BIOMETRICS 55, 1085–1092
December 1999

Sequential Methods for Comparing Years of Life Saved in the Two-Sample Censored Data Problem

Susan Murray

Department of Biostatistics, University of Michigan School of Public Health,
1420 Washington Heights, Ann Arbor, Michigan 48109-2029, U.S.A.
email: skmurray@umich.edu

and

Anastasios A. Tsiatis

Department of Statistics, North Carolina State University,
Raleigh, North Carolina 27695-8203, U.S.A.

SUMMARY. This research develops nonparametric strategies for sequentially monitoring clinical trial data where detecting years of life saved is of interest. The recommended test statistic looks at integrated differences in survival estimates during the time frame of interest. In many practical situations, the test statistic presented has an independent increments covariance structure. Hence, with little additional work, we may apply these testing procedures using available methodology. In the case where an independent increments covariance structure is present, we suggest how clinical trial data might be monitored using these statistics in an information-based design. The resulting study design maintains the desired stochastic operating characteristics regardless of the shapes of the survival curves being compared. This offers an advantage over the popular log-rank-based design strategy since more restrictive assumptions relating to the behavior of the hazards are required to guarantee the planned power of the test. Recommendations for how to sequentially monitor clinical trial progress in the nonindependent increments case are also provided along with an example.

KEY WORDS: Group sequential monitoring; Nonparametric; Survival; Two-sample test.

1. Introduction

In clinical trials, we are ethically obligated to periodically monitor the performance of the treatments under study. It is important, for instance, to detect overly toxic effects that may occur and to protect the patients in the study accordingly. Also, if the new treatment under study performs substantially better than standard treatments, interim analyses of the data detect these differences early. For instance, this is a motivation for sequential analysis in most AIDS clinical trials, where there is a strong need to identify life-enhancing treatments quickly. In one such pivotal study described by Fischl et al. (1990), patients were randomized to high-dose ($n = 262$) versus low-dose AZT ($n = 262$) with wide reaching effects on standard of care for AIDS patients upon the study's conclusion. The higher dose of AZT, which was the standard of care prior to the 1990 publication, had been shown to be an effective treatment compared to placebo but with substantial toxicity. The results observed by Fischl et al. suggested that a lower dose of AZT was more effective and less toxic than the standard. The scientists agreed at the design phase of the study that repeated analyses would preserve the best interests of the patient population.

However, a statistical price must be paid for this monitoring process. By repeatedly looking at the data, we increase our type I error probability. Hence, when analyzing data in this way, we must make adjustments in the sequential boundaries of our test statistics to maintain our type I error. Among popular stopping methods for sequential clinical trials are those suggested by Pocock (1977) and O'Brien and Fleming (1979). More recently, flexible error spending functions have been proposed by Lan and DeMets (1983) to maintain the type I error throughout the trial.

One important requirement for using error spending functions in repeated analyses is an understanding of the joint distribution of the repeated statistical tests. For instance, the joint distribution of sequential log-rank (LR) tests as well as a framework for analyzing these repeated tests were studied by Tsiatis (1981, 1982), Slud (1984), and later Gu and Lai (1991). Because sequential methodology has been developed and studied for the LR test, it has become the primary nonparametric test applied to survival data in repeatedly moni-

tored trials. In fact, this methodology was used in the high-dose versus low-dose AZT study described above. The LR test is especially useful in situations where the treatment difference can be expressed as a hazard ratio assumed constant over time. For such proportional hazards alternatives, the LR test is the optimal nonparametric test. However, it often behaves poorly when hazards are not proportional. In fact, in the nonproportional hazards setting, the LR test statistic does not have a useful interpretation with respect to any treatment difference of particular interest.

An alternative to relative risk is to express treatment differences in terms of years of life saved (YLS) between treatments. That is, regardless of the behavior of the survival hazards across time, an alternative of interest in the AZT study could be defined through the YLS on the superior treatment. Since most clinical trials have limited follow-up, we would define years of life, by necessity, up to some fixed point in time after treatment. This quantity can be estimated easily with differences between areas under estimated survival curves. This resulting YLS test statistic, $\mathcal{T}$, belongs to a larger class of statistics studied by Pepe and Fleming (1989) for the nonsequential setting.

In our remaining discussion, we have three scientific goals. We shall first set up the framework for sequential methods using the $\mathcal{T}$ statistic. An important step in establishing the sequential nature of the $\mathcal{T}$ statistic is characterizing the joint distributions of the tests and survival estimates computed at different interim analyses. This will enable us to compute stopping boundaries for preserving desired operating characteristics. Jennison and Turnbull (1985) studied sequentially computed Kaplan–Meier (KM) survival estimates. Notation required in the sequential framework and results relating to the KM estimate at two separate analysis times are presented in Section 2. Subsequently, we develop methods for adapting the $\mathcal{T}$ statistic to the sequential trials setting in Section 3. In particular, we will discuss under what circumstances the $\mathcal{T}$ statistic has a joint independent increments structure. This is important because, if this is the case, we can immediately apply standard group-sequential methods to this problem.

As a second scientific goal, we shall indicate in Section 4 through simulation the performance of the $\mathcal{T}$ statistic in an information-based monitoring design when the interim test statistics have independent increments. This information-based monitoring technique introduced by Lan and Zucker (1993) and recently formalized by Kim, Boucher, and Tsiatis (1995) and Scharfstein, Tsiatis, and Robins (1997) follows patients until the total information (TI) required to achieve the desired power is collected. During the course of the trial, error spending functions are defined according to the proportion of information collected at each interim analysis. In defining the TI required for the $\mathcal{T}$ statistic to achieve a particular power, the alternative to be detected is defined in terms of the YLS between the two treatments without any additional assumptions regarding the shapes of the underlying survival curves.

Finally, this research instructs practitioners on use of the $\mathcal{T}$ statistics in the more general nonindependent increments setting, where information-based monitoring is difficult to implement. An example in the context of the aforementioned AIDS clinical trial comparing low-dose versus high-dose AZT is given along with further simulation results in Section 5. A discussion follows in Section 6.

2. Sequential Methods for the Kaplan–Meier Estimate

To understand concepts relating to sequential theory, we require an explanation of notation. We borrow notation used by Tsiatis, Boucher, and Kim (1995). Assume that n individuals will enter the trial at times $E_1, \ldots, E_n$ during the accrual period. These possibly different entry times will be bounded positive random variables measuring calendar time from the start of the study, which are independently and identically distributed with distribution function $G(e) = P(E \leq e)$. Each individual i has an internal survival time, T_i, measured from the time of entry to the study. Individuals who have not failed at the time of analysis are censored. For instance, if the data are analyzed at calendar time t, we censor all individuals i such that $T_i > t - E_i$. The notation V_i, measured in internal patient time from study entry, refers to the potential censoring time due to random loss to follow-up. We assume that E_i, V_i, and T_i are independent within each treatment group. If the data are analyzed at calendar time t, then the observable internal patient time random variables are $\{X_i(t), \Delta_i(t)\}$ for all $i = 1, \ldots, n$ such that $E_i \leq t$, where $X_i(t) = \min(T_i, V_i, t - E_i)$ is the observed time on study at analysis time t and $\Delta_i(t) = I\{T_i \leq \min(t - E_i, V_i)\}$ denotes the failure indicator at time t.

Define the KM estimate of the survival function $S(x)$, if we used the survival data available at time t, as $\hat{S}(t, x)$. We emphasize that the first index, t, refers to calendar time measured from the start of the study. The second index, x, refers to the internal patient time measured from a patient's entry into the study. Let t_1 denote the calendar time of the first analysis and t_2 a later analysis time. In the Appendix, we show that

$$\operatorname{cov}\left[\{\hat{S}(t_1, x_1) - S(x_1)\}, \{\hat{S}(t_2, x_2) - S(x_2)\}\right] \approx \frac{S(x_1)S(x_2)}{n(t_2)} \int_0^{\min(x_1, x_2)} \frac{\lambda(u)du}{H(t_2, u)S(u)}, \quad (2.1)$$

where $\lambda(x)$ is the hazard function at internal patient time x, $n(t) = \Sigma_{i=1}^{n} I(E_i \leq t)$ is the total sample size enrolled at calendar time t, and $H(t, x) = P(E_i < t - x, V_i > x)/P(E_i \leq t)$ is the censoring survival distribution among individuals entered by calendar time t. This is the usual covariance of the KM estimate at the later interim look time conditional on those entered by time t_2.

3. Sequential Methods for the $\mathcal{T}$ Statistic

The sequential notation from the last section will be employed in this section as well. Temporarily, however, let us consider the case where one analysis is conducted at the end of a study. Since in this case only one analysis is performed, we submerge the calendar time subscript, t, and focus our attention on the internal patient time subscripts and random variables.

The mean survival time, $E(T)$, is equal to $\int_0^\infty S(u)du$. Hence, if consistent survival estimates were available for $u \in (0, \infty)$, the area under this survival curve would consistently estimate the mean survival time in the presence of censored survival data. However, both theoretical and practical arguments might dictate an upper limit of integration, τ, different from ∞. For instance, clinical investigators might be interested in studying internal patient survival time during the

first τ years on study. On a more practical note, the study design may not allow for estimating $S(x)$ beyond the study time frame so that τ might be the last point where a consistent survival estimate $\hat{S}(\tau)$ may be defined.

In both of these cases, it would be useful to consider a truncated mean, $E\{\min(T,\tau)\} = \int_0^\tau S(x)dx$. One interpretation of this truncated mean is that the area under the survival curve from 0 to τ is the average years of life lived during τ years on study. Hence, if n_g is the sample size in group g at the end of the study, $n = n_1 + n_2$, and $\hat{S}_g$ is the KM survival estimate in group g at the end of the study, the test statistic

$$\mathcal{T} = \left(\frac{n_1 n_2}{n}\right)^{\frac{1}{2}} \int_0^\tau \{\hat{S}_1(u) - \hat{S}_2(u)\}du,$$

nonparametrically compares YLS between treatments 1 and 2 during the first τ years on study. Note that, notationally, with one only analysis at time t, the censoring survival function $H(t,x) = H(x)$. To ensure that the variance of this test statistic is bounded, we shall always require that τ satisfy the requirement $\hat{H}(\tau) > 0$, where $\hat{H}(x)$ estimates $H(x)$. When τ is chosen close to the minimum of the largest event times from groups 1 and 2, this test statistic belongs to a larger class of statistics studied by Pepe and Fleming (1989).

Define $H_g(x)$, $g = 1,2$, as the probability of remaining uncensored at time x for group g. Also, let $A^\tau(x) = \int_x^\tau S(u)du$. When only one statistical analysis is performed, Pepe and Fleming (1989) showed that $\mathcal{T} \xrightarrow{\mathcal{D}} N(0,\sigma^2)$, where, under the null hypothesis of no treatment difference, $\sigma^2 = \Sigma_{g=1}^2 \pi_{3-g} \times \int_0^\tau \{S(u)H_g(u)\}^{-1}\{A^\tau(u)\}^2 \lambda(u)du$, π_g is the probability of falling in group g, and $\lambda(x)$ and $S(x)$ are the hazard and survival functions, respectively, common to both groups under the null hypothesis. Define $\tilde{S}(x)$ as the pooled KM estimator from groups 1 and 2. Let $\hat{A}^\tau(x) = \int_x^\tau \tilde{S}(u)du$, $\hat{H}_g$ be the estimate for the censoring time survival probability, $\tilde{N}(x)$ be the observed number of deaths at time x, $\tilde{Y}(x)$ be the observed number of individuals still at risk at time x, and $\hat{\pi}_g = n_g/n$. Then the variance of the $\mathcal{T}$ statistic may be estimated by $\sigma^2 = \hat{\Sigma}_{g=1}^2 \hat{\pi}_{3-g} \int_0^\tau \{\tilde{S}(u)\hat{H}_g(u)\tilde{Y}(u)\}^{-1}\{\hat{A}^\tau(u)\}^2 d\tilde{N}(u)$. This corresponds to the variance estimate recommended by Pepe and Fleming (1989).

Note that, under the alternative hypothesis, $\hat{\Delta} = \int_0^\tau \{\hat{S}_1(u) - \hat{S}_2(u)\}du$ has a nonzero mean, $\Delta = \int_0^\tau \{S_1(u) - S_2(u)\}du$, which measures the true integrated differences between survival curves in the two groups. Suppose we want to detect a clinically important treatment difference Δ with power $1-\beta$ and size α using the standardized test statistic $\mathcal{T}/\mathrm{SE}(\mathcal{T}) = \hat{\Delta}/\mathrm{SE}(\hat{\Delta})$, where SE denotes standard error. This would require $\{\mathrm{var}(\hat{\Delta})\}^{-1} = \Delta^{-2}(z_{\alpha/2} + z_\beta)^2$, where z_* represents the $1-*$ quantile of the standard normal distribution. The right-hand side of this equation is known as the total statistical information (TI) required to achieve power $1-\beta$. If test statistics computed at different analysis times have an independent increments covariance structure, the left-hand side of the equation suggests a natural way to measure the degree of information in a test statistic monitored across time, i.e., the inverse of the estimated variance for $\hat{\Delta}$ can be monitored until it matches the TI required at the last analysis time. Since the TI required to achieve power $1-\beta$ is defined in terms of Δ, or the average YLS that is clinically important to detect, it is robust to the actual shapes of the survival curves.

Using the results of the sequentially computed KM estimator, we now derive the sequential distribution of the $\mathcal{T}$ statistic using previous notation. If we define the $\mathcal{T}$ test evaluated at time t_j as $\mathcal{T}(t_j) = \{n^*(t_j)\}^{\frac{1}{2}} \int_0^{\tau_j} \{\hat{S}_1(t_j,u) - \hat{S}_2(t_j,u)\}du$, where $n^*(t_j) = n_1(t_j)n_2(t_j)/n(t_j)$, $j = 1,2$, then under the null hypothesis, the asymptotic distribution of $\mathcal{T}(t_1)$ and $\mathcal{T}(t_2)$ is multivariate normal with mean zero and asymptotic covariance to be derived later in this section. In particular, we investigate two scenarios involving the upper limits of integration at analysis times t_1 and t_2, where either $\tau_1 = \tau_2$ or $\tau_1 < \tau_2$. These two special cases shall drive the asymptotic behavior of the sequentially evaluated statistics. The special case where $\tau_1 = \tau_2 = \tau$ may occur in any scenario where a particular window of time within a study is of interest. This would occur, e.g., in a clinical trial with a long accrual period and quick event times or in a case where therapy is given over a short period of time with a treatment effect expected to be realized over some time τ smaller than the duration of the study. For example, patients who have experienced a myocardial infarction may be given anticoagulants for a period of 30 days with survival benefit not expected to extend beyond an initial 6-month period of high risk for the patients in the study. After 6 months, survivors should have similar risk regardless of initial treatment. In such a case, the data may be monitored periodically after 6 months using $\tau = 6$. When primary interest lies in detecting YLS during the entire study period for an event that takes a longer time to be observed, we may increase τ_2 at interim analysis time t_2 to compensate for an increase in the range of data observed beyond time τ_1.

For $a \le \tau_1 \le \tau_2$, consider the expression

$$\{n^*(t_1)n^*(t_2)\}^{\frac{1}{2}} \mathrm{cov}\left[\int_0^{\tau_1} \{\hat{S}_1(t_1,u_1) - \hat{S}_2(t_1,u_1)\}du_1, \int_a^{\tau_2} \{\hat{S}_1(t_2,u_2) - \hat{S}_2(t_2,u_2)\}du_2\right],$$

which becomes the asymptotic covariance of $\mathcal{T}(t_1)$ and $\mathcal{T}(t_2)$ in the case $a = 0$. Under the null hypothesis, this expression becomes

$$\begin{aligned}
&\sum_{g=1}^2 \{n^*(t_1)n^*(t_2)\}^{\frac{1}{2}} \\
&\quad \times \int_0^{\tau_1}\int_a^{\tau_2} \mathrm{cov}\{\hat{S}_g(t_1,u_1), \hat{S}_g(t_2,u_2)\}du_2du_1 \\
&\quad = \sum_{g=1}^2 \{n^*(t_1)n^*(t_2)\}^{\frac{1}{2}} \\
&\qquad \times \int_0^{\tau_1}\int_a^{\tau_2} \mathrm{cov}\{\hat{S}_g(t_2,u_1), \hat{S}_g(t_2,u_2)\}du_2du_1 \\
&\quad = \sum_{g=1}^2 \{n^*(t_1)n^*(t_2)\}^{\frac{1}{2}} \\
&\qquad \times \int_0^{\tau_1}\int_a^{\tau_2} \frac{S(u_1)S(u_2)}{n_g(t_2)} \\
&\qquad \times \int_0^{\min(u_1,u_2)} \frac{\lambda(u)du}{H_g(t_2,u)S(u)} du_2du_1. \qquad (3.1)
\end{aligned}$$

To simplify notation, define $A^{\tau_j}(x) = \int_x^{\tau_j} S(u)du, j = 1, 2$, which at any particular analysis time t would be estimated with $\hat{A}^{\tau_j}(t, x) = \int_x^{\tau_j} \hat{S}(t, u)du$. Continuing from (3.1), we arrive at

$$\sum_{g=1}^{2} \frac{\{n^*(t_1)n^*(t_2)\}^{1/2}}{n_g(t_2)} \int_0^{\tau_1} \frac{A^{\tau_2}(\max(u, a))A^{\tau_1}(u)\lambda(u)}{H_g(t_2, u)S(u)}du. \tag{3.2}$$

This covariance converges in probability to the variance identified by Pepe and Fleming (1989) in the only one analysis case where $a = 0$, $t_1 = t_2$, and $\tau_1 = \tau_2$.

In the special case where $a = 0$ and $\tau_1 = \tau_2$, result (3.2) reveals the asymptotic covariance of $\mathcal{T}(t_1)$ and $\mathcal{T}(t_2)$ to be $\{n^*(t_1)\}^{1/2}\{n^*(t_2)\}^{-1/2}\text{var}\{\mathcal{T}(t_2)\}$, so that this term relates directly to the variance of the $\mathcal{T}$ statistic at interim look time t_2. This statistic can easily be redefined with an independent increments structure. To see this, consider the statistic $\mathcal{T}^*(t) = \{n^*(t)\}^{1/2}[\text{var}\{\mathcal{T}(t)\}]^{-1}\mathcal{T}(t)$. The standardized statistics $\mathcal{T}(t)/[\text{var}\{\mathcal{T}(t)\}]^{\frac{1}{2}}$ and $\mathcal{T}^*(t)/[\text{var}\{\mathcal{T}^*(t)\}]^{1/2}$ are algebraically equivalent definitions. Hence, for testing purposes, we may use them interchangeably. Also note that $\text{cov}\{\mathcal{T}^*(t_1), \mathcal{T}^*(t_2)\} = n^*(t_1)[\text{var}\{\mathcal{T}(t_1)\}]^{-1}[\text{var}\{\mathcal{T}(t_2)\}]^{-1}\text{var}\{\mathcal{T}(t_2)\} = n^*(t_1)[\text{var}\{\mathcal{T}(t_1)\}]^{-1} = \text{var}\{\mathcal{T}^*(t_1)\}$ fulfills the standard definition of independent increments. Most of the test statistics currently used in group sequential monitoring have an independent increment structure. Hence, with little additional effort, the $\mathcal{T}$ test statistic may be adapted for sequential trials using currently available software based on the recursive numerical integration strategy of Armitage, McPherson, and Rowe (1969).

In the case where $\tau_1 < \tau_2$, the covariance of $\mathcal{T}(t_1)$ and $\mathcal{T}(t_2)$ from (3.2) for $a = 0$ can be described in two pieces. The first piece uses $a = 0$ and $\tau_1 = \tau_2$ so that currently available software can be used in estimation; the remaining piece is identified using (3.2) with $a = \tau_1$. Hence, in this case, the covariance of $\mathcal{T}(t_1)$ and $\mathcal{T}(t_2)$ is estimated under the null hypothesis with

$$\left(\frac{n^*(t_1)}{n^*(t_2)}\right)^{(1/2)} \hat{\sigma}^2(t_2) + \sum_{g=1}^{2} \frac{\{n^*(t_1)n^*(t_2)\}^{1/2}}{n_g(t_2)} \times \int_0^{\tau_1} \frac{\hat{A}^{\tau_1}(t_2, u)\hat{A}^{\tau_2}(t_2, \tau_1)d\tilde{N}(t_2, u)}{\hat{H}_g(t_2, u)\tilde{S}(t_2, u)\tilde{Y}(t_2, u)},$$

where the integral in $\hat{\sigma}^2$ has τ_1 as an upper limit and the pooled estimator of survival is used in calculating $\hat{A}^{\tau_j}(t, x)$. This second scenario results in a statistic that does not have an independent increments structure.

To calculate sequential boundaries in the nonindependent increments case, we may use simulation techniques. First, a suitable spending function is selected, such as the O'Brien–Fleming (OF) style of spending using the function $\alpha_{of} = 2 - 2\Phi(z_{\alpha/2}/v^{1/2})$, where v corresponds to some surrogate for the proportion of information collected at an interim analysis time. Then we estimate the covariance structure between the current and all previous $\mathcal{T}$ test statistics calculated during the course of the trial using the observed data. Multivariate normal random variables with the observed covariance structure are simulated to estimate appropriate cutoff points for the statistics at the different analysis times that spend the predetermined type I errors as described above. An example is given in Section 5. Alternatively, software called MULNOR written by Schervish (1984) can be used to calculate sequential boundaries of the $\mathcal{T}$ statistic that satisfy the study design.

4. Application of Information-Based Design (Independent Increments)

We illustrate this methodology via a simulation experiment where a new treatment is compared to a standard in a randomized clinical trial. We expect treatment benefit to extend for a period of 2 years. Clinical investigators indicate that 3 months of life saved over the 2-year period ($\Delta = .25$ years) is a clinically important difference to detect with 90% power at the 5% level of significance. We assume a slow but constant accrual rate of 75 eligible patients each year so that monitoring of the trial will not begin before we have at least 2 years of follow-up on some patients. Hence, $\tau = 2$ years becomes a natural truncation point at each of the interim analyses resulting in an independent increments testing structure where standard software can be employed for calculating error spending functions. A maximum information trial as recently studied by Kim et al. (1995) and Scharfstein et al. (1997) would plan the final analysis at a time when information accrued matches the TI needed at the end of the study to meet power and size requirements. Using an OF error spending function with three planned analyses, the TI required at the end of the study for 90% power and 5% size is approximately $1.05\Delta^{-2}(1.96 + 1.28)^2$, where $\tau = 2$ and 1.05 is an inflation factor typically used for this style of error spending. Note that no specific forms for the survival distributions of either therapy were used to determine Δ or the TI required to meet study design requirements. When the sequentially computed test statistic has an independent increments covariance structure, the information accumulated is approximated using $(\widehat{\text{var}}\,\hat{\Delta})^{-1}$ at each analysis time so that dividing this quantity by the desired TI gives the information proportion for group sequential monitoring. At the last analysis time, we would like $(\widehat{\text{var}}\,\hat{\Delta})^{-1} \approx \text{TI}$.

To plan analysis times and obtain a rough idea for how long the study will need to accrue patients before the the necessary TI is collected, we may investigate various choices for the survival distribution of the experimental therapy. From past experience, we assume the standard treatment follows an exponential distribution with hazard one. For planning purposes, we assume the survival distribution for patients on the new treatment follows an exponential with hazard .655, resulting in the desired $\Delta = .25$ years. Using Maple software, we determine that $(\text{var}\,\hat{\Delta})^{-1} \approx \text{TI}$ five years from the beginning of the study if information accrues at the anticipated rate. The stopping time for the trial will be the first analysis time where the observed YLS statistic extends beyond the sequential boundaries provided from standard sequential software or the time at which we collect our TI if no rejection occurs before this time. With our first analysis planned for year 3, we monitor the percent of information accrued at monthly intervals from year 3 until the end of the trial. If at any time the TI is reached, a final analysis is performed. If the TI is not yet reached at year 4, a second interim analysis

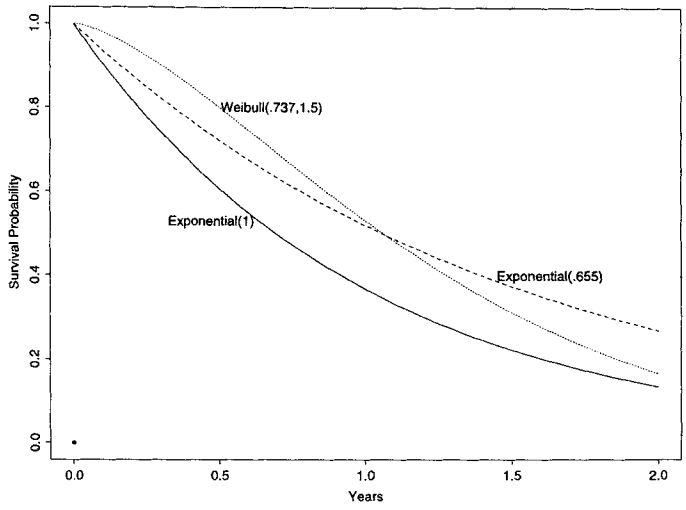

Figure 1. Survival curves used in simulation. The standard treatment survival distribution corresponds to the exponential with hazard one. For study design purposes, the exponential with hazard .655 was postulated for the experimental survival distribution. However, we also consider the case where the true alternative distribution corresponds to the Weibull with scale parameter .737 and shape parameter 1.5. Notice that the areas between either the postulated alternative and the standard treatment curve or the true alternative and the standard treatment curve are equivalent. Also, the median survival is equivalent for the two alternatives under consideration.

is performed and the monthly check of the percent information accrued continues after this analysis. The final analysis time may extend beyond 5 years if the TI has not yet been reached. This monitoring strategy allows for a flexible study design, which may be preferred to using fixed analysis times if the alternative distribution is much different from what was expected. We again emphasize that, although we make assumptions on the shape of the survival distributions for planning purposes, the stochastic operating characteristics of the resulting information-based group sequential procedure do not depend on these assumptions being true in the independent increments setting. However, the duration of the trial may be affected by information being collected at a different rate than anticipated in the design, as we shall witness in simulations described later in this section.

We simulated the monitoring strategy mentioned above 500 times (1) under the null hypothesis, (2) under the hypothesized exponential alternative used for planning purposes, and (3) using a Weibull(.737, 1.5) alternative that also gives us .25 YLS so that the YLS alternative is unchanged (see Figure 1). Note that the median survival for (2) and (3) also match exactly. In planning a sequential analysis using the LR test, currently available software queries the user for median survival in the groups being compared and makes the same exponential distribution assumptions. Using the LR test with OF-style boundaries, the TI required for 90% power and 5% type I error is reached when 239 events have occurred in this scenario. Under (1), sequential monitoring resulted in type I errors of .046 and .052 for the YLS and LR procedures, respectively. Tables 1 and 2 display power, a listing of when studies were stopped, and also a listing of when studies attained their maximum information for 500 simulations using information-based monitoring with the YLS statistic and the LR statistic. Under (2) in Table 1, we verify approximately 90% power using either sequential testing criteria. In this table, the LR test tends to end the trial a bit earlier than the YLS statistic. Under (3), most sequential trials end earlier using the YLS method (see Table 2). Table 2 reveals that the YLS statistic in an information-based design maintains its operating characteristics when the alternative turns out to be from a Weibull distribution instead of the planned for exponential distribution. However, this slight shift in the distribution of the alternative reduces the power of the LR testing procedure to 73.8%.

5. Example of Monitoring in Nonindependent Increments Case

When the upper limit τ_j at the jth interim analysis varies with j, the YLS statistic no longer has the nice independent increments structure described in the previous section. We illustrate how such trials are monitored using survival data

Table 1

Power and frequency of stopping times (with frequency of times maximum information reached in parentheses) for 500 replications using O'Brien–Fleming (OF) spending function in maximum information design in the case where the alternative is truly exponential.

	YLS (power = .904)				Log rank (power = .904)			
	Reject H_0							
Frequency of stopping times	Yes		No		Yes		No	
<4.00 years	184	(0)	0	(0)	214	(0)	0	(0)
[4.00,4.25) years	183	(0)	0	(0)	205	(19)	4	(4)
[4.25,4.50) years	0	(1)	1	(1)	18	(232)	33	(33)
[4.50,4.75) years	2	(23)	6	(6)	14	(195)	9	(9)
[4.75,5.00) years	18	(95)	16	(16)	1	(6)	2	(2)
[5.00,5.25) years	37	(196)	18	(18)	0	(0)	0	(0)
[5.25,5.50) years	27	(126)	7	(7)	0	(0)	0	(0)
≥5.5 years	1	(11)	0	(0)	0	(0)	0	(0)

Table 2
Frequency of stopping times (with frequency of times maximum information reached in parentheses) for 500 *replications using O'Brien–Fleming (OF) spending function in maximum information design when the alternative turns out to be Weibull.*

	YLS (power = .916)				Log rank (power = .738)			
	Reject H_0							
Frequency of stopping times	Yes		No		Yes		No	
<4.00 years	261	(7)	0	(0)	198	(0)	0	(0)
[4.00,4.25) years	163	(41)	7	(7)	142	(73)	19	(19)
[4.25,4.50) years	5	(134)	16	(16)	22	(252)	96	(96)
[4.50,4.75) years	19	(182)	15	(15)	7	(44)	16	(16)
[4.75,5.00) years	9	(87)	4	(4)	0	(0)	0	(0)
[5.00,5.25) years	1	(7)	0	(0)	0	(0)	0	(0)
[5.25,5.50) years	0	(0)	0	(0)	0	(0)	0	(0)
≥5.5 years	0	(0)	0	(0)	0	(0)	0	(0)

from the AIDS clinical trial mentioned in the introduction, in which patients were assigned low-dose ($n = 262$) or high-dose ($n = 262$) zidovudine regimens. We have recreated the data that would have been available for interim analyses in 1987, 1988, and 1989. Of statistical interest at each of these analyses is the average YLS on study, i.e., for the analysis in $1987+i, i = 0, 1, 2$, we would like to know if there is a statistically significant difference in the average years of life lived from study entry to the end of $1987+i$ dependent on treatment. Since the AIDS patients in this study are likely to live longer than 1 year, our upper limit of integration in the $\mathcal{T}$ test statistic will increase with each additional interim analysis. Since we do not have an independent increments testing structure, the methods discussed toward the end of Section 3 for the more complex covariance structure are employed. In particular, we use the OF spending function for type I error recommended earlier where we use calendar time as a surrogate for v. Hence, we plan to spend approximately $(.0007, .0157, .0336)$ type I error at the three respective analysis times.

Using the YLS approach on the low-dose treatment arm, we observed an average of $\hat{\Delta} = 14, 34$, and 63 days of life saved (DLS) from entry into the study until the analysis times in late 1987, 1988, and 1989, respectively. The observed covariance matrix for these estimated means at the various analysis times was

$$\begin{pmatrix} 99.95 & 68.17 & 70.00 \\ 68.17 & 385.23 & 341.17 \\ 70.00 & 341.17 & 655.74 \end{pmatrix}.$$

In order to determine proper cutpoints for statistical significance with these repeated tests, we generated 10,000 multivariate normal random variables with means centered at zero and covariance equal to the observed covariance above. Estimated quantile-based cutpoints for the average number of DLS corresponding to the type I errors allowed at each analysis time were approximately 35, 49, and 55 days. Since we estimated 63 DLS on average from study entry to the analysis time in late 1989, we achieve statistical significance at the .05 level at this analysis favoring the low-dose group. This same conclusion is reached using an LR-based sequential analysis, which maintains an independent increments structure in this case. Using the proportion of total deaths observed by the last analysis time to measure information accrued, the LR test with OF test-statistic boundaries $\{4.68, 2.35, 1.68\}$ also rejects the null hypothesis at the 1989 analysis time, with observed test statistics $\{1.27, 1.65, 2.04\}$.

Simulations verifying the properties of the $\mathcal{T}$ test statistic used in conjunction with this type of study design and the OF spending function assumed uniform$(0, 1)$ entry times for patients randomized to two treatment arms. Once entered, the exponential distribution was used to simulate survival times of the two treatment groups with common mean of 1 year under the null hypothesis and an exponential alternative conferring an average of roughly 3 months of life saved around 3 years after patient recruiting began. Analyses planned at 1, 2, and 3 years from beginning of patient recruitment used the same type I error allocations as in the above example. In order to detect 3 months of life saved using three interim analyses with 90% power, a sample size of 291 per treatment group is required. In calculating this sample size, we compared the inverted variance of $\hat{\Delta}$ at the last analysis time for the distributions described to the TI required at the last analysis time. After 500 repetitions, we estimated a type I error of 4.6% and power of 89.8% under these conditions.

6. Discussion

We have indicated how the YLS statistic can be incorporated in the sequential clinical trials setting. In cases where the upper limits of integration in the sequential tests are equivalent at the interim analyses, the sequential boundaries may be calculated with widely available software since an independent increments structure is present for these statistics. In cases where the primary event of interest takes a longer time to occur, the upper limits of integration of our sequential tests may increase with each additional analysis time to account for increases in the range of interesting observable survival differences. In this case, we may use a simulation strategy to calculate sequential boundaries for our tests. In either case little additional work is required to gain the benefits associated with each test in the sequential survival setting.

The sequentially monitored $\mathcal{T}$ test statistic gives an alternative to the LR test that may be more appropriate when

treatment differences are more readily described through YLS over some period of follow-up. Also Pepe and Fleming (1989) have demonstrated in their research that tests focused on differences between integrated survival curves are more powerful than the LR in many situations where proportional hazards assumptions are violated. In the presence of censoring, Murray's (1994) thesis introduces an adjusted YLS statistic that increases power when prognostic covariates are available. This modified version of the YLS statistic also reduces bias associated with informative censoring when covariates related to the censoring mechanism are available. In Murray's thesis, it has been shown that the adjusted version of the test has a similar theoretical structure to the original $\mathcal{T}$ test in the sequential trials framework, i.e., an independent increments structure to the adjusted test statistic is also available when the upper limits of integration τ remain constant across analysis times. In her thesis, Murray has also completely specified the joint distribution of the adjusted $\mathcal{T}$ statistics across analysis times so that this test can be extended to more general scenarios. Since the adjusted test corrects power loss and bias in the presence of censoring, its use would be most advantageous during the early periods of a clinical trial when censoring is heaviest.

Acknowledgements

This work was supported in part by the National Cancer Institute, the National Institute of Allergy and Infectious Diseases, the National Institutes of Health, the American Cancer Society, and the International Breast Cancer Study Group.

Résumé

Cette recherche propose le développement de stratégies nonparamétriques pour monitoriser de façon séquentielle les données d'essais cliniques, où l'accent est mis sur la détection d'années de vie gagnées. Le test statistique proposé est focalisé sur les différences intégrées des estimations de survie durant la période de temps considéré. Dans de nombreuses applications pratiques, le test statistique présente une structure de covariances avec incréments indépendants. Dès lors, avec un peu d'effort, on pourrait appliquer ces tests avec les méthodes existantes. Dans le cas où une structure de covariances d'incréments indépendants existe, nous montrons comment les données de l'essai clinique peuvent être monitorisées grâce aux tests statistiques proposés dans un protocole basé sur l'information. Le plan d'étude qui en résulte maintient les caractéristiques d'opération stochastiques souhaitées, quelle que soit la forme des courbes de survie comparées. Ceci présente un avantage sur la méthode classique du logrank, puisque des hypothèses plus restrictives en relation avec le comportement des fonctions de risque sont nécessaires pour garantir la finesse souhaitée du test. Des recommandations pour monitoriser séquentiellement les résultats des essais cliniques dans le cas d'incréments non-indépendants sont également proposées avec un exemple.

References

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

Breslow, N. and Crowley, J. (1974). A large sample study of the life table and product limit estimates under random censorship. *Annals of Statistics* **2,** 437–453.

Fischl, M. A., Parker, L. B., Petinelli, C., et al. (1990). A randomized controlled trial of a reduced daily dose of zidovudine in patients with the Aquired Immunodeficiency Syndrome. *The New England Journal of Medicine* **323,** 1009–1014.

Fleming, T. R. and Harrington, D. P. (1991). *Counting Processes and Survival Analysis.* New York: Wiley.

Gu, M. G. and Lai, T. L. (1991). Weak convergence of time-sequential censored rank statistics with applications to sequential testing in clinical trials. *Annals of Statistics* **19,** 1403–1433.

Jennison, C. and Turnbull, B. W. (1985). Repeated confidence intervals for the median survival time. *Biometrika* **72,** 619–625.

Kim, K., Boucher, H., and Tsiatis, A. A. (1995). Design and analysis of group sequential log-rank tests in maximum duration versus information trials. *Biometrics* **51,** 988–1000.

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

Lan, K. K. G. and Zucker, D. M. (1993). Sequential monitoring of clinical trials: The role of information and Brownian motion. *Statistics in Medicine* **12,** 753–765.

Murray, S. (1994). Nonparametric estimation and testing for survival data in the two sample censored data problem incorporating longitudinal covariates. Sc.D. dissertation, Department of Biostatistics, Harvard University, Cambridge, MA.

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

Pepe, M. S. and Fleming, T. R. (1989). Weighted Kaplan–Meier statistics—A class of distance tests for censored survival data. *Biometrics* **45,** 497–507.

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

Scharfstein, D. O., Tsiatis, A. A., and Robins, J. M. (1997). Semiparametric efficiency and its implication on the design and analysis of group-sequential studies. *Journal of the American Statistical Association* **92,** 1342–1350.

Schervish, M. J. (1984). Multivariate normal probabilities with error bound (with corrections in 1985). *Applied Statistics* **33,** 81–94.

Sellke, T. and Siegmund, D. (1983). Sequential analysis of the proportional hazards model. *Biometrika* **70,** 315–326.

Slud, E. V. (1984). Sequential linear rank tests for two-sample censored survival data. *Annals of Statistics* **12,** 551–571.

Tsiatis, A. A. (1981). The asymptotic joint distribution of the efficient scores test for the proportional hazards model calculated over time. *Biometrika* **68,** 311–315.

Tsiatis, A. A. (1982). Repeated significance testing for a general class of statistics used in censored survival analysis. *Journal of the American Statistical Association* **77,** 855–861.

Tsiatis, A. A., Boucher, H., and Kim, K. (1995). Sequential methods for parametric survival models. *Biometrika* **82,** 165–173.

Received June 1998. *Revised March* 1999.
Accepted March 1999.

Appendix

In the notational framework of Section 2, we define the sequential behavior of $\hat{S}(t_1, x_1)$ and $\hat{S}(t_2, x_2)$ using martingale methodology. We first note a result derived in Breslow and Crowley (1974) that $n^{1/2}[\hat{S}(t,x) - \exp\{-\hat{\Lambda}(t,x)\}] \xrightarrow{\mathcal{P}} 0$, where $\hat{\Lambda}(t,x)$ denotes the Nelson–Aalen (NA) estimate of the cumulative hazard function. For the moment, we shall restrict the problem to the one-sample case. We may rewrite the NA estimator in counting process notation as

$$\hat{\Lambda}(t,x) = \int_0^x \{Y(t,u)\}^{-1} dN(t,u),$$

where

$$\mathrm{N}(t,x) = \sum_{i=1}^{n} I\{X_i(t) \le x, \Delta_i(t) = 1\} \quad \text{for } 0 \le x \le t$$

and

$$Y(t,x) = \sum_{i=1}^{n} I\{X_i(t) \ge x\}.$$

So

$$\{\hat{\Lambda}(t,x) - \Lambda(x)\} = \int_0^x dM(t,u)/Y(t,u),$$

where

$$M(t,x) = N(t,x) - \int_0^x \lambda(u) Y(t,u) du.$$

If we define

$$Z_n(t,x) = n^{1/2} \int_0^x dM(t,u)/Y(t,u) \quad \text{for } x \le t$$

and $t = t_1, t_2$, then according to Tsiatis et al. (1995), the two processes $Z_n(t_1, x)$ and $Z_n(t_2, x)$ as functions of x are martingale processes with respect to the filtration that includes the time of entry as well as all the failure and censoring information that is observed to occur for all the individuals within x units after they enter the study. This filtration is similar to that used by Sellke and Siegmund (1983). An application of the multivariate central limit theorem described in Fleming and Harrington's (1991) theorems 5.3.4 and 5.3.5 shows us these processes are jointly normal mean zero processes with

$$\begin{aligned}
&\mathrm{cov}\{Z_n(t_1,x_1), Z_n(t_2,x_2)\} \\
&\quad = \lim_{n\to\infty} \int_0^{\min(x_1,x_2)} n\{Y(t_2,u)\}^{-1} \lambda(u) du \\
&\quad = \int_0^{\min(x_1,x_2)} [P\{X(t_2) > u\}]^{-1} \lambda(u) du.
\end{aligned}$$

Since $X_i(t_2) = \min(T_i, V_i, t_2 - E_i)$, $P\{X_i(t_2) > x\} = P(T_i > x, V_i > x, t_2 - E_i > x) = S(x)C(t_2,x)$, where $C(t_2,x) = P(E_i < t_2 - x, V_i > x)$, we find that

$$\mathrm{cov}\{Z_n(t_1,x_1), Z_n(t_2,x_2)\} = \int_0^{\min(x_1,x_2)} \frac{\lambda(u)du}{S(u)C(t_2,u)}$$

asymptotically. Note that $n(t)/n = \Sigma_{i=1}^{n} I(E_i \le t)/n$ converges in probability to $P(E_i \le t) = G(t)$. Using the delta method together with the Breslow and Crowley result gives us asymptotically

$$\begin{aligned}
&\mathrm{cov}\left[\{\hat{S}(t_1,x_1) - S(x_1)\}, \{\hat{S}(t_2,x_2) - S(x_2)\}\right] \\
&\quad \approx n^{-1} S(x_1) S(x_2) \mathrm{cov}\{Z_n(t_1,x_1), Z_n(t_2,x_2)\} \\
&\quad \approx \{n(t_2)\}^{-1} S(x_1) S(x_2) \\
&\qquad \times \int_0^{\min(x_1,x_2)} \{H(t_2,u)S(u)\}^{-1} \lambda(u) du,
\end{aligned}$$

as claimed in Section 2.