Determination of Power and Sample Size in the Design of Clinical Trials with Failure-time Endpoints and Interim Analyses

Minggao Gu[1], **Ph.D** and **Tze Leung Lai, Ph.D**

Department of Statistics, The Chinese University of Hong Kong, Hong Kong 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 probability and to assess the performance of different interim analysis strategies. It also presents two 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.

[1]Mailing Address: Dr. Minggao Gu, Department of Statistics, The Chinese University of Hong Kong, Shatin, NT, Hong Kong

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 probability under various scenarios of baseline survival, censoring pattern, noncompliance and accrual rate. 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 in [2] another simulation program for these test statistics incorporating group sequential testing using either the Pocock [3] or the O'Brien-Fleming [4] boundary. Their program can be used to compute both the Type I error probability 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 probability 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, Lakatos [5, 6] and Shih [7] also developed sample size calculation programs for a fixed-duration trial. Instead of using Monte-Carlo simulations, their programs are based on Markov models and normal approximations to the distributions of the rank statistics.

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". Earlier, Slud and Wei [10] proposed to simply decompose the Type I error probability α as $e_1 + \cdots + e_k$ and to choose a boundary so that the test statistic has a probability of e_j of first crossing the boundary at the jth interim analysis under the null hypothesis.

In this paper we present two new simulation programs that broaden 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, and Haybittle-type boundaries, plus any other boundary specified by the user. The Haybittle-type boundary, which will be described in detail in the next section, is based on refinements of an idea due to Haybittle [11], extending it to failure-time data. Secondly, while [2] does not incorporate patient withdrawal (censoring) and noncompliance (crossover), our programs allow the user to incorporate these considerations by specifying the censoring distributions and crossover rates of the two treatment groups. Thirdly, our programs give the user an additional way to specify 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^ρ statistics introduced by Harrington and Fleming [17], which in turn belong to the Beta family of statistics proposed by Self [18]. Our programs allow the user to choose the parameters ρ and τ 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 the recent developments 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 the two computer programs implementing these methods, which are available on the World Wide Web. The first program evaluates the power of a test under the specified baseline and alternative distributions for a given sample size, while the second program finds the sample size of a test having the prespecified power at given baseline and alternative distributions. In the final two sections, we first give some examples to demonstrate the accuracy of the programs and then present a case study in applying the methods in the previous sections to the design of a myocardial infarction trial.

METHODS FOR GENERATING BOUNDARIES

In this section we first give some background and explanation of the Slud-Wei and Lan-DeMets boundaries, and then describe an alternative method to generate the stopping boundary, which we shall refer to as the Haybittle-type boundary as it is based on modifications and refinements of an idea due to Haybittle [11]. These three types of boundaries are available as options in our programs, 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 ith interim analysis, for which the boundary b_i is determined by

$$e_i = \mathrm{pr}(|W_1| \leq b_1\sqrt{V_1}, \ldots, |W_{i-1}| \leq b_{i-1}\sqrt{V_{i-1}}, W_i > b_i\sqrt{V_i}), \tag{1}$$

where W_i denotes the test statistic at the ith interim analysis and V_i denotes the corresponding variance estimate. Since our two programs take 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 the recursive numerical integration method of Armitage, McPherson and Rowe [19]. Formula (5.2) of [20] is used by our programs for the variance estimate V_i. It is shown in [20] to provide a consistent estimate under the null hypothesis and local alternatives.

In the Lan-DeMets [8] approach to determine the stopping boundary, the user specifies a non-decreasing function $u : [0, 1] \to [0, \alpha]$, where α is a prescribed Type I error and $u(0) = 0$, $u(1) = \alpha$. Concerning how u should be chosen, see [8, 9, 21, 22]. Let v be a prespecified positive number representing the maximum allowable null variance of the test statistic. The user who has difficulties specifying v can simply use our programs' default option that sets v to be $V(t_k, \rho, \tau)$ defined by (4) in the next section. At the ith interim analysis, stopping occurs and the null hypothesis is rejected if $|W_i| \geq b_i\sqrt{V_i}$, where b_i is determined by the equation

$$u\left(V_i/v\right) = \mathrm{pr}(|W_j| \geq b_j\sqrt{V_j}, \text{ for some } j \leq i). \qquad (2)$$

The study also stops 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 $\alpha(= u(1))$. As pointed out by Lan and DeMets [9], the V_i can be regarded as the "observed information" up to the calendar time of the ith interim analysis, and $u(V_i/v)$ represents how much of α is used (or spent) for this fraction of the maximum information v. In our programs, the user can choose one of three 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 $u(t) = \alpha \log\{1 + (e-1)t\}$, and the linear use function $u(t) = \alpha t$ considered by Kim and DeMets [22], where Φ denotes the standard normal distribution function and $1 - \Phi(z_{\alpha/2}) = \alpha/2$.

The use function approach for survival data requires pre-specification of 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 [23] a simple but statistically 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 θ of independent normal observations $Z_1, Z_2 \ldots$ is 0, assuming that $var(Z_i) = 1$ is known. The Neyman-Pearson test rejects $\theta = 0$ at level α 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 α 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 [23] the following procedure involving a maximum of k repeated significance tests. First determine b such that $\mathrm{pr}(|\sum_{i=1}^{j} Z_i| \geq$

$b\sqrt{j}$, for some $j \leq k-1$) is less than $\alpha/3$, where the probability is evaluated under the null hypothesis $\theta = 0$. With b thus chosen we have the following Haybittle-type boundary for the test statistics W_i of the survival data that have appeared in (1) and (2). The boundary points for the first $k-1$ interim analyses are $b\sqrt{V_j}$, $j = 1, \cdots, k-1$, while the last boundary point is $c\sqrt{V_k}$ so that

$$\mathrm{pr}(|W_j| \geq b\sqrt{V_j} \text{ for some } j \leq k-1, \text{ or } |W_k| \geq c\sqrt{V_k}) = \alpha. \tag{3}$$

Our programs automatically solve for c after the user provides the input parameter b. The user actually does not need to solve carefully for b, and can use the default option of typing 0 for its value. In fact, for any input $b \leq 1$, the programs reset b as 3 to be as the user's input parameter. They first compute $\mathrm{pr}(|\Sigma_{i=1}^{j} Z_i| \geq b\sqrt{j}$, for some $j \leq k-1$). If this probability is less than $\alpha/3$, then the input parameter is taken as it is. Otherwise b is reset as $b + 0.1$, and the probability is recomputed to check if it is less than $\alpha/3$. This procedure is repeated until the probability indeed falls below $\alpha/3$. The output shows the user the actual value of b that is finally chosen.

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 $z_{\alpha/2}$ as described above, the Haybittle-type boundary is able to avoid both (i) and (ii). As shown in Section 4 of [23] and in Examples 1 and 5 below, typical values of b range from 2.6 to 3 when $\alpha = .05$. Although c is adjusted upwards from $z_{\alpha/2}$ in (3) to give an overall significance level α, the upward adjustment is usually small since $\mathrm{pr}(|W_j| \geq b\sqrt{V_j}$ for some $j \leq k-1$) is typically only a small fraction of α.

In addition to the Slud-Wei, Lan-DeMets and Haybittle-type methods, our programs also give one the option of specifying one's own boundary. When there are k interim analyses, the user-specified boundary consists of the boundary values $b_1, \ldots, b_k$ (for $W_i/\sqrt{V_i}$ in a one-sided test and for $|W_i/\sqrt{V_i}|$ in a two-sided test). The power simulation program can be used to check if the Type I error of the test with this boundary is indeed close to α; see Example 3 below.

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. It consists of weighted logrank statistics with weight $\hat{S}^{\rho}(x)(1-\hat{S}(x))^{\tau}$ at each uncensored failure time x, where $\hat{S}$ is the Kaplan-Meier survival curve of the combined sample. In the case $\tau = 0$, it reduces to the G^{ρ} 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 $\tau = 0$). Our programs allow the user to choose any member of this family by specifying the values of ρ and τ.

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

$$V(t_i, \rho, \tau) = -\frac{1}{2}\int_0^{t_i} S(u)^{2\rho}(u)(1-S(u))^{2\tau} w(u)\#(t_i - u)dS(u), \tag{4}$$

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) 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.

SPECIFICATION OF BASELINE AND ALTERNATIVE DISTRIBUTIONS

In our programs, 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 $S(t_1) > \cdots > S(t_l) > 0$. 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 $t_0 = 0$

and $S(0) = 1$. For $t > t_l$, the value of $\log S(t)$ is determined by the linear extrapolation of the values specified for the interval $[t_{l-1}, t_l]$. In particular, to specify an exponential survival function with $S(t^*) = .5$, one set $l = 1$, $t_1 = t^*$ 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 programs. 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 $[s_{i-1}, s_i)$, $i = 1, \cdots, p+1$ ($s_0 = 0$, $s_{p+1} = \infty$). 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

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

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 $t^* > 0$.

SPECIFICATION OF NONCOMPLIANCE RATES

The noncompliance rates, or crossover rates, are specified by the drop-out rate (from treatment group to control group) and drop-in rate (from the control group to treatment group); see [24]. They govern the proportion of subjects in each specified time period who change their lifetime distribution to that of the other group. In our programs, 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 crossover are assumed to occurred only. The caution is that the crossover rate at each of these points (which are endpoints of subintervals) cannot be specified by *rate for the time interval/number of subintervals* since the correct formula is

$$r_1 + (1 - r_1)r_2 + \ldots + (1 - r_1) \cdots (1 - r_{k-1})r_k = r, \tag{6}$$

where r is the rate on the entire interval (assuming continuous crossover) and $r_1, \ldots, r_k$ are the crossover rates of its k consecutive subintervals. In particular, if $r_1 = \cdots = r_k$, then $r_1 = 1 - (1-r)^{1/k}$.

THE PROGRAM FOR SIMULATING POWER

The program *dssp2.exe* (Design of Survival Studies: Power, Version 2) has been developed to simulate power and expected duration of a test. Its main components 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 linear.

The subroutine *probbdy* determines the current boundary point b_j at each time of interim analysis. For two sided-tests, let $f_1(x) = \phi(x/\sqrt{V_1})/\sqrt{V_1}$, and for $j = 2, \cdots, k$, define recursively

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

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 the Gaussian quadrature [25] 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 *rinvsu*. The value of the inverse of the alternative survival function is returned by *hinvsu*, 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 *mcalstat* 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 DOS executable version of the program and the sample input file are available on the World Wide Web (http://www.sta.cuhk.edu.hk/minggao/ctrials.htm). To run the program, first download the files *dssp2.exe* and *dssp2.inp* and then type

dssp2.exe < *dssp2.inp* at the DOS command line. The sample input file *dssp2.inp* contains detailed instructions on how to change the input parameters. In addition, there is a Windows version of the program at the above Web site, where further documentation and illustrative examples can also be found. For those who prefer to run the program in a UNIX environment, the FORTRAN source code *dssp2.f* is also available at the same Web site. However, the user has to compile the code to produce a UNIX executable version of the program. If the UNIX executable file is named *a.out*, then type *a.out* < *dssp2.inp* to start running the program. The user who encounters difficulties can send electronic mail to minggao@cuhk.edu.hk for help.

THE PROGRAM FOR DETERMINING SAMPLE SIZE

Also available at the above Web site is another program *dsss1.exe* (Design of Survival Studies: Sample size, Version 1), which finds the sample size of a test having the prespecified power at given baseline and alternative survival distributions. It uses the main components of *dssp2.exe* together with an iterative procedure, which will be described below, to solve for the sample size. The DOS executable version of the program, together with the input file *dsss1.inp*, and the FORTRAN source code *dsss1.f* for running it in a UNIX environment are available at the same Web site, where further documentation can be found.

Since the sample size is defined as the product of the accrual period (number of years in which accrual occurs) and the accrual rate (number of subjects accrued per year). Therefore, the user of *dssp2.exe* inputs both the accrual period and the accrual rate to specify the sample size of a trial. On the other hand, to determine the sample size that yields some desired power P^*, the user of *dsss1.exe* has to input the accrual period and upper and lower bounds on the accrual rate. There are usually practical constraints on the sample size in clinical trials. For example, budget considerations would lead one to a natural upper bound on the sample size, while publication and administrative considerations would lead one to a natural lower bound on the sample size. After dividing by the accrual period, one therefore obtains upper and lower bounds, U and L, on the accrual rate.

The iterative procedure used by the program begins with computing the power $p(U)$ at accrual rate U by simulation. If $p(U) \leq P^*$, the procedure stops. The output then gives the sample size (product of U and the accrual length) if $p(U)$ differs from P^* by no more than twice the standard error of the simulations (abbreviated by se below), or gives the power $p(U)$ at accrual rate U otherwise so that the user can then input a larger upper bound on the accrual rate and rerun the program to find the required sample size.

Let $\delta = (L/U)^{1/3}$. If $p(U) > P^*$, let $L' = \delta U$ (which is larger than L). The program then computes the power $p(L')$ at accrual rate L' by simulation. If $p(L')$ still exceeds P^*, change L' to $L'' = \delta L'$, otherwise reset L as L'. If $p(L')$ and $p(L'')$ both exceed P^*, change L'' to $L(= \delta L'')$. If $p(L') > P^* \geq p(L'')$, reset L as L''. In the event $p(L')$, $p(L'')$ and $p(L)$ all exceed P^*, the procedure stops and the output tells the user the power $p(L)$ at accrual rate L. The user can then either give a smaller lower bound on the accrual rate to rerun the program, or simply replace P^* by the higher power $p(L)$. The procedure also stops if the simulated power $p(L)$ at accrual rate L (possibly redefined as L' or L'') satisfies $P^* \geq p(L) \geq P^* - 2se$.

If the procedure does not stop in this first phase, it has bracketed the desired accrual rate between U and L so that $p(L) < P^* < p(U)$. The program then uses successive secant approximations to find the desired sample size. Treating the accrual rate as a continuous variable x, let $\widehat{p}(x)$ be a linear function with $\widehat{p}(L) = p(L)$ and $\widehat{p}(U) = p(U)$, and let $\widehat{x}$ be the root of $\widehat{p}(x) = P^*$. Since $p(L) < P^* < p(U)$, $\widehat{x}$ lies between L and U. The program computes the power $p(\widehat{x})$ at accrual rate $\widehat{x}$ by simulation. It resets L as $\widehat{x}$ if $p(\widehat{x}) < P^*$, or resets U as $\widehat{x}$ if $p(\widehat{x}) > P^*$, and repeats the procedure. The iterations terminate when $p(\widehat{x})$ differs from P^* by no more than $2se$ so that $\widehat{x}$ can be used as the desired accrual rate, or when a maximum number N of iterations is reached. The program chooses $N = 20$. The program output gives the product of $\widehat{x}$ and the accrual period as the sample size, rounding off to the nearest integer. However, if the procedure terminates with N iterations without convergence to the desired accrual rate, the output tells the user about this and the new upper and lower bounds U and L at the end of the N iterations, together with the values $p(L)$ and $p(U)$, so that the user can either rerun the program with these much tighter upper and lower bounds on the accrual rate or stop with a somewhat higher (or lower) power by using the accrual rate U (or L).

This iterative scheme typically converges to the desired accrual rate after a few iterations. In the unlikely event that it does not converge after $N = 20$ iterations, the output still gives the user valuable information, providing tighter bounds U and L and the power at either bound. Since the program uses the secant method only after the bracketing $p(L) < P^* < p(U)$ is achieved, successive iterations also preserve this bracketing and give U, L values shrinking towards the desired accrual rate.

EXAMPLES

Example 1. As an illustration of the accuracy of our program for simulating power, we compare the results on power and expected study duration obtained by the program with those reported by Siegmund [26] in a simulation study of the performance of his modification, for logrank statistics, of Haybittle's 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 $t = 1, 1.5, \cdots, 5.5$ (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 [26].

Example 2. The program *dssp2.exe* 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 [5] 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 [5].

Table 1. Simulated Power on CVD and Cancer Trials

	CVD Trial		Cancer Trial	
Sample Sizes	2654	2651	164	169
Power	.93	.93	.88	.89

Example 3. By inputting the hazard ratio to be constantly equal to 1, we can use the program *dssp2.exe* to compute the Type I error probability 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 failure rate of the control group is 0.1 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 thus chosen has a Type I error probability of 5 % based on normal approximation. We can also determine this Type I error probability by Monte Carlo simulation under the specified distribution for the control group, assuming that the distribution for the treatment group also has this specified distribution. We used the program to compute from 50,000 simulations the Type I error probability, which was found 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 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 the program *dssp2.exe* using 10,000 simulations are reported in the following table.

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

	Fixed Duration (5.5 years)	Sequential Design	
	Power	Power	Expected Duration
No Noncompliance	.961	.938	3.2
Noncompliance (.3)	.492	.499	4.3
Noncompliance (.4)	.355	.392	4.5
Noncompliance (.5)	.252	.305	4.7

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 [23], 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).

Example 5. In this example, we demonstrate how program *dsss1.exe* can be used to determine the sample size in designing a clinical trial with 5% Type I error and 90% power at a specified alternative. Similar to Example 3, we consider a 5-year clinical trial with uniform accrual in the first 3 years. The interim analyses are scheduled at 1, 1.5, ..., 4.5, 5 years for a total of 9 interim analyses. The Haybittle method is selected with $b = 3$. The baseline survival function is the exponential survival function with 50% failure in 5 years. For the alternative survival function, we assume that the failure rate is 4% for the first year, 8% for the second and for the third year, and 10% for the fourth and and for the fifth year. We also assume that there is 7.7% censoring (loss to follow-up) for the control group and 7.3% censoring for the treatment group in the 5-year period. In addition, we assume that there is a 3% drop-in rate per year together with a 10% drop-out rate per year, and that these cross-overs occur at the beginning of each year. The logrank statistic is used.

We enter the above specifications of the trial into the sample input file *dsss1.inp* by modifying the appropriate entries. The modified sample input file *exam1.inp* and the output file *exam1.log*, together with other input and output files in the last part of this example, can be found on the same Web page where the files *dsss1.exe* and *dsss1.inp* are. An important input is in line 88 of the input file, where one specifies the accrual length (in years) and the lower and upper bounds of the accrual rate (number of subjects per year). In this example, the accrual length is 3 years. We initially used a lower bound of 300 and upper bound of 500 for the accrual rate. Another important input parameter is the number of simulations used in each iteration, located in line 37 of the input file. We recommend to start with a moderate number, 500, so that the program can finish fast. If one wants a smaller standard error for the simulated power, one can rerun the program later using a larger number of simulations per iteration but tighter lower and upper bounds for the accrual rate based on the outputs of the initial run.

It took a PC with Pentium II 300 processor less than 1 minute to get the result.

The program printed out a message stating that the lower bound on the accrual rate is too large for the desired power. At iteration 1, with accrual rate 500, the simulated power is 1. At iteration 4, with accrual rate 300, the simulated power is .94. Since 300 is the lower bound on the accrual rate, the program stopped without finding the accrual rate for the desired power. Based on the output of the program showing the simulated powers for accrual rates of 300 and 400, we changed the lower and upper bounds for the accural rate to 200 and 400 in the second run of the program. This time, after about 40 seconds of running time on the PC, the program stopped with an accrual rate of 252.0 to yield a simulated power .892, after reaching within 2 standard errors of the desired power of .9. The input and output files of this second run are called *exam2.inp* and *exam2.log*.

In order to get more accurate results, we undertook a third run of the program with 5000 simulations per iteration, after changing the lower and upper bounds on the accrual rate to 200 and 300. This time, after about 5 minutes of running time on the PC, the program found the accrual rate to be 262.1 with a simulated power .904 for the logrank statistic. The input and output files of this final run are called *exam3.inp* and *exam3.log*. To find the sample size for the Peto-Prentice statistic, we changed the input in line 30 of *exam3.inp* from “0” to “1”, and call the new input file *exam4.inp*. It took another 5 minutes of running time on the PC to generate the output file *exam4.log*, which gives an accrual rate of 228.9 with a simulated power of .906, showing the Peto-Prentice statistic to be more efficient than the logrank statistic in this situation.

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%.

Applying our program *dssp2.exe* with 500,000 replications to 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 the 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 statistic (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 % (with standard error of 0.1 %) based on 50,000 simulations.

Two interesting points arise from 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 5 of 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. In fact, the logrank statistic is no longer asymptotically efficient under proportional hazards alternatives when there is noncompliance to perturb the actual hazards that are relevant to the test statistic.

DISCUSSION

Our simulation programs enable 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. They can be used to choose efficient test statistics and to determine the power of the test or the 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 showns 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 programs are 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 for simulating power does not have an analysis module that can be applied directly for interim analysis of actual clinical trial data. Whereas its main components can serve as basic building blocks for an analysis module, we are still working towards a user interface that is easy to use and flexible enough to accomodate different forms of input data.

REFERENCES

1.Halpern, J., Brown, B. Wm J. Designing clinical trials with arbitrary specification of survival functions and for the logrank or generalized Wilcoxon test. *Controlled Clin Trials* 1987; 8: 177-189.

2.Halpern, J., Brown, B. Wm J. A computer program for designing clinical trials with arbitrary survival curves and group sequential testing. *Controlled Clin Trials* 1993; 14: 109-122, 1993.

3.Pocock, S. J. Group sequential methods in the design and analysis of clinical trials. *Biometrika* 1977; 65: 191-199.

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

5.Lakatos, E. Sample size determination in clinical trials with time-dependent rates of losses and noncompliance. *Controlled Clin Trials* 1986; 7: 189-199.

6.Lakatos, E. Sample size based on the logrank statistic in complex clinical trials. *Biometrics* 1988; 44: 229-241.

7.Shih, J. H. Sample size calculation for complex clinical trials with survival endpoints. *Controlled Clin Trials* 1995; 16: 395-407.

8.Lan, K. K. G., DeMets, D. L. Discrete sequential boundaries for clinical trials. *Biometrika* 1983; 70: 659-663.

9.Lan, K. K. G., DeMets, D. L. Group sequential procedures: calendar versus information time. *Statisti. Med.* 1989; 8: 1191-1198.

10.Slud, E., Wei, L. J. Two-sample repeated significance tests based on the modified Wilcoxon statistic. *J. Am. Statist. Assoc.* 1982; 77: 862–868.

11.Haybittle, J. L. Repeated assessments of results in clinical trials of cancer treatment. *British J. Radiology* 1971; 44: 793-797.

12Gehan, E. A generalized Wilcoxon test for comparing arbitrarily singly censored samples. *Biometrika* 1965; 52: 203-223.

13.Lan, K. K. G., Wittes, J. T. Rank tests for survival analysis: a comparison by analogy with games. *Biometrics* 1985; 41: 1063–1069.

14.Gu, M. G., Lai, T. L., Lan, K. K. G. Rank tests based on censored data and their sequential analogues. *Amer. J. Math. & Management Sci.* 1991; 11: 147–176.

15.Peto, R., Peto, J. Asymptotically efficient rank test procedures (with discussion). *J. Roy. Statist. Soc. Ser. A* 1972; 135: 185-206.

16.Prentice, R. L. Linear rank tests with right censored data. *Biometrika* 1978; 65: 167-179.

17.Harrington, D.P., Fleming, T.R. A class of rank test procedures for censored survival data. *Biometrika* 69: 553-566, 1982.

18.Self, S. G. An adaptive weighted logrank test with application to cancer prevention and screening trials. *Biometrics* 1991; 47: 975–986.

19.Armitage, P., McPherson, C. K., Rowe, B. C. Repeated significance tests on accumulating data. *J. Roy. Statist. Soc. Ser. A* 1969; 135, 235–244.

20.Gu, M. G., Lai, T. L. Weak convergence of time-sequential censored rank statistics with applications to sequential testing in clinical trials. *Ann. Statist.* 1991; 19: 1403-1433.

21.Li, Z., Geller, N. L. Changing frequency of interim analysis in sequential monitoring. *Biometrics* 1989; 45: 1017-1020.

22.Kim, K., DeMets, D. L. Design and analysis of group sequential tests based on the type I error spending rate function. *Biometrika* 1987; 74: 149-154.

23.Gu, M. G., Lai, T. L. Repeated significance testing with censored rank statistics in interim analysis of clinical trials. *Statistica Sinica* 1998; 8: 411-428.

24.Wu, M., Fisher, M., DeMets, D. L. Sample sizes for long-term medical trials with time-dependent noncompliance and event rates. *Controlled Clin Trials* 1980; 1: 109-121.

25.Abramowitz, M., Stegun, I. A. *Handbook of Mathematical Functions.* Washington, D.C.: National Bureau of Standards, 1964.

26. Siegmund, D. *Sequential Analysis: Tests and Confidence Intervals.* New York, NY: Springer-Verlag, 1985.

27. DeMets, D. L., Hardy, R., Friedman, L. M., Lan, G. K. K. Statistical aspects of early termination in the Beta-Blocker Heart Attack Trial. *Controlled Clin Trials* 1984; 5: 362-372.

28. Beta-Blocker Heart Attack Trial Research Group: *Beta-Blocker Heart Attack Trial: design, methods, and baseline results.* Bethesda, MD: Clinical Trial Branch, National Heart, Lung, and Blood Institute, 1982.

29. Halperin, M., Rogot, E., Gurian, J., Ederer, F., Sample size for medical trials with special reference to long-term therapy. *J. Chronic Disease* 1968; 21: 13-24.

30. Iwane, M., Palensky, J., Plante, K. A user's review of commercial sample size software for design of biomedical studies using survival data. *Controlled Clin Trials* 1997; 18: 65-93.

31. Emerson, S.S. Statistical packages for group sequential methods. *Am. Stat.* 1996; 50: 183-192.

This research was supported by the National Science and Engineering Research Council of Canada and the National Science Foundation. Part of the work was done while the first author was on sabbatical leave visiting the Office of Biostatistics Research, National Heart, Lung and Blood Institute, where Drs. Nancy Geller, Dean Follmann and Joanna Shih had provided many stimulating discussions and valuable suggestions. The authors are also grateful to the referees and to Dr. Farid AitSahlia of Hewlett-Packard Laboratories for their helpful comments on the manuscript and the documentation of the programs.