DETERMINATION OF POWER AND SAMPLE SIZE
IN THE DESIGN OF CLINICAL TRIALS WITH
FAILURE-TIME ENDPOINTS AND INTERIM ANALYSES

by

MINGGAO GU
Department of Mathematics and Statistics
McGill University

and

TZE LEUNG LAI
Department of Statistics
Stanford University

October 1998

This research was supported in part by
National Science Foundation grant DMS-9704324

Department of Statistics
Sequoia Hall
STANFORD UNIVERSITY
Stanford, California 94305-4065

http://www-stat.stanford.edu
Determination of Power and Sample Size in the Design of Clinical Trials with Failure-Time Endpoints and Interim Analyses

Minggao Gu, Ph.D and Tze Leung Lai, Ph.D

Department of Mathematics and Statistics, McGill University, Montreal, Canada and Department of Statistics, Stanford University, Stanford, California

Abstract

An important but difficult task in the design of a clinical trial to compare time to failure between two treatment groups is determination of the number of patients required to achieve a specified power of the test. Since patients typically enter the trial serially and are followed until they fail or withdraw from the study or until the study is terminated, the power of the test depends on the accrual pattern, the noncompliance rate and the withdrawal rate in addition to the actual survival distributions of the two groups. Incorporating interim analyses and the possibility of early stopping into the trial increases its complexity, and although normal approximations have been developed for computing the significance level of the test when the logrank or other rank statistics are used, there are no reliable analytic approximations for evaluating the power of the test. This article presents methods, based on Monte Carlo simulations and recent advances in group sequential testing with time-to-event responses, to choose appropriate test statistics, to compute power and sample size at specified alternatives, to check the adequacy of commonly used normal approximations of the Type I error and to assess the performance of different interim analysis strategies. It also presents a computer program implementing these methods.

KEY WORDS: Group sequential tests, loss to follow-up, time dependent hazard ratios, noncompliance rates.

RUNNING TITLE: Group sequential tests of failure times.
INTRODUCTION

The Monte Carlo simulation method has been recognized as a flexible and practical way to compute power and sample size in clinical trials comparing time to failure between two treatment groups, in which patients enter serially and are followed until they fail or withdraw from the study or until the study is terminated. Moreover, it also enables one to compute the expected duration of the trial that may be stopped prior to its prescheduled end at times of interim analysis of the data, and to check the adequacy of the normal approximation for the Type I error under various scenarios of baseline survival, censoring pattern, noncompliance and accrual rates. To provide the working clinical trial designer with a tool to perform these Monte Carlo computations, Halpern and Brown [1] developed a simulation program to evaluate the sample size required to achieve some specified power in a fixed-duration trial (without interim analyses) for the logrank or Gehan's test. Subsequently they presented another simulation program for these test statistics in [2] to incorporate group sequential testing with either the Pocock [3] or the O'Brien-Fleming [4] boundary. Their program can be used to compute both the Type I error and the power of the group sequential test under user-specified null and alternative hypotheses of the survival distributions which are assumed to be either piecewise exponential or cure-rate models. Other parameters to specify include the number of subjects entering the study per unit time up to the end of a user-specified accrual period, length of follow-up, and the times of the interim analyses. Moreover, for given Type I error and power, the program also computes the stopping boundary and required sample size for the user. Incorporating time-dependent rates of loss to follow-up and noncompliance, Shih [5] and Lakatos [6,7] also developed, for fixed-duration trials, sample size calculation programs which are based on Markov models and normal approximations to the distributions of the rank statistics, instead of using Monte Carlo simulations.

At the time Halpern and Brown developed their program that was later published in [2], a number of fundamental issues concerning group sequential testing with time-to-event responses remained unresolved. It was widely recognized that the Pocock and the O'Brien-Fleming boundaries, which were developed for immediate responses under
the assumption of an equal number of responses for each time interval between two adjacent interim reviews of the data, could not be justified for censored failure-time responses. Lan and DeMets [8,9] proposed a more flexible way of choosing the stopping boundaries via the concepts of "use function" and "information time". To circumvent difficulties with specifying the maximum information time for censored failure-time data in the Lan-DeMets approach, Slud and Wei [10] proposed to adopt a crude "use function" which simply decomposes the Type I error $\alpha$ as $e_1 + \cdots + e_k$ and chooses a boundary so that the test statistic has a probability of $e_j$ of first crossing the boundary at the $j$th interim analysis under the null hypothesis.

In this paper we describe an alternative approach based on refinements of an idea due to Haybittle [11], extending it to failure-time data. This approach is shown to have many desirable features. We also present a new simulation program that broadens the scope of the Halpern-Brown program [2] in several major ways to provide the investigator with a flexible and powerful tool to plan clinical trials with failure-time endpoints and interim analyses. First, instead of the Pocock and O'Brien-Fleming boundaries, there are four options available to the user for choosing the stopping boundary: Slud-Wei, Lan-DeMets, Haybittle-type boundaries mentioned above, plus any other boundary specified by the user. Secondly, while the Halpern-Brown program does not incorporate patient withdrawal (censoring) and noncompliance (crossover), our program allows the user to incorporate these considerations by specifying the censoring distributions and crossover rates of the two treatment groups. Thirdly, our program gives the user an additional way of specifying the survival distribution of the new treatment by specifying its time-dependent hazard ratio (which is assumed to be piecewise constant) relative to the baseline survival distribution. Finally, in the Halpern-Brown program, the user can only choose between the logrank statistic and Gehan's [12] generalization of the Wilcoxon statistic which, as pointed out by Lan and Wittes [13] and Gu, Lai and Lan [14], is inexorably linked to the censoring pattern and is much more difficult to implement for group sequential testing but has worse performance than the alternative generalization of the Wilcoxon statistic to censored data by Peto and Peto [15] and Prentice [16]. Both the logrank statistic and the Peto-Prentice generalized Wilcoxon statistic are special cases of the class of $G^p$ statistics introduced by Harrington and
Fleming [17], which in turn belong to the Beta family of statistics proposed by Self [18]. Our program allows the user to choose the parameters $\rho$ and $\tau$ in the Beta family of statistics (with $\rho = 0$ and $\tau = 0$ corresponding to the logrank statistic and $\rho = 1$ and $\tau = 0$ corresponding to the Peto-Prentice generalized Wilcoxon statistic), and a number of these statistics can be simulated together in a single simulation run to compare their performance under the specified parameters.

In the next two sections, we describe in detail the methods, based on recent advances in group sequential testing with time-to-failure responses, to choose appropriate stopping boundaries and test statistics. This is followed by three sections describing how the survival functions and noncompliance rates are specified and other main features of our program. In the final two sections, we first give some examples to demonstrate the accuracy of the program and to illustrate its applications to sample size determination, and then present a case study in applying these methods to the design of a myocardial infarction trial.

METHODS FOR GENERATING BOUNDARIES

In this section we describe a simple, efficient method to generate stopping boundaries, which we shall refer to as Haybittle-type boundaries as they are based on modifications and refinements of an idea due to Haybittle [11]. We begin by giving some background and explanation of the Slud-Wei and Lan-DeMets boundaries. These three types of boundaries are available as options in our program, and there is an additional option for the user to specify the boundary. We shall only describe two-sided tests here, since there are obvious analogues for one-sided tests.

The method of Slud and Wei [10] requires the user to specify positive numbers $e_1, \cdots, e_k$ (where $k$ is the number of interim analyses) such that $e_1 + \cdots + e_k = \alpha$, the overall Type I error. Thus $e_i$ is the amount of Type I error spent at the $i$th interim analysis, for which the boundary $b_i$ is determined by

$$e_i = \Pr(|W_i| \leq b_i \sqrt{V_i}, \ldots, |W_{i-1}| \leq b_{i-1} \sqrt{V_{i-1}}, |W_i| > b_i \sqrt{V_i}).$$  \hspace{1cm} (1)

Here and in the sequel, $W_i$ denotes the test statistic at the $i$th interim analysis and $V_i$
denotes the corresponding variance estimate. Since our program takes the \( W_i \) from the
Beta family of statistics whose asymptotic distributions under the null hypothesis are
normal with independent increments, the probability in (1) can be computed, under
this normal approximation, by using the recursive numerical integration method of
Armitage, McPherson and Rowe [19]; moreover, formula (5.2) of [20], which is used by
the program as the variance estimate \( V_i \), is consistent under the null hypothesis and
local alternatives, as shown in [20].

In the Lan-DeMets [8] method to determine the stopping boundary, the user spec-
ifies a non-decreasing function \( u : [0,1] \rightarrow [0,\alpha] \), where \( \alpha \) is a prescribed Type I error
and \( u(0) = 0, u(1) = \alpha \). Concerning how \( u \) should be chosen, see [9, 21]. The user also
has to specify a positive number \( v \) representing the maximum allowable null variance of
the test statistic. At the \( i \)th interim analysis, stopping occurs and the null hypothesis
is rejected if \( |W_i| > b_i\sqrt{V_i} \), where \( b_i \) is determined by the equation
\[
    u(V_i/v) = \Pr(|W_j| > b_j\sqrt{V_j}, \text{ for some } j \leq i).
\]  \hspace{1cm} (2)

The study also stops at the \( i \)th interim analysis if \( V_i \geq v \), in which case the null
hypothesis is rejected if \( |W_i| \geq b_i\sqrt{V_i} \), where \( b_i \) is determined by (2) with the left-
hand side replaced by \( u(1) \). In our program, the user can choose one of four use
functions discussed in [8] and [22]. They are the O'Brien-Fleming use function \( u(t) = 2 - 2\Phi(z_{\alpha/2}/\sqrt{t}) \),
the Pocock use function \( u(t) = \alpha \log(1 + (e - 1)t) \), linear use functions
and another use function in [22] of the form \( u(w/v) = A(w) \), where \( A(w) = 0 \) for
\( w \leq v_0 \),
\[
    A(w) = (a - a^{-1})\phi(a) \log(w/v_0) + 4a^{-1}\phi(a)
\]
for \( v_0 < w < v \), in which \( \phi \) (or \( \Phi \)) is the standard normal density (or distribution)
function and \( a \) is so chosen that \( A(v) = \alpha \), the overall Type I error. The function \( A(w) \)
was introduced by Siegmund [23, p.76], and the use function \( u(w/v) = A(w) \) is called
Siegmund's use function in our program.

The use function approach for survival data requires pre-specification of the maxi-
mum information time given by the null variance \( v \) of the test statistic at the calendar
time when the trial is scheduled to end, which is often difficult to evaluate and to
explain to clinical investigators. We have recently developed in [22] a simple but sta-
tistically efficient method that incorporates information time implicitly via a refinement and extension of Haybittle's [11] boundary to failure-time data, instead of explicitly via a use function. To begin with, consider the problem of testing whether the mean \( \theta \) of independent normal observations \( Z_1, Z_2, \ldots \) is 0, assuming that \( \text{var}(Z_i) = 1 \) is known. The Neyman-Pearson test rejects \( \theta = 0 \) at level \( \alpha \) if \( |Z_1 + \ldots + Z_k| \geq z_{\alpha/2} \sqrt{k} \), where the fixed sample size \( k \) is chosen to attain some given power \( 1 - \beta \) at a specified alternative \( \theta \neq 0 \). The basic idea behind Haybittle's [11] repeated significance test is to keep \( k \) and \( \alpha \) as the maximum sample size and significance level but to allow for early stopping when the data are monitored sequentially, at the expense of some loss in power. This leads to the stopping boundary (for \( |Z_1 + \ldots + Z_j| \)) of the form \( b\sqrt{j} \) if \( j < k \), but of the form \( c\sqrt{j} \) if \( j = k \), with \( c \) near \( z_{\alpha/2} \). In view of this, we have proposed in [22] the following procedure involving a maximum of \( k \) repeated significance tests. First determine \( b \) such that

\[
\Pr\{|\sum_{i=1}^{j} Z_i| \geq b\sqrt{j}, \text{ for some } j \leq k - 1\} = \epsilon \alpha,
\]

where \( 0 < \epsilon < 1 \) is small and the probability is evaluated under the null hypothesis \( \theta = 0 \). The probability in the left-hand side can be evaluated by recursive numerical integration, starting with an approximation formula in [22] as an initial value of \( b \) and using a few iterations to solve the above equation approximately. With \( b \) thus chosen, we have in the case of survival data the following Haybittle-type boundary for the test statistics \( W_i \) that have appeared in (1) and (2). The boundary points for the first \( k - 1 \) interim analyses are \( b\sqrt{V_j}, j = 1, \ldots, k - 1 \), while the last boundary point is \( c\sqrt{V_k} \) so that

\[
\Pr\{|W_j| \geq b\sqrt{V_j}, \text{ for some } j \leq k - 1, \text{ or } |W_k| \geq c\sqrt{V_k}\} = \alpha. \tag{3}
\]

In practice, clinical trialists carrying out interim analysis would like to avoid (i) stopping the trial too early with the consequence that the evidence in favor of a treatment may be unconvincing, and (ii) failing to reject the null hypothesis at the prescheduled end of the trial when a fixed-duration trial rejects the hypothesis. By choosing \( b \) considerably larger than \( c \) as described above, the Haybittle-type boundary is able to avoid both (i) and (ii). As shown in [22], it has similar statistical properties as Sigmund's use function in which \( a \) and \( v \) are chosen in some optimal way that depends
on knowledge of the null variance of the test statistic at the terminal date $t_k$. On the other hand, the Haybittle-type boundary does not require such prior knowledge for its implementation since it does not involve $v$; it accomplishes this via the choice of $b$ described above.

In addition to the Slud-Wei, Lan-DeMets and Haybittle-type methods, our program also gives one the option of specifying one’s own boundary. For example, one may want to try asymmetric boundaries of the type described in Section 3.3 of DeMets and Ware [24], terminating the study early when the data suggest inferiority of the new therapy, even though the evidence may be substantially less than that required to claim the treatment to be beneficial.

CHOICE OF TEST STATISTICS

In [14] and [20] the importance of using appropriate test statistics to achieve good performance of a group sequential test is explained. The Beta family proposed by Self [18] provides a rich class of good candidates. The family essentially consists of weighted logrank statistics with weight $\hat{S}(s)(1 - \hat{S}(s))^\tau$ at $s$, where $\hat{S}$ is the Kaplan-Meier survival curve of the combined sample. To be more specific, suppose that for the $i$th patient, $E_i$ is the entry time, $T_i$ is the lifetime of interest, $Z_i$ is the treatment indicator (0 or 1) and $C_i$ is the censoring time. Using $a^+$ to denote the positive part of a number $a$ ($a^+ = a$ if $a \geq 0$, $a^+ = 0$ otherwise) and $1_A (=1$ or $0)$ to denote the indicator of the occurrence of an event $A$, the data available at calendar time $t$ consist of $(X_i(t), Z_i, \delta_i(t))$, $i = 1, \ldots, n$, where $X_i(t) = \min(T_i, (E_i - t)^+, C_i)$ and $\delta_i(t) = 1_{\{X_i(t) = T_i\}}$. For any fixed $\rho$ and $\tau$, Self’s test statistic for comparing the two groups of lifetimes at time $t$ has the form

$$W(t; \rho, \tau) = \sum_{i=1}^{n} \delta_i(t) \hat{S}_i(X_i(t))^{\rho}(1 - \hat{S}_i(X_i(t)))^{\tau} \left( Z_i - \frac{M_{1,t}(X_i(t))}{M_{0,t}(X_i(t))} \right),$$

where $M_{j,t}(x) = \sum_{i=1}^{n} Z_i^j 1_{\{X_i(t) \geq x\}}$ for $j = 0, 1$, and $\hat{S}_i$ is the Kaplan-Meier survival curve based on the combined sample $(X_i(t), Z_i, \delta_i(t))$, $i = 1, \ldots, n$. Note that this family of statistics in the case $\tau = 0$ reduces to the $G^\rho$ family of Harrington and Fleming [17]. In particular, it includes the logrank statistics ($\rho = 0$ and $\tau = 0$) and
the Peto-Prentice generalized Wilcoxon statistics ($\rho = 1$ and $\tau = 0$).

One important advantage of the Beta family of statistics is that the asymptotic joint distribution of $W(t_i; \rho, \tau)$ evaluated at different times $t_1 < \cdots < t_k$ (measured from the beginning of the trial) of interim analysis is, under the null hypothesis, normal with independent increments and with variances

$$V(t_i, \rho, \tau) = -\frac{1}{2} \int_0^{t_i} S(u)^{2\rho} (1 - S(u))^{2\tau} w(u) \#(t_i - u) dS(u),$$

for $i = 1, \cdots, k$, where $S$ is the common survival function of the $T_i$ under the null hypothesis, $\#(t)$ is the expected number of patients accrued up to time $t$ after the beginning of the trial, $w_j(s)$ is the censoring (withdrawal) survival function of patients at $s$ units of time after entry into the study for treatment group $j$ ($j = 1, 2$), and $w(s) = w_1(s) w_2(s)/(w_1(s) + w_2(s))$, cf. Eq. (3.10) of [20]. In the case $w_1 = w_2$, which is often assumed in the literature, and for the logrank statistic ($\rho = \tau = 0$), (4) reduces to 1/4 times the expected number of failures up to time $t_i$, a commonly used variance formula for the logrank statistic. Note that if one uses the Lan-DeMets method to determine the stopping boundary for the Beta family of statistics, then $v$ in (2) can be taken to be $V(t_k, \rho, \tau)$.

**SPECIFICATION OF BASELINE AND ALTERNATIVE DISTRIBUTIONS**

In our program, the baseline (or control-group) survival function is specified by a non-increasing piecewise log-linear function. The user is asked to input the times $0 < t_1 < \cdots < t_l$ and the values of the survival function $S(t_1) > \cdots > S(t_l) > 0$. The program sets $S(0) = 1$. For $t_{i-1} \leq t \leq t_i$, $i = 1, \ldots, l$, the value of $\log S(t)$ is determined from the user-specified values at the end-points of the interval by linear interpolation, with $t_0 = 0$. For $t > t_l$, the value of $\log S(t)$ is determined by the linear extrapolation of the values specified for the interval $[t_{i-1}, t_i]$. In particular, to specify an exponential survival function with $S(t^*) = .5$, one sets $l = 1$, $t_1 = t^*$ and $S(t_1) = .5$.

The alternative distribution function (of the treatment group under the alternative hypothesis) can be specified by two ways in our program. The first way is the same as
that used to specify the baseline survival function. The second way is to specify it via the ratio of the alternative hazard function to the baseline hazard function. The ratio is assumed to be constant on the intervals \([s_{i-1}, s_i), i = 1, \ldots, p+1\) \((s_0 = 0, s_{p+1} = \infty)\). The \(s_i\) need not be the same as the \(t_j\). In particular, if one has proportional hazards alternatives, then one should choose \(s_p\) to be larger than the maximum study duration so that the ratio on \([s_p, \infty)\) becomes irrelevant. If the hazard ratio on \([s_{i-1}, s_i]\) is \(r_i\), then the alternative survival function \(S_A\) can be expressed in terms of the baseline survival function \(S\) by

\[
S_A(x) = S(s_1)^{r_1-r_2} \cdots S(s_{i-1})^{r_{i-1}-r_i} S(x)^{r_i} \quad \text{for} \ x \in [s_{i-1}, s_i).
\] (5)

The censoring (due to loss to follow-up) distribution functions for the control and treatment groups are specified in the same way as the baseline survival function. In particular, if there is no loss to follow-up, the user can set \(S_c(t^*) = 1\) for the survival function \(S_c\) of the censoring distribution at some \(t^*\) exceeding the prescheduled duration of the trial.

**SPECIFICATION OF NONCOMPLIANCE RATES**

The noncompliance rates, or crossover rates, are specified by the drop-out rate (from treatment group to control group) and the drop-in rate (from the control group to treatment group); see [25]. They govern the proportion of subjects in each specified time period who change their lifetime distribution to that of the other group. In our program, the user specifies this by inputting the times and the rates of crossover for each group at these times, assuming that crossovers can only happen at the specified time points. This restriction is not important since one can approximate the case in which crossover occurs continuously over an interval by specifying enough points (in the interval) at which the only crossovers occur. The crossover rates at these points (which are endpoints of subintervals) are related to the rate \(r\) on the entire interval (assuming continuous crossover) via the formula

\[
r_1 + (1 - r_1)r_2 + \ldots + (1 - r_1)\cdots(1 - r_{k-1})r_k = r,
\] (6)
where $r_1, \ldots, r_k$ are the crossover rates of $k$ consecutive subintervals. In particular, if $r_1 = \cdots = r_k$, then $r_1 = 1 - (1 - r)^{1/k}$.

**THE PROGRAM**

The main components of this program are the following five subroutines: *vari*, *probbdy*, *gd*, *handle*, *mcalstat*.

The subroutine *vari* calculates of the theoretical variances (4) of the test statistics under the null hypothesis at different times of interim analysis according to the baseline survival distribution, censoring distributions, accrual length and accrual rate. Numerical integration using Simpson’s rule is used, noting that the functions $S$ and $w$ in (4) are piecewise log-linear.

The subroutine *probbdy* determines the current boundary point $b_j$ at each time of interim analysis. For two sided-tests, let $f_1(x) = \phi(x/\sqrt{V})/\sqrt{V}$, and for $j = 2, \cdots, k$, define recursively

$$f_j(x) = \int_{-b_{j-1}}^{b_j} f_{j-1}(y)\phi((y-x)/\sqrt{V_j-V_{j-1}}) dy.$$  

For the Slud-Wei or Lan-DeMets method for generating boundaries, $b_j$ is determined by

$$1 - \int_{-b_j}^{b_j} f_j(y) dy = \alpha_j.$$  

In particular, if the Slud-Wei method is specified, $\alpha_j = e_1 + \cdots + e_j$. If the Lan-DeMets method is specified, then $\alpha_j = u(V_j/v)$, where $u$ is the use function. The numerical integration in this subroutine is done by Gaussian quadrature [26] with 24 points.

The subroutine *gd* generates the raw data. It uses the idea that if $U$ is uniformly distributed in $[0,1]$ and $S$ is a survival function, then $S^{-1}(U)$ has $S$ as its survival function. The value of $S^{-1}(u)$ for the baseline survival function is returned by the routine *rinusu*. The value of the inverse of the alternative survival function is returned by *hinusu*, which uses formula (5).

The subroutine *handle* uses the output of *gd* and the interim analysis time to generate the observations for that specific interim analysis.
The subroutine *mcalcstat* takes the observations generated by *handle* and calculates the statistics in the Beta family described above.

The default number of simulations in a run is set to be 1000. This number is chosen to accommodate two conflicting aims; there should be enough runs so that the power simulated has a reasonably small standard error and it should not take too long to complete a simulation. If the power is above 80%, then a simulation run with 1000 simulations corresponds to a standard error of less than 1.3%.

The program is written in FORTRAN. It can be run in a UNIX environment or on IBM-PC compatibles with a FORTRAN compiler. The program is currently designed to run with a formatted input file. Both the program and a sample input file are available on the World Wide Web (http://www.math.mcgill.ca/~minggao/). The sample input file contains instructions for how to change the inputs to obtain desired results. The user who has difficulties running the program can send electronic mail for help to the first author: minggao@math.mcgill.ca

**EXAMPLES AND SAMPLE SIZE DETERMINATION**

*Example 1.* As an illustration of the accuracy of our program, we compare the results on power and expected study duration obtained by the program with those reported by Siegmund [23] in a simulation study of the performance of his modification, for logrank statistics, of Haybittle's [11] repeated significance tests based on normally distributed sample means. In this simulation study, a total of 350 patients arrive independently and uniformly over a 3-year interval, to be randomized to a treatment or placebo. There are \( k = 10 \) periodic reviews at times \( t = 1, 1.5, \ldots, 5.5 \) (years). The survival distribution of the placebo group is assumed to be exponential with mean 3 (years), so the hazard rate is \( \lambda = 1/3 \), and that of the treatment group is exponential with hazard rate \( \lambda/1.8, \lambda/1.65, \lambda/1.5, \lambda/1.4 \). Siegmund takes \( b = 2.85, k = 10 \) for the Haybittle-type boundary in (3). Despite some minor differences between Siegmund's modification of Haybittle's rule and the more flexible version used in our program, with 1000 simulations (instead of Siegmund's 400) in each case, the program gave the same
numerical results (to two significant digits) on the power and expected study duration for the aforementioned alternatives as Table 5.6 in Chapter 4 of [23].

**Example 2.** Our program can be used for power calculations in fixed-duration trials, for which the user of the program needs only to input 1 when queried about the number of interim analyses. In this example, we compare some numerical results on the power of logrank tests obtained by Lakatos [7] using normal approximations of the sampling distributions of logrank statistics under local alternatives with those obtained by the program using simulations to evaluate the exact sampling distributions, in a cardiovascular (CVD) trial and in a cancer trial with loss to follow-up and noncompliance. The results on power obtained by the program with 2000 simulations are listed in Table 1 below, to be compared with the value of .9 in two cases (labeled by their sample sizes) in Table 2 (CVD trial) and Table 3 (Cancer trial) of [7].

<table>
<thead>
<tr>
<th>Table 1. Simulated Power on CVD and Cancer Trials</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample Size</td>
</tr>
<tr>
<td>------------</td>
</tr>
<tr>
<td>Power</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

**Example 3.** By inputting the hazard ratio to be constantly equal to 1, we can use the program to compute the Type I error of the test and thereby to check the adequacy of the normal approximations (under the null hypothesis) that are used to compute the group sequential boundaries. For example, suppose we wish to design a 5-year clinical trial in which 900 patients are accrued at a uniform rate during the first three years. The control group failure rate is 0.10 per year. For the noncompliance or crossover rates, assume that there is a constant drop-in rate of 3% per year and the drop-out rates are 10% for the first year and 5% for the last four years. Also assume that there is a 5% loss to follow-up for both groups during the 5-year period and that interim analysis is scheduled at 1, 1.5, 2, 2.5, ..., 4.5 and 5 years. Suppose that the Lan-DeMets approach with the O'Brien-Fleming use function is used for interim analyses based on the logrank statistics. The boundary chosen by (2) has a Type I error of 5% based on normal approximation. We can also determine this Type I error by Monte Carlo simulation under the specified distribution for the control group, assuming the null hypothesis that the distribution for the treatment group also has this
specified distribution. We used the program to compute the Type I error from 50,000 simulations, incorporating the same features on loss to follow-up and noncompliance as in the demonstration but with 50,000 simulations. We found the Type I error to be .051 (with standard error of .001), showing very good agreement with the value .05 obtained by normal approximation.

Example 4. Note that Examples 2 and Example 3 both incorporate noncompliance while Example 1 assumes that there is no noncompliance. We next consider the effect of noncompliance on the power calculated in Example 1 in the case when the hazard ratio is 1.65. Besides the Haybittle-type group sequential trial we also consider the fixed-duration trial that terminates at \( t = 5.5 \) (years). A constant drop-in rate of .03 per year is assumed. In addition, for each of the three noncompliance scenarios considered in Table 2, there is a constant drop-out rate of .3, .4 and .5 per year, respectively. These crossovers are assumed to occur only in the middle of each year. We use here the Haybittle-type boundary (3) with \( b = 2.6 \) in the sequential design. The other parameters are assumed to be the same as in Example 1. The results on power and expected duration obtained from our program using 10,000 simulations are reported in the following table.

**Table 2. Simulated Power and Expected Duration With or Without Noncompliance**

<table>
<thead>
<tr>
<th>Noncompliance</th>
<th>Fixed Duration (5.5 years)</th>
<th>Sequential Design</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Power</td>
<td>Power Expected Duration</td>
</tr>
<tr>
<td>No Noncompliance</td>
<td>.961</td>
<td>.938</td>
</tr>
<tr>
<td>Noncompliance (.3)</td>
<td>.492</td>
<td>.499</td>
</tr>
<tr>
<td>Noncompliance (.4)</td>
<td>.355</td>
<td>.392</td>
</tr>
<tr>
<td>Noncompliance (.5)</td>
<td>.252</td>
<td>.305</td>
</tr>
</tbody>
</table>

The results in Table 2 show the following patterns in the present case of proportional hazards alternatives: (i) When there is no noncompliance, the fixed-duration logrank test is more powerful than the group sequential test. (ii) When noncompliance is sufficiently high, this is reversed, yielding both savings in duration and increase in power for the group-sequential test. (iii) Power decreases monotonically with the noncompliance rate, in both fixed-duration and group-sequential tests.
The fact that a group sequential logrank test can have more power and a shorter expected duration than its fixed-duration counterpart in the case of nonproportional hazards alternatives has been noted in [20] and [22], where an asymptotic theory that explains this phenomenon is also given. In the present example, although we have proportional hazards alternatives, noncompliance essentially makes the "effective hazards" nonproportional. We now give a simple illustration of this point. Suppose the failure times in the control and treatment groups are exponentially distributed, with respective rates (constant hazards) 1 and 1/2. For simplicity, assume the simultaneous entry scenario in which all subjects enter the study at time \( t = 0 \). Suppose that there is no drop-in but that 100% drop-out occurs at time \( t = 2\ln 2 \). Thus the actual hazard function of the treatment group is 1/2 on the interval \( 0 \leq t < 2\ln 2 \) and is 1 on the interval \( t \geq 2\ln 2 \) because of this crossover. Since the fixed-duration logrank test terminating at \( t^* = 2\ln 2 \) is more powerful than that terminating at \( t^* = 10 \), it is not surprising that a group sequential logrank test with two scheduled analyses at \( t = 2\ln 2 \) and \( t = 10 \) is more powerful than the fixed-duration test at \( t^* = 10 \); see [20] for a similar example dealing with nonproportional hazards alternatives (without noncompliance).

The emphasis of our program is to evaluate the performance and power of different statistics and boundaries under the same baseline and alternative survival distributions, rather than just determining the sample size. However, since it is relatively fast to complete a medium sized simulation run on a workstation (about 1 minute for a simulation with 1000 replications, 1000 patients, 2 statistics and 3 interim analyses on a SPARCstation 2), the user can always restart the program with another sample size (accrual time or accrual rate) to achieve the desired power or to investigate the sensitivity of a design to slight changes in assumptions.

In particular, one can apply a bisection search to determine the sample size that achieves the desired power at a given alternative, using the program to compute the power associated with each sample size selected in the bisection search. Specifically one starts with some upper and lower bounds \( n_1 > n_2 \) for the sample size. There are usually practical constraints on the sample size in clinical trials. For example, budget considerations would lead one to a natural upper bound \( n_1 \) for the sample size, while
publication and administrative considerations would lead one to a natural lower bound \( n_2 \). If \( n_2 \) already yields power (computed using our program) exceeding the desired level, then the sample size can be chosen to be \( n_2 \). If a sample size as large as \( n_1 \) still falls short of achieving the desired power, then one may consider the practicality of carrying out the trial given the available resources. Typically, the power for the sample size \( n_2 \) falls below the desired level, and that for the sample size \( n_1 \) exceeds the desired level. Then one uses the program to compute the power for the sample size which is the integer nearest to \((n_1 + n_2)/2\). Replace \( n_1 \) by \((n_1 + n_2)/2\) if this power exceeds, or by \( n_2 \) if the power falls below, the desired level. This procedure is repeated until one finds a sample size that achieves the desired power within 2 standard errors of the Monte Carlo simulation. In order to save time, we suggest to start running the program with 100 to 300 simulations, and when the power obtained falls in some neighborhood of desired level, to increase the number of simulations to 1000 or more so that the standard error is reasonably small.

**APPLICATION TO THE DESIGN OF A MYOCARDIAL INFARCTION TRIAL**

The Beta-Blocker Heart Attack Trial (BHAT) was a randomized double-blind trial to assess the effect of propranolol on reducing mortality in patients who had at least one myocardial infarction. The trial started in June 1978 and was terminated on October 2, 1981, due to strong evidence in favor of the treatment (propranolol). More details can be found in [27].

The design issues of the actual trial have been discussed by the Beta-Blocker Heart Attack Trial Research Group [28]. More specifically, the trial was designed to run for four years and, based on other studies, the 3-year mortality rate in the placebo group was estimated to be 18%. The treatment, propranolol, was assumed to reduce the mortality rate by 28%. This amounts to a proportional hazards alternative with hazard ratio \( \log(1 - .18(1 - .28))/\log(1 - .18) = .699 \). The noncompliance rate in the treatment group, or the drop-out rate, was assumed to be 12% in the first year, 8% in the second year and 6% in the third year. The noncompliance rate in the
placebo group, or the drop-in rate, was assumed to be 7% for each year for three years. Assuming no loss to follow-up and a single analysis at the end of the four-year study, the original design, developed from calculations based on an asymptotic formula for the binomial test of Halperin et al. [29], required a total of 4200 patients recruited during a two-year period to attain a power of 90%.

We applied our program to the design of this trial. Using 500,000 replications for the fixed-duration design as in the original protocol gave 87.50% power (with standard error of 0.05%) for the logrank test and 88.19% power for the Peto-Prentice test. We chose 500,000 replications to give tight confidence bounds on the power. Although [28] considers the binomial test and uses asymptotic normality to compute the power, the 90% power obtained there is in good agreement with the preceding simulation results on the power of the logrank and generalized Wilcoxon tests.

More importantly, we can now simulate the power for group sequential designs using our program. As indicated in the protocol, the data monitoring committee of BHAT was scheduled to meet semi-annually to review the trial. Adopting the Lan-DeMets approach with the O'Brien-Fleming use function, we simulated the trial with the same parameters (baseline, alternative, accrual pattern, noncompliance rates, etc.) as those in the protocol. In the simulation, the data analyses were done every 6 months except for the first (the first interim analysis was at 12 months for a total of 7 analyses). After 500,000 simulations, the program gave a 87.65% power for the logrank (with standard error of 0.05%) and 88.14% power for the Peto-Prentice statistic. We also used the program to simulate the Type I error of the group sequential design to check the adequacy of the normal approximation that gives the nominal 5% Type I error. Under the same parameters for the survival distribution of the control group, accrual, noncompliance, etc., we found the Type I error to be 4.99% based on 500,000 simulations.

Two interesting points are revealed by the results of this simulation study. First, the average duration of the group sequential trial under the alternative hypothesis in the original protocol is 37.6 months for the logrank test, which is very close to the actual duration of the trial (40 months), and is about 10 months shorter than
the 4-year fixed-duration study. On the other hand, there is no loss in power of the group sequential logrank test (87.65 %) compared with the fixed-duration logrank test (87.50 %), analogous to the situation in Example 4 of the last section. The second point is that despite the proportional hazards alternatives assumed, the fixed-duration (and also the group sequential) logrank test is somewhat less powerful than the Peto-Prentice generalized Wilcoxon test. These two phenomena can both be explained by noncompliance. As pointed out in the paragraph following Example 4, the logrank statistic is no longer asymptotically efficient under proportional hazards alternatives when there is noncompliance to perturb the actual hazards.

DISCUSSION

The feature of simultaneously simulating several statistics at a time makes our program somewhat more complicated than that in [2], which is designed for the user to simulate either the logrank or Gehan's statistic. This complication is worthwhile because of the savings in the number of simulations in situations where one wants to compare several test statistics during the design stage of a trial to decide which one to use. Our simulation program also enables one to evaluate the effect of complex factors such as noncompliance, loss to follow-up, patient accrual patterns, non-proportional hazards alternatives, etc., in the design of a clinical trial with a failure-time endpoint and interim analyses. It can be used to determine the power of the test and sample size to attain some desired power under various scenarios at the design stage.

A recent review [30] of commercial software packages for sample size and power calculations in the design of clinical trials with failure-time endpoints shows that they are based on statistical methods in the 1980s and have not incorporated interim analysis strategies. Another recent review [31] of software packages for group sequential methods shows that their design modules consider relatively simple trials (such as exponential survival models, no noncompliance) and are restricted to logrank statistics and simple stopping rules that were developed in 1980s for survival data. In contrast, our program is based on recent advances in group sequential methods for survival data, and the underlying statistical methods for interim analysis of survival data are described in the
first three sections of the paper.

Unlike the software packages reviewed in [31], our program does not have an analysis module that can be applied directly for interim analysis of actual clinical trial data. Although the main components of the program can serve as basic building blocks for an analysis module, we are still working towards a user interface that is flexible enough to accommodate different forms of input data, and we are awaiting feedback from users of the program whose current version was recently used to aid the design of a cancer prevention trial at the National Cancer Institute.

REFERENCES


The research of the first author was supported by the National Science and Engineering Research Council of Canada, and that of the second author was supported by the National Science Foundation. Part of the work was done while the first author was on leave from McGill University, visiting the Office of Biostatistics Research, National Heart, Lung and Blood Institute. The authors are grateful to Dr. Nancy Geller, Dr. Dean Pollmann and Dr. Joanna Shih, who have provided many stimulating discussions and valuable suggestions.