planning purposes.

3.1 Two-sided tests

For a two-sided test, the hypotheses are:

$$H_0\colon\ \theta = 0 \qquad \text{versus} \qquad H_A\colon\ \theta \neq 0.$$

We set Type I and Type II error probability constraints:

$$\Pr_{\theta=0}\{\text{Reject } H_0\} \;=\; \alpha, \tag{2}$$

$$\Pr_{\theta=\pm\delta}\{\text{Reject } H_0\} \;=\; 1-\beta. \tag{3}$$

Here α and β are specified (typical values might be $\alpha = 0.05$ and $\beta = 0.1$ or 0.2), and δ is a given effect size that it is important to detect. A fixed sample test ($K = 1$) that meets these requirements would reject H_0 when $|Z| \geq z_{\alpha/2}$ and requires information

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

where z_γ denotes the upper $100\,\gamma$ percentage point of the standard normal distribution.

The decision boundary for a procedure with a maximum of K looks takes the form:

$$\begin{array}{l} \text{After group } k = 1, \ldots, K-1 \\ \qquad \begin{array}{ll} \text{if } |Z_k| \geq c_k & \text{stop, reject } H_0 \\ \text{otherwise} & \text{continue to group } k+1, \end{array} \\ \text{After group } K \\ \qquad \begin{array}{ll} \text{if } |Z_K| \geq c_K & \text{stop, reject } H_0 \\ \text{otherwise} & \text{stop and report failure to reject } H_0, \text{ i.e., "accept" } H_0. \end{array} \end{array} \tag{5}$$

A typical boundary for $K = 5$ is illustrated in Figure 1.

[Figure 1 about here.]

From (1) we see that, in this case of equally spaced information increments, the joint distribution of $Z_1, \ldots, Z_K$ under H_0 does not depend on $\mathcal{I}_K$. Therefore the Type I error rate depends solely on the choice of $c_1, \ldots, c_K$. Once these are chosen to satisfy (2), $\mathcal{I}_K$ can be chosen to satisfy (3). However, there are still many ways to choose the $\{c_k\}$ to satisfy (2).

Wang and Tsiatis (1987) suggested a family of boundaries indexed by the parameter γ, in which $c_k = C(k/K)^\gamma$ for $k = 1, \ldots, K$. The value of C is determined by constraint (2) and depends on K, α and the value of γ. Taking $\gamma = 0$ yields a Pocock (1977) boundary where c_k remains constant over k. We denote by $C_P(K, \alpha)$ the value of C for this test with K analyses and type I error probability α, then $c_k = C_P(K, \alpha)$ for each $k = 1, \ldots, K$. The case $\gamma = -1/2$ yields an O'Brien and Fleming (1979) boundary. For this case, we denote the value of C in the Wang and Tsiatis formula by $C_B(K, \alpha)$ and the boundary is $c_k = C_B(K, \alpha)\sqrt{(K/k)}$ for $k = 1, \ldots, K$. These values of c_k decrease with k — similar to those depicted in Figure 1. Tables of the constants $C_P(K, \alpha)$ and $C_B(K, \alpha)$ can be found in the papers referred to above and in JT, Chap. 2. The tables in JT also give values of the so-called *inflation factor*, denoted by R. Then the value of $\mathcal{I}_K$ needed to satisfy (3) can be found from the formula $\mathcal{I}_K = R\mathcal{I}_{f,2}$, where $\mathcal{I}_{f,2}$ is given by (4). We reproduce

Tables 2.1 to 2.4 of JT, Chap. 2 here for ease of reference as Tables 1 to 4. The constants $C_P(K,\alpha)$ and $C_B(K,\alpha)$ are given in Tables 1 and 2 and inflation factors $R_P(K,\alpha)$ and $R_B(K,\alpha)$ in Tables 3 and 4. We discuss the construction of the entries in these tables of constants in Section 4.

[Tables 1 to 4 about here.]

As an example, suppose we specify an O'Brien and Fleming GSP with a maximum of $K = 5$ analyses and $\alpha = 0.05$. From Table 2.3 of JT, we see that for $\gamma = -1/2$ the constant $C = 2.04$ and the boundary values are $c_1 = 4.56$, $c_2 = 3.23$, $c_3 = 2.63$, $c_4 = 2.28$ and $c_5 = 2.04$. These values can be compared with the fixed sample critical value of $z_{0.025} = 1.96$. The wider boundary values are to compensate for the fact that the test statistic is being examined multiple times (here five). Suppose we additionally ask for power $1 - \beta = 0.9$ at effect size $\pm\delta$. From Table 2.4 of JT, the inflation factor is $R = 1.026$, which means that the maximum information needed will be 2.6% more than the fixed sample test would require. For the prototype two-sample normal problem described at the beginning of Section 2, the fixed sample information $\mathcal{I}_{f,2} = (z_{\alpha/2} + z_\beta)^2/\delta^2$ corresponds to a sample size $n_f = 2\,\sigma^2\,\mathcal{I}_{f,2}$ per treatment arm. Of course, with the group sequential stopping rule there is a good possibility of stopping earlier than stage K. For example, if $\mu_A - \mu_B = \delta$ the expected information and number of observations at termination are only 76% of their values for the fixed sample test. For $\mu_A - \mu_B = 1.5\,\delta$ this proportion is 56%; see JT, Table 2.5. The modest increase in maximum information ($R > 1$) is a small price to pay for the advantages of possible early stopping.

Note that this decision boundary does not permit early stopping to accept H_0, i.e., for *futility*. It is possible to have an "inner wedge" boundary that does allow such a feature (see JT, Chap. 5), but we shall not discuss this further here.

3.2 One-sided tests

Here we test

$$H_0\colon\ \theta = 0 \qquad \text{versus} \qquad H_A\colon\ \theta > 0.$$

We set Type I and Type II error probability constraints:

$$\Pr_{\theta=0}\{\text{Reject } H_0\} \;=\; \alpha, \tag{6}$$

$$\Pr_{\theta=\delta}\{\text{Reject } H_0\} \;=\; 1 - \beta. \tag{7}$$

Typical values might be $\alpha = 0.025$ and $\beta = 0.1$ or 0.2. A fixed sample test ($K = 1$) that meets these requirements would reject H_0 when $Z \geq z_\alpha$ and requires information

$$\mathcal{I}_{f,1} = (z_\alpha + z_\beta)^2/\delta^2. \tag{8}$$

The decision boundary for a procedure with a maximum of K looks takes the form:

$$\begin{array}{ll}
\text{After group } k = 1, \ldots, K-1 & \\
\quad \text{if } Z_k \geq b_k & \text{stop, reject } H_0 \\
\quad \text{if } Z_k \leq a_k & \text{stop, accept } H_0 \\
\quad \text{otherwise} & \text{continue to group } k+1, \\
\text{After group } K & \\
\quad \text{if } Z_K \geq b_K & \text{stop, reject } H_0 \\
\quad \text{if } Z_K < a_K & \text{stop, accept } H_0,
\end{array} \tag{9}$$

where $a_K = b_K$ to ensure termination at analysis K — see Figure 2. Typically, tests are designed with analyses at equally-spaced information levels (or "group sizes") so $\Delta_1 = \ldots = \Delta_K$ where $\Delta_k = \mathcal{I}_k - \mathcal{I}_{k-1}$, $k = 2, \ldots, K$, and $\Delta_1 = \mathcal{I}_1$. 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 satisfy (6) and (7). Several suggestions for choice of boundary values are described in JT, Chap. 4 and results presented there show savings in expected sample size are achieved under $\theta = 0$ and $\theta = \delta$, and also at intermediate values of θ. We shall discuss this one-sided testing problem in more detail later when we look at the case of unequal and unpredictable increments of information between looks.

[Figure 2 about here.]

3.3 One-sided tests with a non-binding lower boundary

The upper boundary in Figure 2 is often called the efficacy boundary and the lower one the futility boundary. Sometimes the futility boundary is considered just a guideline — that is, somewhat arbitrary and *non-binding*, so that investigators may decide to continue a study even though the futility boundary has been crossed with $Z_k < a_k$. In order to protect the Type I error in this case, the left hand side of (6) must be computed assuming $a_1 = \ldots = a_{K-1} = -\infty$, i.e., with no lower boundary. This leads to a higher efficacy boundary and a small decrease in power, but the Type I error probability will be maintained whatever futility boundary is actually used — a feature which

is often important. If the efficacy boundary is constructed this way, it is still useful to have a futility boundary in mind as a guide to stopping for negative results, but applying this futility boundary can only decrease the Type I error and so (6) is always assured.

3.4 Other boundaries

Various other shapes of boundaries could be considered. For example, in a one-sided or two-sided test it may only be desirable to stop early only for futility and not for efficacy. Alternatively, the goal might be to demonstrate equivalence or non-inferiority. We shall not consider these more specialized situations further here but direct the reader to the references cited in Section 1.

4 Computations for Group Sequential Tests — Armitage's Iterated Integrals

This section covers some technical computational details and may be omitted at a first reading. We discuss how the Type I error, power, expected information (or sample size) and other statistical properties of GSPs like the one-sided and two-sided tests discussed in Section 3 can be computed. Our calculations are relevant for any of the models for accumulating data for which the canonical representation (1) applies.

Let $\mathcal{C}_1, \ldots, \mathcal{C}_K$ be subsets of the real line $\Re = (-\infty, \infty)$ representing the continuation regions of a group sequential test. That is, if stage k has been reached, then the procedure stops if $Z_k \notin \mathcal{C}_k$, but otherwise it continues to stage $k+1$. The sets $\mathcal{C}_k$ need not be intervals, but we must have $\mathcal{C}_K = \emptyset$, the empty set, to ensure termination by stage K. The stopping region at stage k is $\mathcal{C}_k^c$, the complement of $\mathcal{C}_k$, and this may be further partitioned into several sets, indicating the appropriate action to be taken upon stopping. If the only action upon stopping is to choose between accepting and rejecting a null hypothesis H_0, then we have two sets, $\mathcal{A}_k$ and $\mathcal{B}_k$ say, where we stop to accept H_0 at stage k if $Z_k \in \mathcal{A}_k$, and stop to reject H_0 at stage k if $Z_k \in \mathcal{B}_k$. In this case $\mathcal{A}_k \cup \mathcal{B}_k \cup \mathcal{C}_k = (-\infty, \infty)$ is a partition of the real line.

As an example, the two-sided procedures of Section 3, are of the above form with

$$\begin{aligned} &\mathcal{A}_k = \emptyset, && \mathcal{B}_k = (-\infty, -c_k) \cup (c_k, \infty), && \mathcal{C}_k = (-c_k, c_k), && \text{for } k = 1, \ldots, K-1, \\ &\mathcal{A}_K = (-c_K, c_K), && \mathcal{B}_K = (-\infty, -c_K) \cup (c_K, \infty), && \mathcal{C}_K = \emptyset. \end{aligned}$$

Similarly, the one-sided procedures of Section 3 are of this form but with

$$\mathcal{A}_k = (-\infty, a_k), \quad \mathcal{B}_k = (b_k, \infty), \quad \mathcal{C}_k = (a_k, b_k), \quad k = 1, \ldots, K,$$

where $a_K = b_K$ so $\mathcal{C}_K = \emptyset$. Decision boundaries of the other procedures we have mentioned (inner wedge designs, equivalence tests, etc.) can also be described in this way.

We define the stopping time T by

$$T = \min\{k\colon Z_k \notin \mathcal{C}_k\}. \tag{10}$$

Note that $1 \le T \le K$ since $\mathcal{C}_K = \emptyset$. We assume $Z_1, \ldots, Z_K$ have the canonical joint distribution (1) and define

$$G_k(z;\theta) = \Pr_\theta\{Z_k \le z,\, T \ge k\}$$

and

$$g_k(z;\theta) = \frac{\partial}{\partial z}\, G_k(z;\theta) \tag{11}$$

for $k = 1, \ldots, K,\ -\infty < z < \infty$ and $-\infty < \theta < \infty$.

From the subdensity $g_k(z;\theta)$ given by (11), we can obtain all the quantities we need. For example, the distribution of the stopping stage is

$$\Pr_\theta\{T = k\} = \int_{\mathcal{C}_k^c} g_k(z;\theta)\,\mathrm{d}z, \quad k = 1, \ldots, K.$$

Similarly, the probability that the procedure stops and takes the action associated with the sets $\{\mathcal{A}_k\}$, say, is

$$\Pr_\theta\{\cup_{k=1}^K \mathcal{A}_k\} = \sum_{k=1}^K \int_{\mathcal{A}_k} g_k(z;\theta)\,\mathrm{d}z.$$

This last expression allows us to compute the size and power of our tests.

The quantities $\{g_k(z;\theta)\}$ involve complicated multi-normal integrals, the numerical computation of which would appear to be quite difficult, especially for larger values of K, say $K > 5$. However the computation is facilitated by noting the Markov structure of the sequence $Z_1,\ Z_2,\ \ldots$. The recursive formulae of Armitage, McPherson and Rowe (1969) can be used to calculate each $g_k(z;\theta)$ in turn. These formulae are:

$$g_1(z;\theta) = \phi(z - \theta\sqrt{\mathcal{I}_1}) \tag{12}$$

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

$$g_k(z;\theta) = \int_{\mathcal{C}_{k-1}} g_{k-1}(u;\theta)\ \frac{\sqrt{\mathcal{I}_k}}{\sqrt{\Delta_k}}\ \phi\left(\frac{z\sqrt{\mathcal{I}_k} - u\sqrt{\mathcal{I}_{k-1}} - \Delta_k\theta}{\sqrt{\Delta_k}}\right)\mathrm{d}u \tag{13}$$

where $\phi(z) = (2\pi)^{-1/2}\exp(-z^2/2)$ denotes the standard normal density.

These equations follow directly from the joint distribution of $\{Z_1, \ldots, Z_K\}$ given by (1). The first equation (12) is immediate and equation (13) follows by noting that (1) implies the sequence

of score statistics $Z_k\sqrt{\mathcal{I}_k}$, $k = 1, \ldots, K$, is Markov with independent normal increments. Thus

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

and is independent of $Z_1, \ldots, Z_{k-1}$. As before, $\Delta_k = \mathcal{I}_k - \mathcal{I}_{k-1}$ denotes the increment in information between analyses $k-1$ and k and for notational completeness we define $Z_0 = \mathcal{I}_0 = 0$.

It follows that only a succession of $K-1$ univariate integrations is needed to evaluate the subdensities $g_k(z;\theta)$ and related probabilities and not a rather complicated K-fold multivariate integral. More details on these computations are given in JT, Chap. 19.

In fact, the computations can be simplified even more by realizing that we only need to carry out the recursive integrations for one value of θ, say $\theta = 0$. Emerson and Fleming (1990) note the following "handy formula" that is useful in converting a sub-density g_k evaluated under one value of θ for computations at another:

$$g_k(z;\theta) = g_k(z;0)\exp(\theta z\sqrt{\mathcal{I}_k} - \theta^2\mathcal{I}_k/2), \quad k = 1, \ldots, K. \tag{14}$$

This result clearly holds for $k = 1$. If we assume the result holds for $k-1$, use of (13) and some algebraic manipulation shows it holds for k and then (14) follows by induction. This is an example of a likelihood ratio identity; see Siegmund (1985, Propositions 2.24 and 3.2).

The computational methods described in this section are used by the various commercial and free software packages that are widely available to aid implementation of group sequential designs and monitoring of accumulating data. More information on available computer software is provided in Section 9.

5 Error Spending Procedures for Unequal, Unpredictable Increments of Information

While monitoring the trial, the increments of information at successive analyses may not be equal. For example, if the meetings of the DMCs are planned for certain calendar times, variations in subject accrual will imply unequal and unpredictable increments in information. In our normal prototype example of Section 2, the information level $\mathcal{I}_k$ at stage k depends on the value of σ^2, which may be unknown and, while the information levels at analysis k can be estimated using the current estimate of σ^2 in the formula $\mathcal{I}_k = n_k/(2\sigma^2)$, this value will not be known in advance. Similarly, if we are collecting binary data and comparing proportions based on a normal approximation to the binomial, the variance will be unknown (JT, Sec. 3.6) and variance estimates from the accumulating

data must be used. In a two-armed trial, when the difference between two treatments is adjusted for baseline covariates, the information at each stage depends on the baseline data which are only observed as subjects enter the study. For survival data endpoints, information is approximately proportional to the total number of events that have occurred so, again, increments are likely to be unequal and unpredictable.

Lan and DeMets (1983) presented two-sided tests which "spend" Type I error as a function of observed information. These methods start with the definition of a (Type I) error spending function $f(\mathcal{I})$. A typical function is depicted in Figure 3. It can be any non-decreasing function with $f(0) = 0$ and $f(\mathcal{I}) = \alpha$ for $\mathcal{I} \geq \mathcal{I}_{max}$. The choice of $\mathcal{I}_{max}$ is discussed below.

[Figure 3 about here.]

The critical value c_k for the stopping boundary at analysis k is chosen to give cumulative Type I error probability $f(\mathcal{I}_k)$ at stage k. The null hypothesis H_0 is accepted if $\mathcal{I}_{max}$ is reached without earlier rejection of H_0. The critical values $\{c_k\}$ are computed iteratively. At the first analysis, the information $\mathcal{I}_1$ is observed and c_1 is obtained by solving the equation: $\Pr_{\theta=0}\{|Z_1| \geq c_1\} = f(\mathcal{I}_1)$. The test stops and rejects H_0 if $|Z_1| > c_1$ and continues otherwise. Now suppose we are at stage $k \geq 2$ and we have observed $\mathcal{I}_1, \ldots, \mathcal{I}_k$. Having already obtained critical values $c_1, \ldots, c_{k-1}$ at the previous analyses, we compute the current critical value c_k by solving for it in the equation:

$$\Pr_{\theta=0}\{|Z_1| < c_1, \ldots, |Z_{k-1}| < c_{k-1}, |Z_k| \geq c_k\} = f(\mathcal{I}_k) - f(\mathcal{I}_{k-1}).$$

This equation can be solved numerically using the computational formulae of Section 4. Note that computation of c_k does *not* depend on future information levels, $\mathcal{I}_{k+1}, \mathcal{I}_{k+2}, \ldots$. In a "maximum information design" the study continues until a boundary is crossed or an analysis reached with $\mathcal{I}_k \geq \mathcal{I}_{max}$. The maximum number of analyses K does not have to be pre-specified in advance but if a particular maximum K is so specified, the study terminates at analysis K with $f(\mathcal{I}_K)$ defined to be α even if $\mathcal{I}_K < \mathcal{I}_{max}$. The value of $\mathcal{I}_{max}$ should be chosen to meet the desired power requirement under a typical or anticipated sequence of information levels. A usual choice is the equally spaced one, i.e., $\mathcal{I}_k = (k/K)\mathcal{I}_{max}$ for $k = 1, \ldots, K$ with a pre-specified K. Chapter 7 of JT presents the family of error spending functions $f(\mathcal{I}) = \min\{\alpha, \alpha(\mathcal{I}/\mathcal{I}_{max})^\rho\}$, $\mathcal{I} \geq 0$, referred to as the "ρ-family", where the index ρ can take positive values. If ρ is large then less of the error is spent early and the decision boundary at $\pm c_k$ is wider there; conversely, early boundaries are narrower when ρ is smaller. For equally spaced looks, the choice of $\rho = 1$ approximates a Pocock (1977) boundary, while $\rho = 3$ approximates an O'Brien and Fleming (1979) boundary.

The error spending construction ensures the Type I error probability is equal to α for any observed sequence of information levels $\{\mathcal{I}_k\}$. This property relies on the fact that the sequence $\{Z_k\}$ follows the canonical joint distribution (1), given the observed $\{\mathcal{I}_k\}$. It is therefore essential that the values of $Z_1, \ldots, Z_{k-1}$ do not affect the next information level $\mathcal{I}_k$ and this precludes, for example, deciding to conduct the next analysis sooner when the current test statistic is close to a boundary. Examples of how such practices can inflate the Type I error probability are given in JT, Sec. 7.4. We shall discuss *adaptive* methods which do allow such response-dependent choice of group sizes in Sections 7.4 and 8, but note for now that these procedures have to be defined in special ways in order to protect Type I error in the presence of such adaptation.

We can construct a one-sided error spending test analogously. We define two non-decreasing functions $f(\mathcal{I})$ and $g(\mathcal{I})$ with $f(0) = g(0) = 0$, $f(\mathcal{I}_{max}) = \alpha$ and $g(\mathcal{I}_{max}) = \beta$, specifying how the Type I and II error probabilities are spent as a function of the accruing information. In a similar manner to the two-sided case, we successively construct pairs of critical values (a_k, b_k) so that

$$\Pr_{\theta=0}\{\text{Reject } H_0 \text{ by analysis } k\} = f(\mathcal{I}_k) \text{ and } \Pr_{\theta=\delta}\{\text{Accept } H_0 \text{ by analysis } k\} = g(\mathcal{I}_k).$$

The value of $\mathcal{I}_{max}$ should be chosen so that the boundaries converge at the final analysis under a typical sequence of information levels, e.g., $\mathcal{I}_k = (k/K)\mathcal{I}_{max}$, $k = 1, \ldots, K$, for an anticipated value K. If we reach $\mathcal{I}_K > \mathcal{I}_{max}$ ("over-running") then solving for a_K and b_K is liable to yield $a_K > b_K$. In this case, keeping b_K as calculated and reducing a_K to b_K guarantees the Type I error rate at α and gains extra power. If the final analysis K is reached with $\mathcal{I}_K$ still less than $\mathcal{I}_{max}$ ("under-running"), again keeping b_K as calculated preserves the Type I error rate at α. However, this time we must increase a_K to b_K and the attained power will be slightly below $1 - \beta$.

Finally we note that, for a fixed sequence of information levels there is a one-to-one correspondence between the procedures of Sections 3.1 and 3.2 defined directly in terms of boundaries for Z-values and the procedures discussed here which are defined by error spending functions.

6 *P*-values and Confidence Intervals

So far we have concentrated on the design and monitoring of a group sequential study in a hypothesis testing framework. However, once we have ended the study, we are usually interested in more than just a decision to accept or reject the null hypothesis. In this section, we shall consider the construction of P-values (which measure the evidence against the null hypothesis) and of

confidence intervals (which give a range of effect sizes θ "consistent" with observed data). Both are to be computed once the procedure has terminated. We shall also, in Section 6.3, describe repeated confidence intervals and repeated P-values which may be used at any interim analysis. The methods described here apply to the one-sided and two-sided procedures of Sections 3.1 and 3.2 which use parametric boundaries, as well as the error spending tests of Section 5.

6.1 *P*-values on termination

We use the notation of Section 4. Let Ω be the sample space defined by the group sequential design, that is, the set of all pairs (k, z) where $z \notin \mathcal{C}_k$ so the test can terminate with $(T, Z_T) = (k, z)$. We denote the observed value of (T, Z_T) by (k^*, z^*). The P-value is the minimum significance level under which a test defined on the sample space Ω can reject H_0 on seeing the observed outcome (k^*, z^*), smaller P-values indicating stronger evidence against H_0. For a continuous response distribution, the P-value should be uniformly distributed under H_0, i.e., $\Pr\{P\text{-value} \leq p\} = p$ for all $0 \leq p \leq 1$.

The P-value for testing H_0 on observing (k^*, z^*) can also be stated as

$$\Pr_{\theta=0}\{\text{Observe } (k, z) \text{ as extreme or more extreme than } (k^*, z^*)\},$$

where "extreme" refers to the ordering of Ω implicit in the construction of tests of H_0 at different significance levels. However, there is no single natural ordering of the points in Ω and several different orderings have been proposed (see JT, Sec. 8.4). Suppose a GSP has continuation regions (a_k, b_k), $k = 1, \ldots, K-1$, then in the "stagewise" ordering of Ω we say (k', z') is higher than (k, z), denoted $(k', z') \succ (k, z)$, if any one of the following three conditions holds:

$$(i)\ k' = k \text{ and } z' \geq z,\ \ (ii)\ k' < k \text{ and } z' \geq b_{k'},\ \ (iii)\ k' > k \text{ and } z \leq a_k.$$

When the GSP is a one-sided test, it is natural to consider a one-sided P-value for testing H_0: $\theta = 0$ versus $\theta > 0$,

$$\Pr_{\theta=0}\{(T, Z_T) \succ (k^*, z^*)\},$$

so higher outcomes in the ordering give greater evidence against H_0.

If the GSP is a two-sided test of H_0: $\theta = 0$ versus $\theta \neq 0$ with continuation regions $(-c_k, c_k)$, we start with the same overall ordering (with $-c_k$ and c_k in place of a_k and b_k in the above definition) but now consider outcomes in both tails of the ordering when defining a two-sided P-value. Consider an O'Brien and Fleming two-sided procedure with $K = 5$ stages, $\alpha = 0.05$ and equal increments

in information. As stated in Section 3.1, the critical values are $c_1 = 4.56$, $c_2 = 3.23$, $c_3 = 2.63$, $c_4 = 2.28$ and $c_5 = 2.04$. The stagewise ordering for this GSP is depicted in Figure 4.

[Figure 4 about here.]

Suppose we observe the values shown by stars in Figure 4, $Z_1 = 3.2$, $Z_2 = 2.9$ and $Z_3 = 4.2$, so the boundary is crossed for the first time at the third analysis and the study stops to reject H_0 with $T = 3$ and $Z_T = 4.2$. The two-sided P-value is given by

$$\Pr_{\theta=0}\{|Z_1| \geq 4.56 \text{ or } |Z_2| \geq 3.23 \text{ or } |Z_3| \geq 4.2\}$$

which can be calculated to be 0.0013, using the methods of Section 4.

Other orderings are possible, but the stagewise ordering has the following desirable properties:

(i) If the group sequential test has two-sided Type I error probability α, the P-value is less than or equal to α precisely when the test stops with rejection of H_0.

(ii) The P-value on observing (k^*, z^*) does not depend on values of $\mathcal{I}_k$ and c_k for $k > k^*$, which means the P-value can still be computed in an error spending test where information levels at future analyses are unknown.

6.2 A confidence interval on termination

We can use a similar reasoning to construct a confidence interval (CI) for θ upon termination. Suppose the test terminates at analysis k^* with $Z_{k^*} = z^*$. A $100(1-\alpha)\%$ confidence interval for θ contains precisely those values θ for which the observed outcome (k^*, z^*) is in the "middle $(1-\alpha)$" of the probability distribution of outcomes under θ.

This can be seen to be the interval (θ_1, θ_2) where

$$\Pr_{\theta=\theta_1}\{(T, Z_T) \succ (k^*, z^*)\} = \alpha/2$$

and

$$\Pr_{\theta=\theta_2}\{\ T, Z_T) \prec (k^*, z^*)\} = \alpha/2.$$

This follows from the relation between a $100(1-\alpha)\%$ confidence interval for θ and the family of level α two-sided tests of hypotheses $H\colon\ \theta = \tilde{\theta}$.

Consider our previous example where an O'Brien and Fleming two-sided procedure with $K = 5$ stages and $\alpha = 0.05$ ended at stage $T = 3$ with $Z_3 = 4.2$ and suppose the observed information levels are $\mathcal{I}_1 = 20$, $\mathcal{I}_2 = 40$ and $\mathcal{I}_3 = 60$. In this case, the computation using Armitage's iterated

integrals (Section 4) yields a 95% CI of (0.20, 0.75) for θ. In contrast, the "naive" fixed sample CI would be (0.29, 0.79) but it is not appropriate to use this interval: failure to take account of the sequential stopping rule means that the coverage probability of this form of interval is *not* $1-\alpha$.

Note that there is a consistency of hypothesis testing and the CI on termination. Suppose a group sequential study is run to test H_0: $\theta = 0$ versus $\theta \neq 0$ with Type I error probability α. Then, a $1-\alpha$ confidence interval on termination should contain $\theta = 0$ if and only if H_0 is accepted. This happens automatically if outcomes for which we reject H_0 are at the top and bottom ends of the sample space ordering — and any sensible ordering does this.

6.3 Repeated confidence intervals and repeated *P*-values

Repeated confidence intervals (RCIs) for a parameter θ are defined as a sequence of intervals I_k, $k = 1, \ldots, K$, for which a simultaneous coverage probability is maintained at some level, $1-\alpha$ say. The defining property of a $(1-\alpha)$-level sequence of RCIs for θ is

$$\Pr_\theta\{\theta \in I_k \text{ for all } k = 1, \ldots, K\} = 1-\alpha \quad \text{for all } \theta. \tag{15}$$

The interval I_k provides a statistical summary of the information about the parameter θ at the kth analysis, automatically adjusted to compensate for repeated looks at the accumulating data. As such, it can be presented to a Data Monitoring Committee (DMC) to be considered with all other relevant information when discussing early termination of a study. The construction and use of RCIs is described in JT, Chap. 9.

If τ is *any* random stopping time taking values in $\{1, \ldots, K\}$, the guarantee of simultaneous coverage (15) implies that the probability I_τ contains θ must be at least $1-\alpha$, i.e.,

$$\Pr_\theta\{\theta \in I_\tau\} \geq 1-\alpha \quad \text{for all } \theta. \tag{16}$$

This property shows that an RCI can be used to summarize information about θ on termination and the confidence level $1-\alpha$ will be maintained, regardless of how the decision to stop the study was reached. In contrast, the methods of Section 6.2 for constructing confidence intervals on termination rely on a particular stopping rule being specified at the outset and strictly enforced.

When a study is monitored using RCIs, the intervals computed at interim analyses might also be reported at scientific meetings. The basic property (15) ensures that these interim results will not be "over-interpreted". Here, over-interpretation refers to the fact that, when the selection bias or optional sampling bias of reported results is ignored, data may seem more significant than

warranted, and this can lead to adverse effects on accrual and drop-out rates, and to pressure to unblind or terminate a study prematurely.

Repeated P-values are defined analogously to repeated confidence intervals. At the kth analysis, a two-sided repeated P-value for H_0: $\theta = \theta_0$ is defined as $P_k = \max\{\alpha\colon\ \theta_0 \in I_k(\alpha)\}$, where $I_k(\alpha)$ is the current $(1-\alpha)$-level RCI. In other words, P_k is that value of α for which the kth $(1-\alpha)$-level RCI contains the null value, θ_0, as one of its endpoints. The construction ensures that, for any $p \in (0,\ 1)$, the overall probability under H_0 of ever seeing a repeated P-value less than or equal to p is no more than p and this probability is exactly p if all repeated P-values are always observed. Thus, the repeated P-value can be reported with the usual interpretation, yet with protection against the multiple-looks effect.

The repeated confidence intervals and P-values defined in this section should not be confused with the CIs and P-values discussed in Sections 6.1 and 6.2, which are valid only at termination of a sequential test conducted according to a strictly enforced stopping rule. Monitoring a study using repeated confidence intervals and repeated P-values allows flexibility in making decisions about stopping a trial at an interim analysis. These methodologies can, therefore, be seen as precursors to more recent adaptive methods designed, also motivated by the desire for greater flexibility in monitoring clinical trials.

7 Optimal Group Sequential Procedures

7.1 Optimizing within classes of group sequential procedures

We have described a variety of group sequential designs for one-sided and two-sided tests with early stopping to reject or accept the null hypothesis. Some tests have been defined through parametric descriptions of their boundaries, others through error spending functions. Since a key aim of interim monitoring is to terminate a study as soon as is reasonably possible, particularly under certain values of the treatment difference, it is of interest to find tests with optimal early stopping properties. These designs may be applied directly or used as benchmarks to assess the efficiency of designs which are attractive for other reasons. In our later discussion of flexible "adaptive" group sequential designs, we shall see the importance of assessing efficiency in order to quantify a possible tradeoff between flexibility and efficiency.

In formulating a group sequential design, we first specify the hypotheses of the testing problem and the Type I error rate α and power $1-\beta$ at $\theta = \delta$. Let $\mathcal{I}_f$ denote the information needed by the

fixed sample test, i.e., $\mathcal{I}_{f,2}$ as given by (4) for a two-sided test with error probabilities α and β, or $\mathcal{I}_{f,1}$ as given by (8) for a one-sided test with error probability constraints (6) and (7). We specify the maximum number K of possible analyses and the maximum information that may be required $\mathcal{I}_{max} = R\mathcal{I}_f$, where R is 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 look within the specified family of GSPs for that one which minimizes the average information on termination $E(\mathcal{I}_T)$ either at one θ value or averaged over several θ values.

To find the optimum procedure for a given sequence of information levels $\{\mathcal{I}_k\}$, we must search for boundary values $\{c_k\}$ for a two-sided test or $\{(a_k, b_k)\}$ for a one-sided test that minimize the average expected sample size criterion subject to the error probability constraints. This involves searching in a high dimensional space. Rather than search this space directly, we create a related sequential Bayes decision problem with a prior on θ, sampling costs, and costs for a wrong decision. The solution for such a problem can be found by a backward induction (dynamic programming) technique. Then, a two-dimensional search over cost parameters leads to a Bayes problem whose solution is the optimal GSP with error rates equal to the values α and β being sought. This is essentially a Lagrangian method for solving a constrained optimization problem; see Eales and Jennison (1992, 1995) and Barber and Jennison (2002) for more details.

7.2 Optimizing with equally spaced information levels

Let us consider one-sided tests with $\alpha = 0.025$, power $1 - \beta = 0.9$, $\mathcal{I}_{max} = R\mathcal{I}_{f,1}$, and K analyses at equally spaced information levels $\mathcal{I}_k = (k/K)\,\mathcal{I}_{max}$. For our optimality criterion, here we shall take $\int f(\theta) E_\theta(\mathcal{I}_T)\,\mathrm{d}\theta$, where $f(\theta)$ is the density of a $N(\delta,\, \delta^2/4)$ distribution and $\mathcal{I}_T$ denotes the information level on termination. This average expected information is centered on θ values around $\theta = \delta$ but puts significant weight over the range 0 to $2\,\delta$, encompassing both the null hypothesis and effect sizes well in excess of the value δ at which power is set. This is a suitable criterion when δ is a minimal clinically significant effect size and investigators are hoping the true effect is larger than this.

[Table 5 about here.]

Table 5 shows the minimum expected value of $\int f(\theta) E_\theta(\mathcal{I}_T)\,\mathrm{d}\theta$ for various combinations of K and R. These values are stated as percentages of the required fixed sample information $\mathcal{I}_{f,1}$ and as such are invariant to the value of the effect size δ. For fixed R, it can be seen that the average $E(\mathcal{I}_T)$ decreases as K increases, but with diminishing returns. For fixed K, the average $E(\mathcal{I}_T)$ decreases as

R increases up to a point, R^* say. For values of $R > R^*$, the larger increments in information (group sizes) implicit in the definition $\mathcal{I}_k = (k/K)R\mathcal{I}_{f,1}$, are sub-optimal. It is evident that including just a single interim analysis ($K = 2$) can significantly reduce $E(\mathcal{I}_T)$. If the resources are available to conduct more frequent analyses, we would recommend taking $K = 4$ or 5 and $R = 1.1$ or 1.2 to obtain most of the possible reductions expected sample size offered by group sequential testing.

We can use our optimal tests to assess parametric families of group sequential tests that have been proposed for this one-sided testing problem. The assessment is done by comparing the criterion $\int f(\theta)E_\theta(\mathcal{I}_T)\,\mathrm{d}\theta$ for each test against that for the corresponding optimal procedure. We consider three families of tests:

A. In Section 5 we introduced the ρ family of error spending tests with Type I and II error spending functions $f(x) = \min\{\alpha, \alpha(x/\mathcal{I}_{max})^\rho\}$ and $g(x) = \min\{\beta, \beta(x/\mathcal{I}_{max})^\rho\}$, respectively. For given $\mathcal{I}_{max}$, the requirement that the upper and lower decision boundaries of a one-sided test meet at $\mathcal{I}_{max}$ determines the value of ρ and *vice versa*. Since $\mathcal{I}_{max} = R\mathcal{I}_{f,1}$ the inflation factor R is also determined by ρ.

B. Hwang et al. (1990) proposed another family of error spending tests in which cumulative error spent is proportional to $(1-e^{-\gamma\,\mathcal{I}_k/\mathcal{I}_{max}})/(1-e^{-\gamma})$ instead of $(\mathcal{I}_k/\mathcal{I}_{max})^\rho$ in the ρ family defined in (A). In this case, the parameter γ determines the inflation factor R and *vice versa*.

C. Pampallona and Tsiatis (1994) proposed a parametric family for monitoring successive values of Z_k. This family is indexed by a parameter Δ and the boundaries for Z_k involve $\mathcal{I}_k^{\Delta-1/2}$. The parameter Δ determines the inflation factor R and *vice versa*.

Figure 5 shows values of $\int f(\theta)E_\theta(\mathcal{I}_T)\,\mathrm{d}\theta$ plotted against R for these three families of tests for the case of $K = 5$ equally sized groups, $\alpha = 0.05$ and $1 - \beta = 0.9$. The fourth and lowest curve is the minimum possible average $E_\theta(\mathcal{I}_T)$ for each value of R, obtained by our optimal tests.

[Figure 5 about here.]

It can be seen that both error spending families are highly efficient but the Pampallona and Tsiatis (1994) tests are noticeably sub-optimal.

7.3 Optimizing over information levels

We can extend the computations of the previous section to permit the optimal choice of cumulative information levels $\mathcal{I}_1, \ldots, \mathcal{I}_K$ with $\mathcal{I}_K \leq R\mathcal{I}_f$, as well as optimizing over the decision boundary

values $\{(a_k, b_k)\}$. In particular, allowing the initial information level $\mathcal{I}_1$ to be small may be advantageous if it is important to stop very early when there is a large treatment benefit — the "home run" treatment. We still use dynamic programming to optimize for a given sequence $\mathcal{I}_1, \ldots, \mathcal{I}_K$, but add a further search over these information levels by, say, the Nelder and Mead (1965) algorithm applied to a suitable transform of $\mathcal{I}_1, \ldots, \mathcal{I}_K$.

Allowing a free choice of the sequence of information levels enlarges the class of GSPs being considered, resulting in more efficient designs. We shall see in the next section that there are tangible benefits from this approach, particularly for $K = 2$.

Although we consider arbitrary sequences $\mathcal{I}_1, \ldots, \mathcal{I}_K$, these information levels and the boundary values (a_k, b_k), $k = 1, \ldots, K$, are still set at the start of the study and cannot be updated as observations accrue. Relaxing this requirement leads to a further enlargement of the candidate procedures which we discuss in the next section.

7.4 Procedures with data dependent increments in information

The option of scheduling each future analysis in a response-dependent manner has some intuitive appeal. For example, it would seem reasonable to choose smaller group sizes when the current test statistic lies close to a stopping boundary and larger group sizes when well away from a boundary. Schmitz (1993) refers to such designs as "sequentially planned decision procedures". Here, 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 the currently available data. The whole procedure can be designed to optimize an efficiency criterion subject to the upper limit $\mathcal{I}_K \leq R\mathcal{I}_f$. There is an uncountable number of decision variables to be optimized as one defines $\mathcal{I}_{k+1}(z_k)$, $a_{k+1}(z_k)$ and $b_{k+1}(z_k)$ for each value of $\mathcal{I}_k$ and *every* z_k in the continuation region $\mathcal{C}_k = (a_k, b_k)$. However, by means of discretization of the $\mathcal{I}_k$ scale, the dynamic programming optimization computation, though still formidable, can be carried out. Note that, while the Schmitz designs are adaptive in the sense that future information levels are allowed to depend on current data, these designs are not "flexible". The way in which future information levels are chosen, based on past and current information levels and Z-values, is specified at the start of the study — unlike the flexible, adaptive procedures we shall discuss in Section 8.

The question arises as to how much extra efficiency can be obtained by allowing unequal but pre-specified information levels (Section 7.3) or, further, allowing these information levels to be data dependent (Schmitz, 1993). Jennison and Turnbull (2006a) compare families of one-sided

tests of H_0: $\theta = 0$ versus H_1: $\theta > 0$ with $\alpha = 0.025$ and power $1 - \beta = 0.9$ at $\theta = \delta$. They use the same efficiency criterion $\int f(\theta) E_\theta(\mathcal{I}_T)\,\mathrm{d}\theta$ we have considered previously, subject to the constraint on maximum information $\mathcal{I}_K \leq R\,\mathcal{I}_f$. We can define three nested classes of GSPs:

1. GSPs with equally spaced information levels,

2. GSPs permitting unequally spaced but fixed information levels,

3. GSPs permitting data dependent increments in information according to a pre-specified rule.

Table 6, which reports cases in Table 1 of Jennison and Turnbull (2006a) with $R = 1.2$, shows the optimal values of the efficiency criterion for these three classes of GSPs as a percentage of the fixed sample information for values of $K = 1$ to 6, 8, and 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 6 about here.]

Although we have focused on a single efficiency criterion $\int f(\theta) E_\theta(\mathcal{I}_T)\,\mathrm{d}\theta$, the same methods can be applied to optimize with respect to other criteria, such as $E_\theta(\mathcal{I}_T)$ at a single value of θ or averaged over several values of θ. Results for other criteria presented in Eales and Jennison (1992, 1995) and Barber and Jennison (2002) show qualitatively similar features to those we have reported here. Optimality criterion can also be defined to reflect both the cost of sampling and the economic utility of a decision and the time at which it occurs; see Liu, Anderson, and Pledger (2004).

8 Tests Permitting Flexible, Data Dependent Increments in Information

8.1 Flexible re-design protecting the Type I error probability

In the GSPs of Section 7.4, the future information increments (group sizes) and critical values are permitted to depend on current and past values of the test statistic Z_k, assuming of course that the stopping boundary had not been crossed. These procedures are not flexible in that the rules for making the choices are pre-specified functions of currently available Z_k values. With this knowledge, it is possible *ab initio* to compute a procedure's statistical properties, such as Type I and II error probabilities and expected information at termination. However, what can be done if an unexpected event happens in mid-course and we wish to make some *ad hoc* change in the

future information increment levels? This is often referred to as flexible *sample size re-estimation* or *sample size modification*.

Consider the application of a classical group sequential one-sided design. The trial is under way and, based on data observed at analysis j, it is desired to increase future information levels. If we were to do this and continue to use the original values (a_k, b_k) for $k > j$ in the stopping rule (9), the Type I error rate would no longer be guaranteed at α. If arbitrary changes in sample size are allowed, the Type I error rate is typically inflated — see Cui et al. (1999, Table A1) and Proschan and Hunsberger (1995). (However, as an exception, note that if it is *preplanned* that increases in sample size are only permitted when the interim treatment estimate is sufficiently high (conditional power greater than 0.5), this implies that the actual overall Type I error rate may be reduced — Chen, DeMets and Lan (2004).)

Suppose, however, that we do go ahead with this adaptation and the cumulative information levels are now $\tilde{\mathcal{I}}^{(1)}, \ldots, \tilde{\mathcal{I}}^{(K)}$; here, $\tilde{\mathcal{I}}^{(k)} = \mathcal{I}_k$ for $k \leq j$ but the $\tilde{\mathcal{I}}^{(k)}$ differ from the originally planned $\mathcal{I}_k$ for $k > j$. Let $\tilde{Z}^{(k)}$ be the usual Z-statistic *formed from data in stage k alone* and $\tilde{\Delta}_k = \tilde{\mathcal{I}}^{(k)} - \tilde{\mathcal{I}}^{(k-1)}$. Again, the $\tilde{\Delta}_k$ are as originally planned for $k \leq j$ but they depart from this plan for $k > j$. We *can* still maintain the Type I error probability using the original boundary if we use the statistics $\tilde{Z}^{(k)}$ in the appropriate way. Note that, even though the information increment $\tilde{\Delta}^{(k)}$ is an ingredient of the statistic $\tilde{Z}^{(k)}$ and, for $k > j$, this can depend on knowledge of the previously observed $\tilde{Z}^{(1)}, \ldots, \tilde{Z}^{(k-1)}$, each $\tilde{Z}^{(k)}$ has a standard normal $N(0,1)$ distribution under $\theta = 0$ conditionally on the previous responses and $\tilde{\Delta}^{(k)}$. It follows that this distribution holds unconditionally under H_0, so we may treat $\tilde{Z}^{(1)}$, $\tilde{Z}^{(2)}$, ... as independent $N(0,1)$ variables. The standard distribution of the $\{\tilde{Z}^{(k)}\}$ under H_0 means we can use the original boundary values in (9) and maintain the specified Type I error rate α, provided we monitor the statistics

$$\tilde{Z}_k \;=\; (w_1\tilde{Z}^{(1)} + \ \ldots\ + w_k\tilde{Z}^{(k)}) \,/\, (w_1^2 + \ \ldots\ + w_k^2)^{1/2}, \quad k = 1, \ldots, K, \tag{17}$$

where the weights $w_k = \sqrt{\Delta_k}$ are the square roots of the *originally planned* information increments. With this definition, the $\tilde{Z}_k$ follow the canonical joint distribution (1) under H_0 that was originally anticipated; see Lehmacher and Wassmer (1999) or Cui et al. (1999). Under the alternative $\theta > 0$, the $\tilde{Z}^{(k)}$ are not independent after adaptation and if information levels are increased, then so are the means of the Z-statistics — which leads to the desired increase in power.

Use of a procedure based on (17) is an example of a *combination test*. In particular (17) is a *weighted inverse normal combination* statistic (Mosteller and Bush, 1954). Other combination test statistics can be used in place of (17), such as the inverse χ^2 statistic proposed by Bauer and

Köhne (1994). However, use of (17) has two advantages: (i) we do not need to recalculate the stopping boundaries $\{(a_k, b_k)\}$, and (ii) if no adaptation occurs, we have $\tilde{Z}_k = Z_k$, $k = 1, 2, \ldots$, and the procedure proceeds as originally planned.

8.2 Efficiency of flexible adaptive procedures

We have seen in Section 8.1 how, by using (17), the investigator has the freedom to modify a study in light of accruing data and still maintain the Type I error rate. 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) discuss the example of a GSP with $K = 5$ analyses testing H_0: $\theta \leq 0$ 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 $\mathcal{I}_f = \mathcal{I}_{f,1}$, as given by (8). Suppose the study is designed as a one-sided test from the ρ-family of error-spending tests, as described in Section 5, and we choose index $\rho = 3$. The boundary values $a_1, \ldots, a_5$ and $b_1, \ldots, b_5$ are chosen to satisfy

$$\Pr_\theta\{Z_1 > b_1 \text{ or } \ldots \text{ or } Z_1 \in (a_1, b_1), \ldots, Z_{k-1} \in (a_{k-1}, b_{k-1}), Z_k > b_k\} = (\mathcal{I}_k/\mathcal{I}_{max})^3\, \alpha,$$

$$\Pr_\theta\{Z_1 < a_1 \text{ or } \ldots \text{ or } Z_1 \in (a_1, b_1), \ldots, Z_{k-1} \in (a_{k-1}, b_{k-1}), Z_k < a_k\} = (\mathcal{I}_k/\mathcal{I}_{max})^3\, \beta$$

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 similar in shape to those in Figure 2.

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 strategy described in Cui et al. (1999), 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 and 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 6 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 Figure 6, and the expected information curves of the two tests are of a similar form. Figure 7 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 6 and 7 demonstrate 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 $4.20\,\mathcal{I}_f$.

[Figure 6 about here.]

[Figure 7 about here.]

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 $\theta > 0$, Jennison and Turnbull (2006a) 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. The plot of the efficiency ratio in Figure 8 shows the adaptive design is considerably less efficient than the simple group sequential test, especially for $\theta > \delta/2$, and this quantifies the cost of delay in learning the study's objective.

Another motivation for sample size modification is the desire to increase sample size on seeing low interim estimates of the treatment effect. Investigators may suppose the true treatment effect is perhaps smaller than they had hoped and aim to increase, belatedly, the power of their study.

Or they may hope that adding more data will make amends for an "unlucky start". We have studied such adaptations in response to low interim estimates of the treatment effect and found inefficiencies similar to, or worse than, those in the preceding example. The second example in Jennison and Turnbull (2006a) concerns such adaptation using the Cui et al. (1999) procedure. We have found comparable inefficiencies when sample size is modified to achieve a given conditional power using the methods of Bauer and Köhne (1994), Proschan and Hunsberger (1995), Shen and Fisher (1999) and Li et al. (2002). When adaptation is limited to smaller increases in sample size, the increase in power is smaller but efficiency loss is still present.

[Figure 8 about here.]

We saw in Section 7.4 that the pre-planned adaptive designs of Schmitz (1993) can be slightly more efficient than conventional group sequential tests. One must, therefore, wonder why the adaptive tests that we have studied should be less efficient than competing group sequential tests, sometimes by as much as 30 or 40%. We can cite three contributory factors:

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. Adaptive procedures using combination test statistics (17) with their unequal weighting of observations are not based on sufficient statistics. Thus, they cannot be optimal designs for any criteria. Since the potential benefits of adaptivity are slight, 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 arising in the optimal adaptive designs of Section 7.4. 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. The qualitative differences between these two types of procedure are illustrated by the typical shapes of sample size functions shown in Figures 9 and 10.

[Figure 9 about here.]

[Figure 10 about here.]

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, often as an assumed value in a conditional power calculation. Since this estimator is highly variable, use of this estimate leads to random variation in sample size which is in itself inefficient; see Jennison and Turnbull (2003) for further discussion of this point in the context of a two-stage design.

Our conclusion in this section is that group sequential tests provide an efficient and versatile mechanism for conducting clinical trials, but it can be useful to have adaptive methods to turn to when a study's sample size is found to be too small. Our first example depicts a situation where a change in objective could not have been anticipated at the outset and an adaptive solution is the only option. While good practice at the design stage should ensure that a study has adequate power, it is reassuring to know there are procedures available to rescue an under-powered study while still protecting the Type I error rate. What we do not recommend is use of such adaptive strategies as a substitute for proper planning. Investigators may have different views on the likely treatment effect, but it is still possible to construct a group sequential design that will deliver the desired overall power with low expected sample size under the effect sizes of most interest; for further discussion of how to balance these objectives, see Schäfer and Müller (2004) and Jennison and Turnbull (2006b).

9 Discussion

In Sections 2 to 6 we described the classical framework in which group sequential tests are set, presented an overview of GSPs defined by parametric boundaries or error spending functions, and discussed inference on termination of a GSP. These classical GSPs are well studied; optimal tests have been derived for a variety of criteria and error spending functions identified which give tests with close to optimal performance.

GSPs adapt to observed data in the most fundamental way by terminating the study when a decision boundary is crossed. Error spending designs have the flexibility to accommodate unpredictable information sequences. In cases where information depends on nuisance parameters that affect the variance of the outcome variable, Mehta and Tsiatis (2001) propose "information monitoring" designs in which updated estimates of nuisance parameters are incorporated in error

spending tests. Overall, classical group sequential methodology is versatile and can handle a number of the problems which more recent adaptive methods have been constructed to solve.

A question which poses problems for both group sequential and adaptive methods is how to deal with delayed responses which arrive after termination of a study. Stopping rules are usually defined on the assumption that no more responses will be observed after the decision to terminate, but it is not uncommon for such data to accrue, particularly when there is a significant delay between treatment and the time the primary response is measured. Group sequential methods that can handle such delayed data and methods for creating designs which do this efficiently are described by Hampson (2009).

We discussed in Sections 7.4 and 8 how data dependent modification of group sizes can be viewed as a feature of both classical GSPs and adaptive designs. It is our view that the benefits of such modifications are small compared to the complexity of these designs. There is also a danger that interpretability may be compromised, indeed, Burman and Sonesson (2006) give an example where adaptive re-design leads to a complete loss of credibility.

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, adaptations can be made following the methods described in Section 8 or, more generally, by maintaining the conditional Type I error probability, as suggested by Denne(2001) and Müller and 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.

It should be clear from our exposition that group sequential and adaptive methods involve significant computation. Fortunately, there is a growing number of computer software packages available for implementing these methods to design and monitor clinical trials. Self contained programs include

EAST (`http://www.cytel.com/Products/East/`),

ADDPLAN (`http://www.addplan.com/`),

PEST (`http://www.maths.lancs.ac.uk/department/research/statistics/mps/pest`),

NCSS/PASS (`http://www.ncss.com/passsequence.html`), and

ExpDesign Studio (Chang, 2008).

Several useful macros written for SAS are detailed in Dmitrienko et al. (2005, Chap. 4). The add-on module S+SeqTrial (`http://www.splus.com/products/seqtrial/`) is available for use with S-PLUS.

A number of websites offer software that can be freely downloaded. The `gsDesign` package (`http://cran.r-project.org/`) is one of several packages for use with R. The website `http://www.biostat.wisc.edu/landemets/` contains FORTRAN programs for error spending procedures. Our own FORTRAN programs, related to the book JT, are available at `http://people.bath.ac.uk/mascj/book/programs/general`. For a review of the capabilities of all these software packages, we refer the reader to the article by Wassmer and Vandemeulebroecke (2006).

Our comments on adaptive design in this chapter relate to sample size modification as this is the key area of overlap with GSPs. Adaptive methods do, of course, have a wide range of further applications — as the other chapters in this book demonstrate.

References:

Armitage, P. (1960). *Sequential Medical Trials, 1st Ed.* Springfield: Thomas.

Armitage, P., McPherson, C.K. and Rowe, B.C. (1969). Repeated significance tests on accumulating data. *J. Royal Statistical Society, A* **132**, 235–244.

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.

Chang, M. (2008). *Classical and Adaptive Clinical Trial Designs Using ExpDesign Studio.* New York: Wiley.

Chen, J.Y.H., DeMets, D.L. and Lan, K.K.G. (2004). Increasing the sample size when the unblinded interim result is promising. *Statistics in Medicine*, **23**(7), 1023-1038.

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

DeMets, D.L., Furberg, C.D. and Friedman, L.M., Eds. (2006). *Data Monitoring in Clinical Trials.* New York: Springer.

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

Dmitrienko, A., Molenberghs, G., Chuang-Stein, C. and Offen, W. (2005). *Analysis of Clinical Trials Using SAS: A Practical Guide*. Cary: SAS Institute Press.

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

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

Ellenberg, S.S., Fleming, T.R. and DeMets, D.L. (2002). *Data Monitoring Committees in Clinical Trials: A Practical Perspective*. West Sussex: Wiley.

Emerson, S.S. and Fleming, T.R. (1990). Parameter estimation following group sequential hypothesis testing. *Biometrika*, **77**, 875–892.

Hampson, L.V. (2009). *Group Sequential Tests for Delayed Responses.* PhD thesis, University of Bath.

Herson, J. (2009). *Data and Safety Monitoring Committees in Clinical Trials.* Boca Raton: Chapman & Hall/CRC.

Hwang, I.K., Shih, W.J. and DeCani, J.S. (1990). Group sequential designs using a family of type I error probability spending functions. *Statistics in Medicine* **9**, 1439–1445.

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

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

Jennison, C. and Turnbull, B.W. (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 nonadaptive 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**, 917–932.

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

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. (2002). A sample size adjustment procedure for clinical trials based on conditional power. *Biostatistics* **3**, 277–287.

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

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

Mosteller, F. and Bush, R.R. (1954). Selected quantitative techniques. In *Handbook of Social Psychology, Vol.1.* Ed. G. Lindsey. Cambridge: Addison-Wesley, pages 289–334.

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.

Nelder, J.A. and Mead, R. (1965). A simplex method for function minimization. *Computer Journal* **7**, 308–313.

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

Pampallona, S. and Tsiatis, A.A. (1994). Group sequential designs for one-sided and two-sided hypothesis testing with provision for early stopping in favor of the null hypothesis. *J. Statistical Planning and Inference* **42**, 19–35.

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

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

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

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.

Siegmund, D. (1985). *Sequential Analysis.* New York: Springer-Verlag.

Tsiatis, A.A. and Mehta, C. (2003). On the inefficiency of the adaptive design for monitoring

clinical trials. *Biometrika* **90**, 367–378.

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

Wang, S.K. and Tsiatis, A.A. (1987). Approximately optimal one-parameter boundaries for group sequential trials. *Biometrics* **43**, 193–200.

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

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

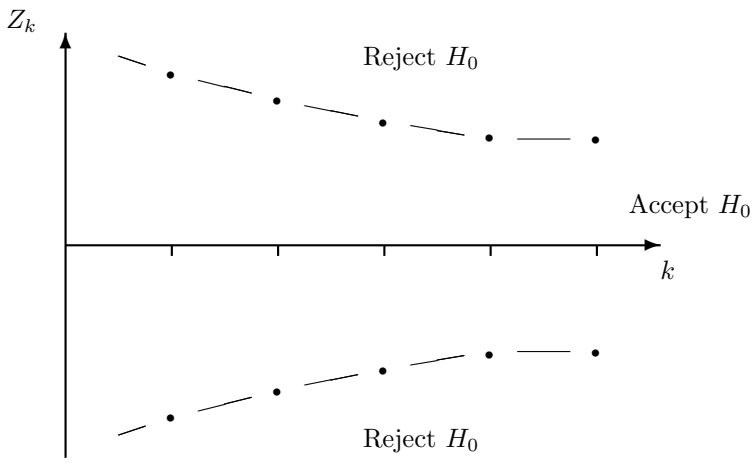

Figure 1: Two-sided decision boundary for $K = 5$.

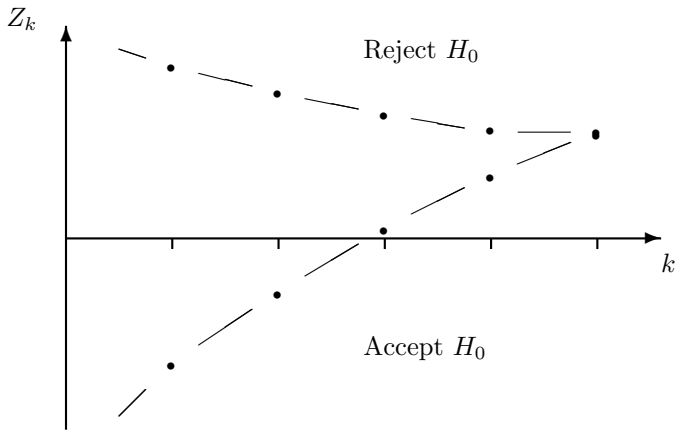

Figure 2: Decision boundary for a one-sided test with $K = 5$.

Figure 3: A typical error spending function.

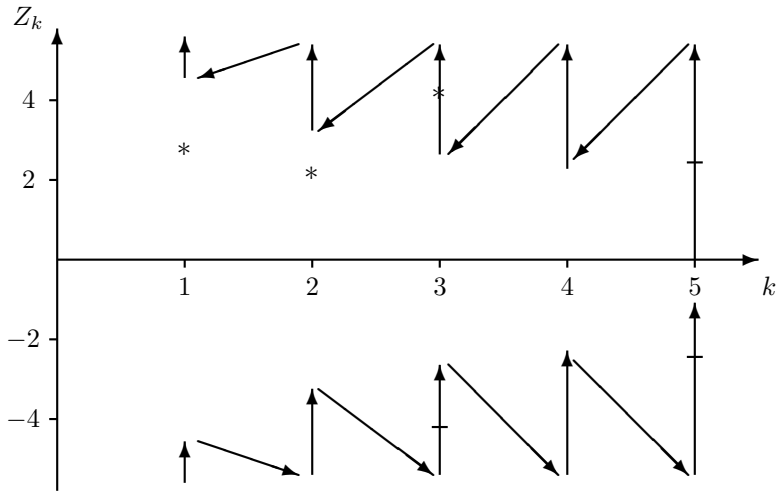

Figure 4: Stagewise ordering for an O'Brien and Fleming design with 5 analyses and $\alpha = 0.05$.

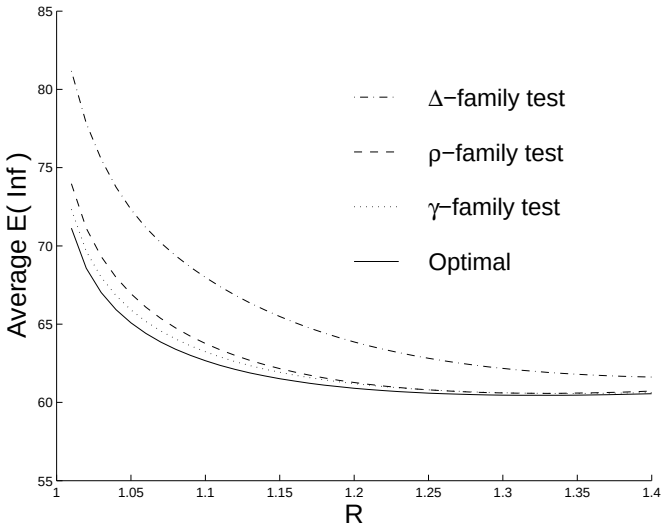

Figure 5: $\int f(\theta) E_\theta(\mathcal{I}_T)\,\mathrm{d}\theta$ as a percentage of $\mathcal{I}_{f,1}$ plotted against the inflation factor R for four families of tests with $K = 5$ equally spaced looks, $\alpha = 0.025$ and $\beta = 0.1$.

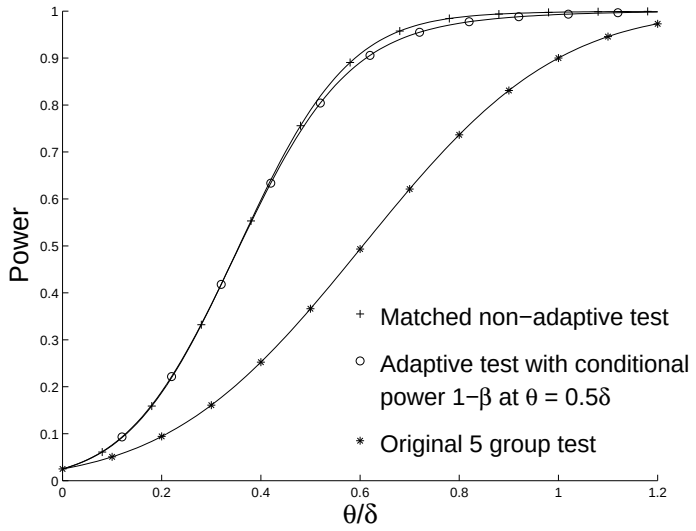

Figure 6: Power curves of the original test, the adaptive design with sample size revised at look 2 to attain conditional power 0.9 at $\theta = 0.5\,\delta$, and the matched non-adaptive test.

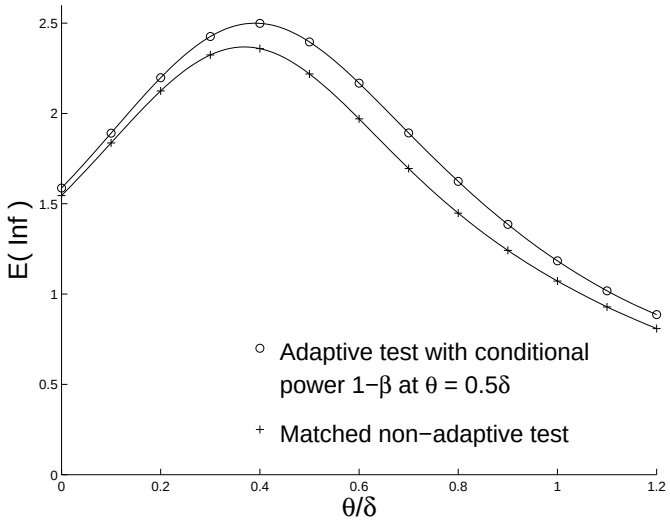

Figure 7: Expected information functions $E_\theta(\mathcal{I}_T)$ of the adaptive design and matched non-adaptive design, expressed in units of $\mathcal{I}_f$.

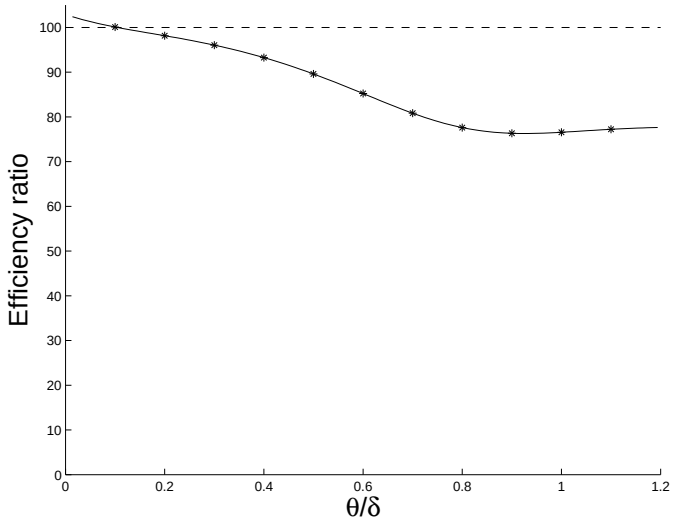

Figure 8: Efficiency ratio between the adaptive design and matched non-adaptive design.

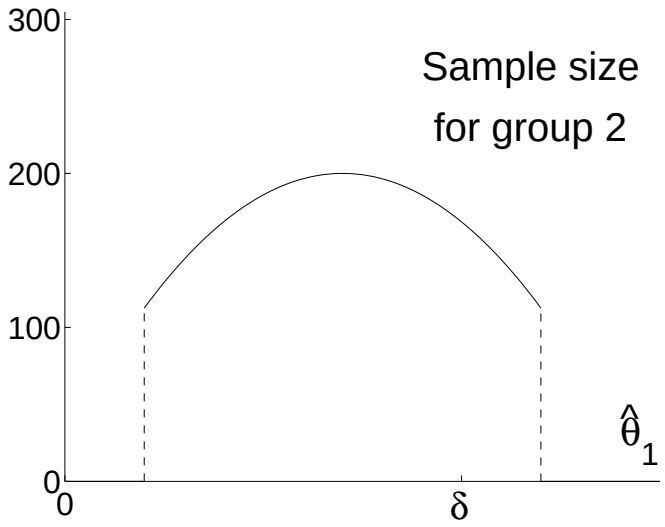

Figure 9: Typical shape of sample size function for an optimal adaptive test.

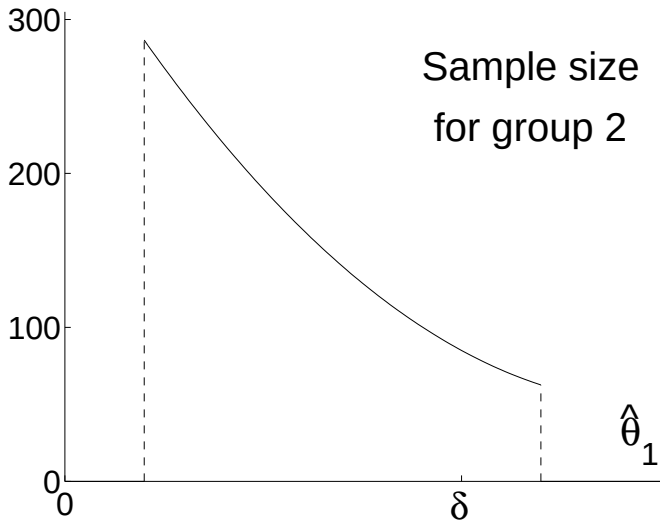

Figure 10: Typical shape of sample size function for a conditional power adaptive design.

Table 1: Constants $C_P(K, \alpha)$ for Pocock two-sided tests with K groups of observations and type I error probability α

	$C_P(K, \alpha)$		
K	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$
1	2.576	1.960	1.645
2	2.772	2.178	1.875
3	2.873	2.289	1.992
4	2.939	2.361	2.067
5	2.986	2.413	2.122
6	3.023	2.453	2.164
7	3.053	2.485	2.197
8	3.078	2.512	2.225
9	3.099	2.535	2.249
10	3.117	2.555	2.270
11	3.133	2.572	2.288
12	3.147	2.588	2.304
15	3.182	2.626	2.344
20	3.225	2.672	2.392

Table 2: Constants $C_B(K, \alpha)$ for O'Brien & Fleming two-sided tests with K groups of observations and type I error probability α

	$C_B(K, \alpha)$		
K	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$
1	2.576	1.960	1.645
2	2.580	1.977	1.678
3	2.595	2.004	1.710
4	2.609	2.024	1.733
5	2.621	2.040	1.751
6	2.631	2.053	1.765
7	2.640	2.063	1.776
8	2.648	2.072	1.786
9	2.654	2.080	1.794
10	2.660	2.087	1.801
11	2.665	2.092	1.807
12	2.670	2.098	1.813
15	2.681	2.110	1.826
20	2.695	2.126	1.842

Table 3: Constants $R_P(K, \alpha, \beta)$ to determine group sizes for Pocock two-sided tests with K groups of observations, type I error probability α and power $1 - \beta$

	$R_P(K, \alpha, \beta)$					
	$1 - \beta = 0.8$			$1 - \beta = 0.9$		
K	α=0.01	α=0.05	α=0.10	α=0.01	α=0.05	α=0.10
1	1.000	1.000	1.000	1.000	1.000	1.000
2	1.092	1.110	1.121	1.084	1.100	1.110
3	1.137	1.166	1.184	1.125	1.151	1.166
4	1.166	1.202	1.224	1.152	1.183	1.202
5	1.187	1.229	1.254	1.170	1.207	1.228
6	1.203	1.249	1.277	1.185	1.225	1.249
7	1.216	1.265	1.296	1.197	1.239	1.266
8	1.226	1.279	1.311	1.206	1.252	1.280
9	1.236	1.291	1.325	1.215	1.262	1.292
10	1.243	1.301	1.337	1.222	1.271	1.302
11	1.250	1.310	1.348	1.228	1.279	1.312
12	1.257	1.318	1.357	1.234	1.287	1.320
15	1.272	1.338	1.381	1.248	1.305	1.341
20	1.291	1.363	1.411	1.264	1.327	1.367

Table 4: Constants $R_B(K, \alpha, \beta)$ to determine group sizes for O'Brien & Fleming two-sided tests with K groups of observations, type I error probability α and power $1 - \beta$

	$R_P(K, \alpha, \beta)$					
	$1 - \beta = 0.8$			$1 - \beta = 0.9$		
K	α=0.01	α=0.05	α=0.10	α=0.01	α=0.05	α=0.10
1	1.000	1.000	1.000	1.000	1.000	1.000
2	1.001	1.008	1.016	1.001	1.007	1.014
3	1.007	1.017	1.027	1.006	1.016	1.025
4	1.011	1.024	1.035	1.010	1.022	1.032
5	1.015	1.028	1.040	1.014	1.026	1.037
6	1.017	1.032	1.044	1.016	1.030	1.041
7	1.019	1.035	1.047	1.018	1.032	1.044
8	1.021	1.037	1.049	1.020	1.034	1.046
9	1.022	1.038	1.051	1.021	1.036	1.048
10	1.024	1.040	1.053	1.022	1.037	1.049
11	1.025	1.041	1.054	1.023	1.039	1.051
12	1.026	1.042	1.055	1.024	1.040	1.052
15	1.028	1.045	1.058	1.026	1.042	1.054
20	1.030	1.047	1.061	1.029	1.045	1.057

Table 5: Minimum values of $\int f(\theta) E_\theta(\mathcal{I}_T)\,\mathrm{d}\theta$ expressed as a percentage of $\mathcal{I}_{f,1}$.

				R			
K	1.01	1.05	1.1	1.15	1.2	1.3	*minimum over R*
2	79.3	74.7	73.8	74.1	74.8	77.1	73.8 at $R = 1.1$
3	74.8	69.0	67.0	66.3	66.1	66.6	66.1 at $R = 1.2$
4	72.5	66.5	64.2	63.2	62.7	62.5	62.5 at $R = 1.3$
5	71.1	65.1	62.7	61.5	60.9	60.5	60.5 at $R = 1.3$
10	68.2	62.1	59.5	58.2	57.5	56.7	56.4 at $R = 1.5$
20	66.8	60.6	58.0	56.6	55.8	54.8	54.2 at $R = 1.6$

Table 6: Optimized $\int f(\theta) E_\theta(\mathcal{I}_T)\,\mathrm{d}\theta$ as a percentage of $\mathcal{I}_{f,1}$ for tests with inflation factor $R = 1.2$.

K	*1. Optimal GSP with K equal group sizes*	*2. Optimal GSP with K optimized group sizes*	*3. Optimal adaptive design of Schmitz*
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