STATISTICS IN MEDICINE
Statist. Med. 2001; **20**:2055–2063 (DOI: 10.1002/sim.843)

Predicting analysis times in randomized clinical trials

Emilia Bagiella and Daniel F. Heitjan*,†

*Division of Biostatistics, Mailman School of Public Health, Columbia University, 622 W. 168th Street,
New York, NY 10032, U.S.A.*

SUMMARY

Randomized clinical trial designs commonly include one or more planned interim analyses. At these times an external monitoring committee reviews the accumulated data and determines whether it is scientifically and ethically appropriate for the study to continue. With failure-time endpoints, it is common to schedule analyses at the times of occurrence of specified landmark events, such as the 50th event, the 100th event, and so on. Because interim analyses can impose considerable logistical burdens, it is worthwhile predicting their timing as accurately as possible. We describe two model-based methods for making such predictions during the course of a trial. First, we obtain a point prediction by extrapolating the cumulative mortality into the future and selecting the date when the expected number of deaths is equal to the landmark number. Second, we use a Bayesian simulation scheme to generate a predictive distribution of milestone times; prediction intervals are quantiles of this distribution. We illustrate our method with an analysis of data from a trial of immunotherapy in the treatment of chronic granulomatous disease. Copyright © 2001 John Wiley & Sons, Ltd.

1. INTRODUCTION

Randomized clinical trials commonly include one or more scheduled interim analyses. At each such analysis, the trial staff prepares a detailed report on results to date and executes a formal test of the primary hypothesis. A data and safety monitoring board (DSMB), comprising experts in biostatistics, medical ethics and relevant clinical areas, reviews the data and judges whether the weight of evidence is sufficient to stop the trial [1]. Conducting such an analysis can be a significant effort and expense, and consequently it is essential that such meetings take place only when they have the potential to materially influence the trial's operation or conclusion.

When the primary outcome is time to an event, the statistical information in the data is proportional to the number of events that have occurred. Consequently, in calculating the sample size and designing an interim analysis scheme, it is common to plan in terms of numbers of events. Thus, one might determine that 100 events are necessary to achieve a

*Correspondence to: Daniel F. Heitjan, Division of Biostatistics, Mailman School of Public Health, Columbia University, 622 W. 168th Street, New York, NY 10032, U.S.A.

†E-mail: dfh5@columbia.edu

Contract/grant sponsor: USPHS; contract/grant numbers: CA-13696, HL-53968

Received July 1999
Accepted August 2000

target level of power, and schedule three interim analyses after the first 25, 50 and 75 events, with the final analysis after the 100th event. Although modest departures from the analysis schedule will typically have little effect on the statistical properties of the interim analysis procedure, for the sake of the trial's credibility it is generally preferable to adhere to the schedule as closely as possible.

Predicting the times of interim and final analyses has long been recognized as an important practical aspect of trial design. Rubinstein *et al.* [2] and Lachin and Foulkes [3] have proposed methods for projecting the length, sample size and number of events from *a priori* estimates of accrual, event and censoring rates. One can readily adapt their formulae to project interim analysis landmark times. Unfortunately, pilot data are often limited in quantity and poor in quality, and consequently these predictions can be seriously in error. Moreover, the existing methods give only point predictions that do not reflect the uncertainty of estimation and prediction.

With the costs at stake, it seems natural to use accumulating data from the trial itself to make the most accurate possible predictions of the times of interim analysis milestones. In this article we use a simple statistical model to make two separate projections: a point prediction that selects the calendar time when the expected number of deaths, given experience to date, equals the milestone number, and a Bayesian interval prediction based on simulation of future accrual and event dates. With the method one can easily produce projections for any milestone at any calendar time. We demonstrate our method with an analysis of data from a trial of immunotherapy for the treatment of chronic granulomatous disease [4].

2. STATISTICAL MODEL

2.1. Notation

We are conducting a clinical trial comparing k treatment arms. Assume that the trial commenced enrolment at calendar time 0. Denote by $N_j(t)$, $j=1,\ldots,k$ the number of subjects enrolled by calendar time t in treatment group j, and let $N(t)=\sum_j N_j(t)$ be the total number enrolled by calendar time t. For the $N_j(t)$ patients enrolled in group j by time t, let the enrolment times be $e_{ji}<t$ and the failure times (possibly censored) be s_{ji}. We measure enrolment times from the opening of the study, and failure times from the date of a subject's enrolment. Let $Y_{ji}(t)$ take the value 1 if patient i in group j is under observation and at risk for failure at time t, and 0 otherwise. Let $D_j(t)$ be the number of patients in treatment group j who failed on study by calendar time t, and $D(t)=\sum_{j=1}^k D_j(t)$ be the total number of observed failures by time t. Let $C_j(t)$ be the number of patients in treatment group j who were lost to follow-up without experiencing an event by time t, and $C(t)=\sum_{j=1}^k C_j(t)$ be the total number of losses by time t.

At the current calendar time, t_0, we are interested in predicting the future course of the trial. Define $Q_j(t_0,t)$ to be the conditional expectation, given the data up to time t_0, of the number of patients in group j who are in the study and alive at time t_0 who will be observed to fail by calendar time $t>t_0$, and define $Q(t_0,t)=\sum_{j=1}^k Q_j(t_0,t)$. Similarly, define $R_j(t_0,t)$ to be the expected number of patients who will enrol in group j and fail on study between times t_0 and t, and let $R(t_0,t)=\sum_{j=1}^k R_j(t_0,t)$. Defining $ED(t_0,t)$ to be the conditional expectation, given the data up to time t_0, of the number of events that will have been observed by time

$t > t_0$, we have that

$$ED(t_0, t) = D(t_0) + Q(t_0, t) + R(t_0, t) \tag{1}$$

2.2. *Theory for a general survival distribution*

Our objective is to predict the calendar time $T^\star$ at which the $D^\star$th event will occur, that is, the minimum $T^\star$ such that $D(T^\star) = D^\star$. Assume that the event times of the patients in the k arms are independent with distributions $F_j(t)$ and densities $f_j(t)$, and that the censoring times in group j are also independent of each other and of the failure times, and have distributions $G_j(t)$ and densities $g_j(t)$. Then the conditional expectation of the number of patients already enrolled in group j and alive and on study at time t_0 who will have been observed to fail by time t is

$$Q_j(t_0, t) = \sum_{i=1}^{N_j(t_0)} Y_{ji}(t_0) \frac{[F_j(t - e_{ji}) - F_j(t_0 - e_{ji})] - \int_{t_0 - e_{ji}}^{t - e_{ji}} G_j(u) f_j(u) \mathrm{d}u}{[1 - F_j(t_0 - e_{ji})][1 - G_j(t_0 - e_{ji})]} \tag{2}$$

$R(t_0, t)$ represents the expected number of patients who will both enrol and experience a failure on study between times t_0 and t. We assume that subjects enrol in the study according to a Poisson process with intensity μ, so that the expected number of new patients who will enrol during a time interval of length s is μs. Assuming that a proportion $1/k$ of patients are to be enrolled in each arm, the expected number who will enrol in group j after t_0 and experience a failure on study before calendar time t is

$$R_j(t_0, t) = \mu k^{-1} \int_0^{t - t_0} \left\{ \int_0^{t - t_0 - u} f_j(s)(1 - G_j(s))\, \mathrm{d}s \right\} \mathrm{d}u \tag{3}$$

2.3. *Theory for exponential survival*

Suppose that the failure times are exponentially distributed with rates λ_j, $j = 1, \ldots, k$, so that $F_j(u) = 1 - \exp(-\lambda_j u)$. Suppose also that the censoring times are exponentially distributed with rates v_j, $j = 1, \ldots, k$, so that $G_j(u) = 1 - \exp(-v_j u)$. Then equation (2) becomes

$$Q_j(t_0, t) = \sum_{i=1}^{N_j(t_0)} Y_{ji}(t_0) \frac{\lambda_j [1 - \exp(-(\lambda_j + v_j)(t - t_0))]}{\lambda_j + v_j} \tag{4}$$

and equation (3) becomes

$$R_j(t_0, t) = \frac{\mu \lambda_j}{k(\lambda_j + v_j)} \left[(t - t_0) - \frac{1 - \exp(-(\lambda_j + v_j)(t - t_0))}{\lambda_j + v_j} \right] \tag{5}$$

Our model generalizes the model of Rubinstein *et al.* [2] and Lachin and Foulkes [3], who assumed a trial of length $V + W$ time units, where accrual takes place only during the first V units. To see this, define $A_j(t_0, t)$ to be the expected number of subjects enrolled in group j between times t_0 and t who are still alive and on study at time t, and $P_j(t_0, t)$ to be the probability that a subject in group j who is at risk at time t_0 will have had an event on study by time t. Then the expected total number of observed events in arm j in the Rubinstein *et al.* design is

$$E[D_j] = R_j(0, V) + A_j(0, V) P_j(V, W) \tag{6}$$

Straightforward integration gives

$$A_j(0,V)=\frac{\mu}{k(\lambda_j+v_j)}[1-\exp(-(\lambda_j+v_j)V)]$$

and

$$P_j(V,W)=\frac{\lambda_j}{\lambda_j+v_j}[1-\exp(-(\lambda_j+v_j)W)]$$

Substituting these expressions into (6) gives equation (A2) of Rubinstein *et al.* [2].

3. POINT PREDICTION OF THE ANALYSIS LANDMARK

A straightforward point prediction of the time $T^{\star}$ by which $D^{\star}$ patients will have experienced an event is the solution (in the argument t) of the equation

$$D^{\star}=\widehat{ED}(t_0,t)=D(t_0)+\hat{Q}(t_0,t)+\hat{R}(t_0,t)$$

where $\hat{Q}$ and $\hat{R}$ denote maximum likelihood (ML) estimates of the quantities Q and R, respectively, from equations (4) and (5) above.

We estimate the failure and accrual rates using the data observed up to time t_0. When enrolment is a homogeneous Poisson process with rate μ, the estimated enrolment rate is

$$\hat{\mu}=N(t_0)/t_0$$

When the event times are exponential, the MLE of the event rate in group j is

$$\hat{\lambda}_j=D_j(t_0)\bigg/\sum_{i=1}^{N_j(t_0)}s_{ji}$$

Similarly, when the censoring times are exponential, the MLE of the censoring rate in group j is

$$\hat{v}_j=C_j(t_0)\bigg/\sum_{i=1}^{N_j(t_0)}s_{ji}$$

We obtain ML estimates $\hat{Q}_j(t_0,t)$ and $\hat{R}_j(t_0,t)$ by inserting the parameter MLEs into (4) and (5).

4. BAYESIAN PREDICTION INTERVALS

4.1. Posterior distributions of model parameters

A second component of our approach involves calculating a Bayesian prediction interval for $T^{\star}$ by means of a simple simulation strategy. We begin by assuming that the exponential event rates λ_j are *a priori* independent with $\Gamma(A_j,B_j)$ prior densities

$$h_j(\lambda_j)\propto\lambda_j^{A_j-1}\exp(-\lambda_jB_j)$$

There are various ways to conceptualize the specification of the prior parameters. One approach is to directly specify the prior mean M_j and variance V_j, which then translate into $A_j = M_j^2/V_j$ and $B_j = M_j/V_j$. Alternatively, one can think of A_j as a prior number of events and B_j as a prior length of follow-up. In either case, the posterior density at time t_0 is

$$\Gamma\left(A_j + N_j(t_0), B_j + \sum_{i=1}^{N_j(t_0)} s_{ji}\right)$$

One can construct priors for the v_j parameters using analogous considerations for censoring events.

Similarly, we assume that the Poisson accrual rate μ is independent of the event rates and has a $\Gamma(A,B)$ prior. One can specify the prior as a mean M and variance V of μ, in which case $A = M^2/V$ and $B = M/V$. Alternatively, one can think of A as the number of accruals in a prior enrolment period of length B; this is a natural approach when data are available from a pilot study. With either method, the posterior density of μ is

$$\Gamma(A + N(t_0), B + t_0)$$

4.2. *Method for computing prediction intervals*

We calculate prediction intervals with a straightforward simulation method that involves repeating the following steps a large number of times:

(i) Draw a random sample from the current posterior of $(\lambda_1, \ldots, \lambda_k, v_1, \ldots, v_k, \mu)$.

(ii) Conditional on the data thus far and the sampled parameter value, draw a complete set of data. This involves completing the failure times for currently censored subjects, drawing enrolment times for up to a fixed target number of new subjects, and then drawing the new subjects' failure and censoring times.

(iii) With the newly completed data, calculate the calendar failure and censoring times of all the subjects. Rank the data and find $t^\star$, the $D^\star$th calendar event time. If the simulated data set does not contain $D^\star$ events, set $t^\star$ to ∞.

The resulting list of values represents a set of draws from the predictive distribution of $T^\star$. We then calculate the $\alpha/2$ and $1-\alpha/2$ quantiles of this distribution, which gives us the limits of a $100(1-\alpha)$ per cent prediction interval for the landmark date.

5. SOFTWARE

We have written functions in S-plus (MathSoft, Inc., Seattle, WA) to compute the point estimate of $T^\star$ (Section 3) and the Bayesian prediction intervals (Section 4) in a two-arm randomized trial. The software and example data are available at the second author's web site (`http://biostat.columbia.edu/~dheitjan/patrct`).

6. EXAMPLE

We have used our method since January 1999 to produce weekly interim analysis predictions for the REMATCH trial [5], a study comparing an implantable ventricular assist device to optimal medical therapy in the treatment of heart failure. Because this trial is ongoing and the data are not yet available for publication, we demonstrate the method using published data from the Chronic Granulomatous Disease (CGD) Study [4].

This study was a randomized trial comparing gamma interferon (γ-IFN) to placebo in the treatment of CGD, an inherited disorder that predisposes patients to recurrent infections. The primary endpoint was time until first infection. From August 1988 to March 1989, the study randomized 128 patients. Our data, taken from Appendix D of Fleming and Harrington [6], consist of an identifier for the hospital and the patient, a censoring indicator, the treatment group, the randomization date (in days from 27 August 1988), and the time to first infection (in days from randomization).

To demonstrate the methods, we have devised an alternative interim analysis plan for the CGD Study. Following ICGDCSG [4], we assume that the rate of first infections on placebo is one per two years, or 1/730 on the scale of time in days. We assume that γ-IFN treatment will reduce this rate by 2/3, to 1/2190, for a hazard ratio of 1/3. Using the Schoenfeld method [7], we calculated that we would need 35 events to have 90 per cent power to detect such a difference. Assuming that a single interim analysis about midway through the study would suffice, we set an interim analysis landmark at $D^{\star}=18$.

We proposed executing the analysis with a two-sided symmetric test using the O'Brien–Fleming spending function [8, 9], and assuming that 51.4 per cent of the information (18/35) is available at the interim analysis. From program `ldbnds` (David M. Reboussin, Wake Forest University), the nominal p-value for significance at the interim look is 0.0036.

We calculated prediction intervals for $D^{\star}=18$ and $D^{\star}=35$, assuming that we would inspect the data at monthly intervals after the beginning of randomization. We also included day 251 (5 May 1989), the date of the actual 18th event, and day 353 (15 August 1989), the date of the actual 35th event. We supposed that the maximum number of subjects to enrol was 128. For the Bayesian analysis, we took as event rate priors $A_0=1$, $B_0=730$ for the placebo arm and $A_1=1$, $B_1=2,190$ for the γ-IFN arm, as though we had seen one infection in 730 days of observing subjects on placebo, and one infection in 2190 days of observing subjects on γ-IFN. We set both censoring rate priors to reflect a pilot trial in which there was one loss in 3650 days (10 years) of patient follow-up. The references for this trial do not describe the *a priori* expectation for accrual, but in the event the observed rate was 128 patients in 206 days, or 0.62 per day. Because most trials overestimate accrual, we suppose that the investigators' prior knowledge was equivalent to a pilot trial that enrolled $A=30$ subjects in $B=15$ days, for an expected rate of $A/B=2$ subjects per day.

Table I presents the numbers of events and the logrank p-values at the monthly update times, and Figure 1 plots the monthly 95 per cent prediction interval against the date of the prediction. The connecting lines give the point predictions as a function of time. The intervals are wide initially, as there were few infections in the early months. As events occur (at first only in the placebo arm), the intervals shrink slightly, although they increase from month 4 to month 5, a period that saw no events take place. They shrink again considerably between months 5 and 6, in response to six new events on placebo compared to only one on active. The 18th event took place on day 251 (5 May 1989), a point which lay within all of the

Table I. Progress of the CGD Study.

Day	Date	First infections		
	(year.month.day)	Placebo	γ-IFN	Logrank P
30	88.09.26	1	0	0.0833
60	88.10.26	2	0	0.1063
90	88.11.25	3	0	0.0630
120	88.12.25	4	0	0.0319
150	89.01.24	4	0	0.0281
180	89.02.23	10	1	0.0027
210	89.03.25	11	2	0.0054
240	89.04.24	13	3	0.0017
251	89.05.05	14	4	0.0045
270	89.05.24	16	5	0.0042
300	89.06.23	18	6	0.0037
330	89.07.23	21	6	0.0006
353	89.08.15	24	11	0.0027
360	89.08.22	24	11	0.0033

preceding prediction intervals. Because the p-value on that date (0.0045) exceeded its critical value (0.0036), our version of the trial would have proceeded to its final analysis.

The 35th recorded event took place on 15 August 1989. Except for the 11-month prediction, which gave the interval 23 August to 24 December, all the other preceding intervals bracketed the actual date. The final prediction failed to anticipate a rash of eight events that occurred between 23 July and 15 August. The largest number of events in any earlier month was seven, which occurred between 24 January and 23 February.

For comparison, we applied formula (A2) of Rubinstein *et al.* [2] under the *a priori* parameter estimates. The formula predicts that the 18th event will occur on day 217, or 1 April 1989 – 34 days earlier than the actual date of 5 May. Although our 23 February (day 180) prediction interval included this date, the interval was wide (15 March to 23 June), reflecting the uncertainty that still existed at this stage of the trial. The pre-trial prediction for the 35th event is 23 November 1989 – 102 days late. The problem with these predictions is that the pre-trial rates overestimated the accrual rate and underestimated the event rate. The prediction of the 18th event time is fairly good because the errors cancel, but by the time of the 35th event the underestimated mortality dominates and causes the prediction to be significantly late.

We have executed a number of analyses to explore the sensitivity of predictions to the prior. Predictably, the prior dominates early in the study, when data are sparse. For example, when we specify the accrual prior as $A = 120$ subjects in $B = 300$ days, the prediction intervals are wide and late because the prior strongly asserts, contrary to the data, that accrual will be slow. In extreme cases the prediction intervals can fail to cover the point predictions, which are based on ML estimates that ignore prior information. As one approaches the analysis landmark, however, the prior loses influence because the timing of the target event becomes less dependent on new accrual and more dependent on subjects already in the study and their pending event times.

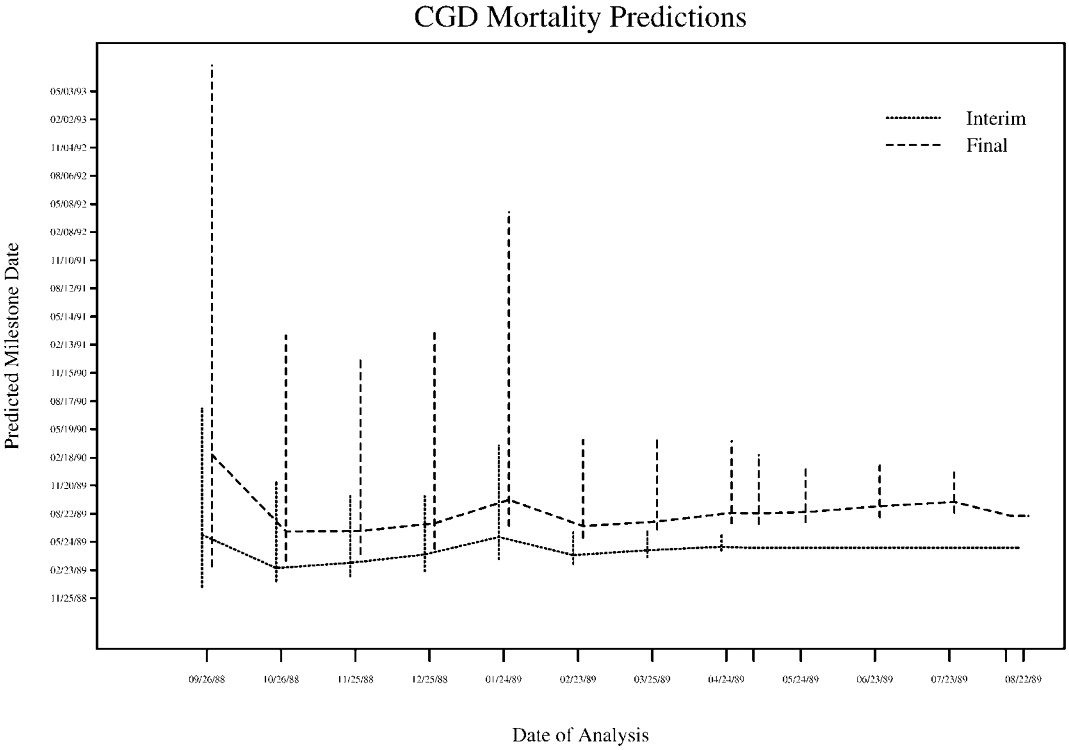

Figure 1. Interval and point predictions of the dates of the interim (18 events) and final (35 events) milestone times in the CGD Trial, computed at monthly intervals from the opening of enrolment. We also include 5 May 1989 (day 251), the date of the 18th event, and 15 August 1989, the date of the 35th event.

7. DISCUSSION

In many trials it is valuable to predict the dates of landmark events. Our method goes beyond existing approaches by combining prior estimates of accrual, event and censoring rates with data from the trial itself, and in producing prediction intervals rather than just point predictions. Although our example used roughly equally spaced landmarks, our model and software can accommodate unequally spaced landmarks as well.

Several features of our model require further development. First, because ours is a parametric model, if the underlying accrual and failure-time distributions differ from our assumptions the predictions are potentially biased or inefficient. Yet because our method involves prediction of future events, some modelling is necessary, and parametric modelling is particularly convenient. A richer family of survival models such as the generalized F [10] could make the predictions more robust.

A second concern is non-homogeneity of accrual rates. Our model assumes a constant accrual rate, whereas in practice many trials accrue slowly in the beginning and more rapidly later as investigators gain experience with the protocol. Slow accrual early would presumably cause overestimation of times to analysis landmarks, which is less damaging than underestimation. In any event, it should be possible to model accrual rates with simple parametric forms that would capture most of this effect.

Our model also assumes homogeneity among centres in accrual, event and censoring rates. When there are large disparities among centres, as sometimes happens, explicitly modelling these features could lead to better calibrated predictions.

Because our model involves simulating completed data sets, we can use it to predict other things besides landmark event times. For example, if interim analyses are scheduled in calendar time rather than event time, one can use the simulations to predict the number of events that will have taken place by prespecified analysis dates. One can also compute the predictive power [11] by executing the primary data analysis on each simulated data set and tabulating the proportion of data sets where the result is statistically significant. One can compute the conditional power [12] by the same method, using an event-rate prior that assigns probability 1 to the alternative-hypothesis rates.

Our experience with the model so far has been favourable. The REMATCH DSMB met in December 1998, and the NHLBI intended to convene it again in the spring of 1999. They agreed to postpone the meeting until October when we predicted (correctly) that we would reach an interim analysis milestone in the summer. This brief delay saved thousands of dollars in project funds and opportunity costs, without harm to the patients or to the integrity of the trial. Moreover, we have found our weekly prediction of the analysis times to be valuable both as a logistical planning tool and as a measure of the trial's progress.

REFERENCES

1. Fleming TR. Evaluating therapeutic interventions: Some issues and experiences. *Statistical Science* 1992; **4**: 428–441.
2. Rubinstein LV, Gail MH, Santner TJ. Planning the duration of a comparative clinical trial with loss to follow-up and a period of continued observation. *Journal of Chronic Diseases* 1981; **34**:469–479.
3. Lachin JM, Foulkes MA. Evaluation of sample size and power for analyses of survival with allowance for nonuniform patient entry, losses to follow-up, noncompliance and stratification. *Biometrics* 1986; **42**:507–519.
4. International Chronic Granulomatous Disease Cooperative Study Group. A controlled trial of interferon gamma to prevent infection in chronic granulomatous disease. *New England Journal of Medicine* 1991; **324**:509–516.
5. Rose EA, Moskowitz AJ, Packer M, Sollano JA, Williams DL, Tierney AR, Heitjan DF, Meier P, Ascheim DD, Levitan RG, Weinberg AD, Stevenson LW, Shapiro PA, Lazar RM, Watson J, Goldstein D, Gelijns AC for the REMATCH investigators. The REMATCH Trial: Rationale, design and endpoints. *Annals of Thoracic Surgery* 1999; **67**:723–730.
6. Fleming TR, Harrington DP. *Counting Processes and Survival Analysis*. Wiley: New York, 1991.
7. Schoenfeld DA. Sample-size formula for the proportional-hazards regression model. *Biometrics* 1983; **39**: 499–503.
8. O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. *Biometrics* 1979; **35**:549–556.
9. Lan KKG, DeMets DL. Group sequential procedures: calendar versus information time. *Statistics in Medicine* 1989; **8**:1191–1198.
10. Kalbfleisch JD, Prentice RL. *The Statistical Analysis of Failure Time Data*. Wiley: New York, 1980.
11. Spiegelhalter DJ, Freedman LS, Blackburn PR. Monitoring clinical trials: Conditional or predictive power? *Controlled Clinical Trials* 1986; **7**:8–17.
12. Halperin M, Lan KKG, Ware JH, Johnson NJ, DeMets DL. An aid to data monitoring in long-term clinical trials. *Controlled Clinical Trials* 1982; **3**:311–323.