COMPUTATIONAL ISSUES IN INFORMATION-BASED GROUP SEQUENTIAL CLINICAL TRIALS

KyungMann Kim*, Anastasios A. Tsiatis† and Cyrus R. Mehta‡

ABSTRACT

Lan and DeMets (Biometrika 1983; 70:659–663) introduced a flexible procedure for monitoring of group sequential clinical trials based on the discretization of the Brownian motion process. Subsequently Kim and DeMets (Biometrika 1987; 74:149–154) developed a general procedure for design of such clinical trials. A number of procedures have been proposed for statistical inference following group sequential tests regarding the P-values and the point and confidence interval estimation of the parameter of interest such as the effect size or the treatment difference in such clinical trials. In this article, computational issues are described for design and monitoring of clinical trials with interim analysis based on group sequential methods for possible early stopping for efficacy or safety and for inference following early stopping of group sequential clinical trials. The computational procedures as implemented in a commercial package EaSt (2000) are illustrated with an example of a lung cancer clinical trial.

1. Introduction

In many long-term chronic disease clinical trials, the primary outcome of interest is either time to an event such as death with possible right censoring or repeated measures taken at successive follow-ups with possibly missing data. Typically patients enter clinical trials serially in a way known as staggered entry, and the final analysis is conducted either after a prespecified follow-up period or after a prespecified number of events or follow-up visits.

For ethical as well as practical reasons these trials are monitored more or less regularly, and the trial may be considered for early stopping if a sufficiently large treatment difference emerges during an interim analysis. When monitoring such clinical trials, multiplicity due to repeated significance testing has to be properly accounted for in order to control the overall type I error probability at a desired level. In order to preserve the operating characteristics of a statistical test applied repeatedly over time, it is necessary to understand the joint distribution of the test statistics at different interim times.

Often this joint distribution turns out to be multivariate normal or at least asymptotically so, and subsequently the group sequential methods require multivariate numerical integration. The MULNOR program by Schervish (1984) can be used to this end, but it involves very intensive numerical computations and can handle multivariate integrations of only up to seven dimensions.

*Department of Biostatistics & Medical Informatics, University of Wisconsin Medical School, 600 Highland Ave, K6/438 CSC, Madison, WI 53792-4675, U.S.A.
†Department of Statistics, North Carolina State University, P.O. Box 8203, Raleigh, NC 27695, U.S.A.
‡Cytel Software Corporation, 675 Massachusetts Avenue, Cambridge, MA 02139, U.S.A.

Key words: Error spending function; Group sequential methods; Inflation factor; Interim analysis
If the increments between sequentially computed test statistics are independent, how-
however, the multivariate numerical integration reduces to univariate numerical integrations
involving a simple recursion as described in Armitage, McPherson and Rowe (1969) and
McPherson and Armitage (1971). Moreover, this allows the use of standard group sequential
methods such as by Pocock (1977), O’Brien and Fleming (1979) and Lan and DeMets
(1983) for design, monitoring and analysis of group sequential clinical trials.

In this article, we describe computational issues in design and monitoring of clinical
trials with interim analysis based on group sequential methods for possible early stopping
for efficacy or safety and in inference following early stopping of group sequential clinical
trials. We review group sequential methods for interim analysis of clinical trials for possible
early stopping for efficacy or safety in Section 2 and inference following early stopping based
on group sequential tests in Section 3. We describe the independent increment structure
critical for application of group sequential tests in Section 4. We introduce the notion of
inflation factor in section 5 and of information-based design and monitoring in Section 6. We
describe computational issues in group sequential methods in Section 7 with regard to design
and monitoring of such clinical trials and inference following early stopping. Finally we give
illustration of the computational procedures as implemented in the commercial package
EaSt (2000) with an example of a lung cancer clinical trial in Section 8. We conclude with
a discussion in Section 9.

2. Group sequential methods

Suppose the response to treatment is normally distributed with means $\mu_A$ and $\mu_B$ for
treatments $A$ and $B$, respectively, with known variance $\sigma^2$ and that one is interested in
testing the null hypothesis of no treatment difference, $H_0 : \mu_A = \mu_B$, against the alternative
hypothesis $H_1 : \mu_A \neq \mu_B$ or equivalently

$$H_0 : \delta = 0 \text{ versus } H_1 : \delta \neq 0$$

where $\delta = \mu_A - \mu_B$. With the standardized test statistic

$$Z = \frac{\bar{X}_A - \bar{X}_B}{\sqrt{2\sigma^2/n}},$$

one would reject $H_0$ if $|Z| > z_{\alpha/2}$, the upper $\alpha/2$ quantile of the standard normal distribution.

Suppose one plans to look at the accumulating data for up to $K$ times after every $n$
subjects per treatment per analysis. Then, for the $k$th interim analysis, one would consider
the partial sum

$$S_k = \sum_{j=1}^{k} Y_j \sim N(\delta^* k, k)$$

where

$$Y_j = \frac{\bar{X}_{A_j} - \bar{X}_{B_j}}{\sqrt{2\sigma^2/n}} \sim N(\delta^*, 1)$$

with $\delta^* = \delta/\sqrt{2\sigma^2/n}$ and $\bar{X}_{A_j}$ and $\bar{X}_{B_j}$, the sample means of $n$ observations accumulated
between the $(j - 1)$st and the $j$th interim analyses. Or, equivalently, one would compute a
standardized test statistic

$$Z_k = \frac{\sum_{j=1}^{k} (\bar{X}_{A_j} - \bar{X}_{B_j})}{\sqrt{k(2\sigma^2/n)}} = \frac{S_k}{\sqrt{k}} \sim N(\delta^* \sqrt{k}, 1)$$
and decide to stop or to continue to the next interim analysis.

If one were to apply the fixed sample test repeatedly, the false positive rate, i.e., type I error probability, is known to become inflated beyond the desired level, ultimately becoming 1. This phenomenon was described as sampling to reach a foregone conclusion by Anscombe (1954). Table 1 summarizes the effect of multiple significance tests at a nominal level of \( \alpha = 0.05 \) and gives the actual type I error probabilities achieved.

<table>
<thead>
<tr>
<th>( K )</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>10</th>
<th>20</th>
<th>50</th>
<th>100</th>
<th>( \infty )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type I Error</td>
<td>0.05</td>
<td>0.083</td>
<td>0.107</td>
<td>0.126</td>
<td>0.142</td>
<td>0.193</td>
<td>0.246</td>
<td>0.320</td>
<td>0.374</td>
<td>1.000</td>
</tr>
</tbody>
</table>

This clearly indicates the need to adjust the criteria in terms of the critical values of the sequential tests. One way is to choose the critical values \( c_1, \ldots, c_K \) so that the overall significance level is \( \alpha \), that is, under the null hypothesis,

\[
\Pr(|Z_1| < c_1, \ldots, |Z_K| < c_K) = 1 - \alpha.
\]

Pocock (1977) suggested using the same critical values at each look and rejecting \( H_0 \) the first time when

\[
|Z_k| \geq c_k \equiv c_P \text{ or equivalently } |S_k| \geq b_k = c_P \sqrt{k},
\]

whereas O’Brien and Fleming (1979) suggested rejecting \( H_0 \) the first time when

\[
|Z_k| \geq c_k = c_O \sqrt{K/k} \text{ or equivalently } |S_k| \geq b_k = c_O \sqrt{K}.
\]

With O’Brien-Fleming procedures, it is hard to reject \( H_0 \) early because of larger critical values at earlier looks, with the final test becoming similar to a fixed sample test.

Wang and Tsiatis (1987) proposed a more general class of group sequential boundaries indexed by \( \Delta \) as follows:

\[
b_k = b(\alpha, K, \Delta)k^\Delta.
\]

This class of group sequential boundaries include as special cases those of O’Brien and Fleming (1979) when \( \Delta = 0 \) and of Pocock (1977) when \( \Delta = 0.5 \).

The standard group sequential methods described above require equal increments of information between interim analyses and a prespecified maximum number of repeated analyses \( K \). In practice, however, there is no guarantee that these two conditions will be met. Thus flexible methods are required. Slud and Wei (1982) suggest that one specify exit probabilities

\[
\pi_k = \Pr(|Z_1| < c_1, \ldots, |Z_{k-1}| < c_{k-1}, |Z_k| \geq c_k)
\]

such that \( \sum_{k=1}^{K} \pi_k = \alpha \), the overall significance level.

Lan and DeMets (1983) instead suggested specifying an error spending function, also known as, use function, to test at arbitrary times or even sporadically. It is defined as a monotonically increasing function, \( \alpha^*(\tau) \), for \( 0 \leq \tau \leq 1 \), such that \( \alpha^*(0) = 0 \) and \( \alpha^*(1) = \alpha \). Note that \( \alpha^*(\tau_k) - \alpha^*(\tau_{k-1}) = \pi_k \) specifies the exit probability above. In order to use the error spending function, the information fractions \( \tau_k \) have to be estimated. For example, the information fraction is defined as \( \tau_k = k/K \) for a two-sample group sequential test.
when equal samples of size $n$ are accumulated between successive interim analyses. With the unequal sample sizes between the $(k-1)$st and the $k$th analyses for $k = 1, \ldots, K$, the information fraction is defined as $\tau_k = n_k/n_K$ where $n_k$ denotes the cumulative sample size at the $k$th interim analysis. For the Pocock (1977) and the O'Brien and Fleming (1979) group sequential method, Lan and DeMets (1983) proposed the following error spending functions,

$$\alpha^*(\tau) = \alpha \log\{1 + (e - 1)\tau\}$$

and

$$\alpha^*(\tau) = 2\{1 - \Phi(z_{\alpha/2}/\sqrt{\tau})\}$$

for one-sided tests, respectively.

3. Inference following early stopping

A number of procedures have been proposed for statistical inference following early stopping regarding the observed significance of the group sequential test, i.e., the so-called $P$-values, and the point and confidence interval estimation of the parameter of interest such as the effect size or the treatment difference in such clinical trials.

One of the procedures based on the ordering of the sample space under the group sequential testing according to Armitage (1958), Fairbanks and Madsen (1981) and Tsiatis, Rosner and Mehta (1984) provides the framework in which $P$-values and point and confidence interval estimates can be determined in the most natural and coherent way. According to this ordering, an earlier stopping and, if early stopping occurs at the same time, a larger value of the test statistic is a more extreme evidence against the null hypothesis.

If the group sequential test terminates at the $k$th interim analysis with a terminal observed value of the test statistic $z^*$, the $P$-value following early stopping based on a group sequential test is defined as follows:

$$P = \sum_{j=1}^{k-1} \pi_j + \pi^*$$

where $\pi_j$ is the exit probability at the $j$th interim analysis and

$$\pi^* = \Pr(\{|Z(t_1)| \leq c_1, \ldots, |Z(t_{k-1})| \leq c_{k-1}, |Z(t_k)| > |z^*|\}).$$

Hence the $P$-value is smaller if the test terminates at an earlier interim analysis and if the test statistics is larger when crossing the group sequential boundary.

In order to generalize this notion of $P$-values and to facilitate the point and confidence interval estimation of the parameter of interest, we define the $P$-value function based on the ordering introduced above as follows:

$$P(\beta) = \Pr((Z(t), t) > (z^*, t^*); \beta)$$

where $t^*$ denotes the time when the group sequential boundary is crossed with the value of the test statistic $z^*$. Then the $P$-value above is simply $P(0)$.

Based on this $P$-value function, the median unbiased estimate $\tilde{\beta}$ of the parameter of interest $\beta$ is defined simply as the solution to

$$P(\beta) = 0.5,$$

while the lower and upper $100(1 - \alpha)%$ confidence limits, $\beta_{L,\alpha/2}$ and $\beta_{U,\alpha/2}$, satisfy

$$P(\beta_{L,\alpha/2}) = \alpha/2 \text{ and } P(\beta_{U,\alpha/2}) = 1 - \alpha/2.$$
4. Independent increment structure

In a general parametric or semiparametric model, statistical information for the parameter of interest $\beta$ is the precision of the best estimator for $\beta$ (meaning the smallest variance) which is approximately $[se(\hat{\beta})]^{-2}$. For example, the information for $\beta = \pi_A - \pi_B$ for a two-sample binomial problem is

$$\left\{ \frac{\pi_A (1 - \pi_A)}{n_A} + \frac{\pi_B (1 - \pi_B)}{n_B} \right\}^{-1}.$$

For the proportional hazards model, the information for the log hazard ratio $\beta$ is proportional to the number of events.

For a general class of parametric and semi-parametric models, treatment effect can be summarized in the parameter $\beta$, and under $H_0 : \beta = \beta_0$, the test statistics satisfies the following distribution

$$Z(t) = \frac{\hat{\beta}(t)}{se\{\hat{\beta}(t)\}} \sim N(\beta_0 I^{1/2}(t, \beta_0), 1).$$

With the transformation,

$$W(t) = I^{1/2}(t, \beta_0)Z(t) \sim N(\beta_0 I(t, \beta_0), I(t, \beta_0)),$$

becomes asymptotically a Brownian motion process where with drift $\beta_0 I(t, \beta_0)$ and variance $I(t, \beta_0)$, which is the information in this case. Furthermore, asymptotically,

$$\frac{[se\{\hat{\beta}(t)\}]^{-2}}{I(t, \beta_0)} \to 1.$$

As summarized in Jennison and Turnbull (1999), independent increment structures have been found in many circumstances case by case. The unifying theory by Scharfstein, Tsiatis and Robins (1997) showed that under a general setting, the joint distribution of the sequential test statistics $\{W(t_1), \ldots, W(t_K)\}$ is asymptotically normal with mean vector

$$\{\beta_0 I(t_1, \beta_0), \ldots, \beta_0 I(t_K, \beta_0)\}$$

and variance-covariance structure

$$\text{Var}\{W(t_k)\} = I(t_k, \beta_0)$$

and

$$\text{Cov}[W(t_k)\{W(t_l) - W(t_k)\}] = 0, \quad k \leq l,$$

namely, that the joint distribution has an independent increment structure. As is shown in the next section, this independent increment structure in the sequentially computed test statistics simplifies the computations involving multivariate normal integrations to those involving recursive univariate integrations. These distributional properties of the sequentially computed test statistics remain true even under the alternative hypotheses albeit contiguous. Jennison and Turnbull (1997) showed the same for a general class of regression models.
5. Inflation factor

If one decides to conduct interim analyses based on group sequential methods, one needs to determine the number of subjects \( n \) per treatment per analysis. For given \( \delta^* \) under the alternative hypothesis, one can compute the power of the group sequential test as

\[
1 - \gamma = 1 - \Pr(|Z_1| < c_1, \ldots, |Z_K| < c_K).
\]

Or conversely, given the desired power \( 1 - \gamma \) of the group sequential test, one can determine the value of \( \delta^* \) that satisfies the above equation. Since \( \delta^* = \delta_A/\sqrt{2\sigma^2/n} \), one can then solve for \( n \) to determine the necessary group size, i.e.,

\[
n = 2(\delta^*)^2 \left( \frac{\sigma}{\delta_A} \right)^2.
\]

Hence the maximum sample size for the group sequential design becomes

\[
2nK = 4(\delta^* \sqrt{K})^2 \left( \frac{\sigma}{\delta_A} \right)^2.
\]

For a fixed sample design with \( K = 1 \), the necessary sample size to detect a difference \( \delta_A \) with power \( 1 - \gamma \) at a two-sided level \( \alpha \) test is

\[
2n = 4(z_{\alpha/2} + z_{\gamma})^2 \left( \frac{\sigma}{\delta_A} \right)^2.
\]

For a group sequential design with \( K > 1 \), the maximum sample size as derived above is

\[
2nK = 4(\delta^* \sqrt{K})^2 \left( \frac{\sigma}{\delta_A} \right)^2
\]

\[
= 4(z_{\alpha/2} + z_{\gamma})^2 \left( \frac{\sigma}{\delta_A} \right)^2 \times \left( \frac{\delta^* \sqrt{K}}{z_{\alpha/2} + z_{\gamma}} \right)^2,
\]

thus a constant multiple of the corresponding sample size for the fixed sample design. This constant multiple

\[
\mathcal{F} = \left( \frac{\delta^* \sqrt{K}}{z_{\alpha/2} + z_{\gamma}} \right)^2
\]

is referred to as the inflation factor. Table 2 gives the inflation factors for the group sequential design with the maximum number of interim analyses including the final analysis, \( K = 2, 3, 4, 5 \), for both Pocock and O’Brien-Fleming designs with power, \( 1 - \gamma = 0.8, 0.90, 0.95 \), at a two-sided test with significance level, \( \alpha = 0.05, 0.01 \).

6. Information-based design and monitoring

Suppose one and only analysis is to be performed at calendar time \( T \) based on

\[
Z(T) = \frac{\hat{\beta}(T)}{se\{\hat{\beta}(T)\}} \sim N(\beta T^{1/2}(T, \beta), 1).
\]

Let \( \beta_A \) denote the treatment difference that is considered to be clinically important to detect. In testing \( H_0 : \beta = 0 \) against \( H_A : \beta = \beta_A \) at a two-sided level \( \alpha \) with power \( 1 - \gamma \) based on the normal distribution for the standardized test statistics \( Z(T) \) above,

\[
\beta_A T^{1/2}(T, \beta_A) = z_{\alpha/2} + z_{\gamma}.
\]
Therefore, for actual implementation, one must convert information into an experimental design which depends on nuisance parameters. Using estimates of the nuisance parameters, one can compute the number of subjects, events, or observations per subject needed to attain the target information. These computations can be done theoretically or by simulations. For example, for a two-sample binomial problem with $n_A = n_B = n$,

$$I = \frac{n}{\pi_A(1 - \pi_A) + \pi_B(1 - \pi_B)} = \left(\frac{z_{\alpha/2} + z_\gamma}{\pi_A - \pi_B}\right)^2,$$

and thus

$$n = \frac{\{\pi_A(1 - \pi_A) + \pi_B(1 - \pi_B)\}(z_{\alpha/2} + z_\gamma)^2}{(\pi_A - \pi_B)^2}.$$

Once the fixed sample size or the fixed information is known, the maximum sample size or the maximum information for the group sequential design can be determined simply as

$$I_{\text{max}} = \left(\frac{z_{\alpha/2} + z_\gamma}{\beta_A}\right)^2 \times F.$$

This works only when there is an independent increment structure in the group sequential test statistics as noted in Section 3.

Now consider monitoring of accumulating data. Instead of analyzing the data only once at the end of the trial, one may compute the test statistics at various interim times $t_1, \ldots, t_K$ for up to $K$ times. The strategy for stopping the trial early is to reject the null hypothesis $H_0$ the first time when the test statistic is sufficiently large, i.e., when

$$|Z(t_k)| \geq c_k.$$

The values $c_1, \ldots, c_K$ are referred to as group sequential critical values.

As shown by Scharfstein, Tsiatis and Robins (1997), any efficient-based test statistics (almost all test statistics used in practice) calculated at interim times $t_1, \ldots, t_K$ behave like

---

Table 2: Inflation factors for group sequential designs

<table>
<thead>
<tr>
<th>$\alpha$</th>
<th>$1 - \gamma$</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>.05</td>
<td>.80</td>
<td>1.11</td>
<td>1.01</td>
<td>1.17</td>
<td>1.02</td>
<td>1.20</td>
</tr>
<tr>
<td>.90</td>
<td>.80</td>
<td>1.10</td>
<td>1.01</td>
<td>1.15</td>
<td>1.02</td>
<td>1.18</td>
</tr>
<tr>
<td>.95</td>
<td>.80</td>
<td>1.09</td>
<td>1.01</td>
<td>1.14</td>
<td>1.02</td>
<td>1.17</td>
</tr>
<tr>
<td>.01</td>
<td>.80</td>
<td>1.09</td>
<td>1.00</td>
<td>1.14</td>
<td>1.01</td>
<td>1.17</td>
</tr>
<tr>
<td>.90</td>
<td>.80</td>
<td>1.08</td>
<td>1.00</td>
<td>1.12</td>
<td>1.01</td>
<td>1.15</td>
</tr>
<tr>
<td>.95</td>
<td>.80</td>
<td>1.08</td>
<td>1.00</td>
<td>1.12</td>
<td>1.01</td>
<td>1.14</td>
</tr>
</tbody>
</table>

$^1$Pocock designs

$^2$O’Brien-Fleming designs
a standardized partial sum of independent normal variables. The variance of this partial sum is directly related to the information at the interim times as

$$I(t_k) = \frac{[se\{\hat{\beta}(t_k)\}]^{-2}}{\sigma}, \quad k = 1, \ldots, K.$$  

This distributional structure with independent increments allows standard calculations for Gaussian random variables according to Armitage, McPherson and Rowe (1969) and McPherson and Armitage (1971) described in the following section.

Suppose one plans to monitor a clinical trial at calendar times $t_1, \ldots, t_K$. Then for the $k$th interim analysis at calendar time $t_k$, one would compute the test statistic

$$Z(t_k) = \frac{\hat{\beta}(t_k)}{se\{\hat{\beta}(t_k)\}},$$

and would terminate the clinical trial and reject $H_0$ if $|Z(t_k)| > c_k$.

In order to compute the critical values $c_k$, one needs first to determine $I_{\max}$ and then to estimate the information fraction as

$$\tau_k = \frac{[se\{\hat{\beta}(t_k)\}]^{-2}}{I_{\max}}.$$

And then apply Lan and DeMets’ procedure based on the pre-specified error spending function $\alpha^*$ and the scaled Brownian motion process on a unit interval $(0, 1)$

$$B(\tau_k) = \frac{W(t_k)}{\sqrt{I_{\max}}} \sim N(\eta\tau_k, \tau_k),$$

where $\eta = \beta\sqrt{I_{\max}}$.

7. Computational issues

For general group sequential tests, one has to deal with a multidimensional numerical integration to compute the probabilities associated with such tests and to determine the group sequential critical values. Suppose that $S_k$ represent the test statistics at the $k$th interim analysis and that up to $K$ interim analyses including the final analysis are planned in advance. Assume that $S_1, S_2, \ldots, S_K$ have a multivariate normal distribution asymptotically. In order to satisfy the overall type I error probability for the group sequential test at a desired significance level, say $\alpha$, one has to determine the group sequential boundary values $b_1, \ldots, b_K$ such that under the null hypothesis

$$\Pr(|S_1| < b_1, \ldots, |S_K| < b_K) = 1 - \alpha.$$  

This calculation involves a numerical integration of a multivariate normal density function. As was mentioned earlier, the MULNOR program by Schervish (1984) can be used for this numerical computation, but it involves very intensive numerical computation and worse yet it can handle multivariate integration of only up to seven dimensions. Slud and Wei (1982) had to use the MULNOR program in developing group sequential methods for the Gehan’s version of Wilcoxon test statistics for censored survival data due to lack of independent increment structure.

When the sequentially computed multivariate normal test statistics have independent increment structure, i.e.,

$$\text{Cov}(S_k, S_l - S_k) = 0 \text{ and } \text{Cov}(S_k, S_l) = k = \text{Var}(S_k), \quad k < l,$$  

— 160 —
or equivalently

\[
\text{Cov}(Z_k, Z_l) = \sqrt{\frac{k}{l}} = \sqrt{\frac{\text{Var}(S_k)}{\text{Var}(S_l)}}, \quad k < l,
\]

the density function for \( S_l \) is completely determined by the density functions for \( S_k \) and the independent increment \( S_k - S_l \) via convolution. Therefore, the following simplification is possible according to Armitage, McPherson and Rowe (1969) and McPherson and Armitage (1971).

Suppose that \( X_1, X_2, \ldots, X_K \) are independent normal observations with mean \( \mu \) and variance 1, and let \( S_k = \sum_{i=1}^{k} X_i \) denote the partial sum and \( Z_k = \frac{S_k}{\sqrt{k}} \) denote the standardized statistics as before. Denote by \( f_k \) the probability density function of \( S_k \) under \( H_0 \) in the sequential procedure. Then

\[
f_k(s) = \int_{-b_{k-1}}^{b_{k-1}} f_{k-1}(u)\phi(s-u)du
\]

is defined recursively where \( f_1 \) is the standard normal density function \( \phi \). Since the probability of stopping at or before the \( k \)th interim analysis is

\[
1 - \Pr(|S_1| < b_1, \ldots, |S_k| < b_k) = 1 - \int_{-b_k}^{b_k} f_k(u)du,
\]

the type I error probability is determined by

\[
\alpha = 1 - \int_{-b_K}^{b_K} f_K(u)du.
\]

Alternatively, with the prespecified exit probability \( \pi_k \) of stopping exactly at the \( k \)th interim analysis, given by

\[
\pi_k = \Pr(|S_1| < b_1, \ldots, |S_{k-1}| < b_{k-1}, |S_k| \geq b_k)
= \int_{-b_{k-1}}^{b_{k-1}} f_{k-1}(u)\{1 - \Phi(b_k - u) + \Phi(-b_k - u)\}du
\]

where \( \Phi \) is the standard normal distribution function, the overall type I error probability becomes \( \alpha = \pi_1 + \cdots + \pi_K \). This computational procedure has been implemented in a publicly available software package as described in Reboussin, DeMets, Kim and Lan (2000) and in a commercial software product EaSt (2000).

Elson (1995) uses the Fast Fourier Transformation algorithm for probability calculations required in group sequential methods, while Schoenfeld (2001) presents a transition matrix-based algorithm for the same calculations. One may also resort to brute force simulations for essentially numerical integrations involved in probability calculations required in group sequential methods. In fact, Schoenfeld (2001) acknowledges that the algorithm is proposed for its ease of programming rather than for its accuracy or speed. With the fast central processing units (CPUs) available on most computers nowadays, the computational efficiency is less of an issue. In fact these various computational algorithms are comparable in their efficiency. However, the algorithm described here requires the least amount of house-keeping in terms of its implementation and also most versatile to handle a variety of computing needs as regards group sequential methods.
8. EaSt implementation

Implementation in the software package EaSt (200) of information-based design and monitoring of group sequential clinical trials and inference following early stopping of such clinical trials is illustrated using a phase III randomized, controlled clinical trial in advanced stage non-small cell lung cancer.

8.1. Information-based design

In this example which was originally designed as a fixed sample design, the primary objective is to determine whether adding a new agent, say X, to an approved chemotherapeutic agent, taxotere, is effective in prolonging overall survival, the primary endpoint of the clinical trial, of patients who are refractory to or have relapsed following platinum-based combination chemotherapy for locally advanced stage IIIb or metastatic stage IV non-small cell lung cancer.

A one-year survival rate is estimated to be 30% among patients treated with taxotere as a second-line therapy in this setting. Improving the one-year survival rate to 40% with taxotere plus X was considered clinical significant in this disease setting. With an accrual of 600 patients over a year with a minimum of a year of follow-up, the clinical trial will have 84% power to detect the stated treatment difference, i.e., the log hazard ratio of 0.273, at a two-sided logrank test at a significance level of 0.05 based on the exponential distribution assumption for the overall survival. This requires observing a total of 469 deaths during the clinical trial, and the end of the clinical trial can be reached as early as 10 months after the last patient is enrolled under the null hypothesis.

Suppose that one is going to design the clinical trial as a group sequential clinical trial with up to four interim analyses including the final analysis according to O'Brien and Fleming (1979). The inflation factor for this group sequential design is approximately 1.02, requiring to observe a maximum of 480 deaths during the clinical trial. Therefore, if one were to keep the same number of patients, the maximum duration becomes 2.09 years.

Figure 1 shows the single-look, i.e., fixed sample design and the four-look group sequential design side by side as they appear in the input/output forms of the EaSt software. The output includes the maximum study duration and the maximum number of events required based on the maximum information for the study. It also provides the operating characteristics of the designs in terms of the expected number of subjects, study duration and number of events under different hypotheses.

Figure 2 displays the two-sided O'Brien-Fleming stopping boundaries for the four-look group sequential design in terms of the critical values for the group sequential tests. Figure 3 shows the exit probabilities at each of the four looks under the null and alternative hypotheses. Note that the probability of stopping early under the null hypothesis is only 0.021. This is a consequence of using the O'Brien-Fleming stopping boundaries.

In order to achieve the desired power with the four-look group sequential design, we need to keep the clinical trial open until 480 events are observed during the study. The maximum study duration (accrual plus follow-up) needed to observe 480 events can be reduced by increasing patient accrual or the number of patients can be reduced by increasing the follow-up duration. Obviously there is a trade-off between the two approaches. For example, if one is willing to enroll a maximum of 600 subjects over one year, then the followed-up duration could be as long as 1.09 years. However, if one is willing to enroll a maximum of 720 subjects over 1.2 years, the follow-up duration could be as long as only 0.5 years. Figure 4 is useful for evaluating the trade-off between decreasing the maximum study duration and decreasing
Information-Based Design and Monitoring

Fig. 1: Fixed sample and four-look group sequential designs

Fig. 2: O'Brien-Fleming stopping boundaries for four-look design
Fig. 3: Exit probabilities under null and alternative hypotheses

Fig. 4: Trade-off between decreasing study duration and accrual time
the number of patients. The computational procedure for design of group sequential clinical trials with failure time data are summarized in detail in Kim and Tsiatis (1990), Kim (1995) and Kim, Boucher and Tsiatis (1995).

8.2. Information-based monitoring

For any chosen group sequential design, there is a corresponding error spending function. EaSt (2000) fits a beta function to the cumulative exit probabilities as an error spending function for group sequential monitoring internally. At interim monitoring, we input the current value of the test statistic along with statistical information at each interim look. The EaSt software then computes the group sequential boundary based on the error spending function corresponding to the chosen group sequential design.

<table>
<thead>
<tr>
<th>Interim Monitoring of Survival Design - Plan2 of Untitled2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Plan Id</strong></td>
</tr>
<tr>
<td><strong>Look Number</strong></td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
</tbody>
</table>

Fig. 5: Information-based monitoring

Figure 5 shows the history of a hypothetical group sequential monitoring of the clinical trial. At the first interim analysis, there are 125 events observed with the group sequential test statistic of $Z = 1.97$. The information fraction for this interim analysis is thus $\tau_1 = 125/480 = 0.260$, and the group sequential boundary was determined to be 3.929, resulting in continuation of the clinical trial.

At the second interim analysis, there are 245 events observed with the group sequential test statistic of $Z = 2.40$. The information fraction for this interim analysis is thus $\tau_2 = 245/480 = 0.510$, and the group sequential boundary was determined to be 2.880, resulting in continuation of the clinical trial.

At the third interim analysis, there were 375 events observed with the group sequential test statistic of $Z = 2.60$. The information fraction for this interim analysis is thus $\tau_3 = 375/480 = 0.781$, and the group sequential boundary was determined to be 2.293, resulting in group sequential boundary crossing and leading to early stopping of the clinical trial.

8.3. Inference following early stopping

The $P$-value and the point and confidence interval estimate of the parameter of interest, which in this example is the log hazard ratio, are displayed in Figure 6. When the clinical trial is terminated after the third interim analysis with $z^* = 2.6$ and 375 events, $P$-value is
0.011. The median unbiased estimate of the log hazard ratio is 0.265, and the 95% confidence interval is (0.056, 0.470). The study was designed to detect the log hazard ratio of 0.274. (Please note the signs are reversed.)

9. Discussion

As described in this article, computations required for design and monitoring of group sequential clinical trials become quite simplified if the increments between the sequentially computed test statistics are statistically independent. With the independent increment structure in the group sequential tests, one can employ the so-called information-based design and monitoring based only on the fixed sample design and the inflation factor for design and on the error spending function approach for monitoring. The general results from Scharfstein, Tsiatis and Robins (1997) and from Jennison and Turnbull (1997) indicate that most efficient-based test statistics indeed have a joint distribution which is multivariate Gaussian with independent increments, allowing ready applications of the group sequential methods for interim analysis in clinical trials with a wide variety of outcome data types and the associated regression models.

Acknowledgements

This work was partially supported by the grant CA52733 from the National Institutes of Health, Department of Health and Human Service, U.S. Government.

REFERENCES


