Group Sequential Designs for Monitoring Survival Probabilities

D. Y. Lin; L. Shen; Z. Ying; N. E. Breslow


Stable URL:
http://links.jstor.org/sici?sici=0006-341X%28199609%2952%3A3C1033%3AGSDFMS%3E2.0.CO%3B2-Y

*Biometrics* is currently published by International Biometric Society.

Your use of the JSTOR archive indicates your acceptance of JSTOR’s Terms and Conditions of Use, available at http://www.jstor.org/about/terms.html. JSTOR’s Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at http://www.jstor.org/journals/ibs.html.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

JSTOR is an independent not-for-profit organization dedicated to creating and preserving a digital archive of scholarly journals. For more information regarding JSTOR, please contact support@jstor.org.
Group Sequential Designs for Monitoring Survival Probabilities

D. Y. Lin,1 L. Shen,1 Z. Ying,2 and N. E. Breslow1

1Department of Biostatistics, Box 357232, University of Washington, Seattle, Washington 98195, U.S.A.
2Department of Statistics, Hill Center, Busch Campus, Rutgers University, Piscataway, New Jersey 08855, U.S.A.

SUMMARY

In the present paper we derive the asymptotic joint distribution of the Nelson–Aalen estimators for the cumulative hazard function (or the Kaplan–Meier estimators for the survival function) calculated at different (possibly random) calendar time points of a survival study. This result is used to develop group sequential designs for clinical trials in which patients are enrolled over an extended period and the survival probability at a given failure time point is the parameter of primary interest. The proposed designs allow monitoring of the survival probability (or probabilities) at interim looks so that the study can be terminated when there is sufficient evidence for a positive or negative result. A detailed illustration is provided with a problem from the National Wilms Tumor Study Group, which motivated this research.

1. Introduction

For censored survival data, linear rank statistics are commonly used to compare two survival distributions or to assess the goodness-of-fit of a hypothesized distribution. These statistics pertain to the survival distribution over the entire follow-up period. In many clinical trials, however, the measure of efficacy is a long-term, say \(\tau\)-year, survival rate, and the shape of the distribution before \(\tau\) is largely irrelevant. For ethical and scientific reasons it is often desirable to review the accumulating survival data during the course of a clinical trial so that the study can be terminated as soon as there is conclusive evidence. Due to staggered patient entry and loss to follow-up, the \(\tau\)-year survival probability must be estimated by the Kaplan–Meier estimator rather than the binomial sample proportion at interim looks. However, sequential methods based on the Kaplan–Meier estimator are not currently available. By contrast, sequential linear rank tests have been well studied (Jones and Whitehead, 1979; Slud and Wei, 1982; Tsiatis, 1982; Slud, 1984; Gu and Lai, 1991).

Our interest in developing sequential methods for monitoring survival probabilities stemmed from the design of the fifth protocol of the National Wilms Tumor Study Group (NWTSG) (D’Angio et al., 1989). Wilms tumor is an embryonal tumor of the kidney that affects about one in every 10,000 children under the age of 15 years. The current treatment consists of combination chemotherapy with or without radiation therapy. In the fifth NWTSG protocol study, no further treatment is planned after surgical resection for one subgroup of patients under two years of age with a tumor of “favorable” histology. The proposed criterion for evaluating this “surgery only” treatment arm is the proportion of patients who remain continuously disease-free and alive for two years following surgery. This has long been used by the NWTSG as an endpoint because very few relapses occur after the two-year point. (A linear rank test comparing the new treatment with historical data would be less appropriate since it may be sensitive to early differences in relapse rates which do not hold up at two years.) The investigators expect the two-year relapse-free survival in the “surgery

Key words: Censoring; Clinical trial; Data monitoring; Interim analysis; Kaplan–Meier estimator; Nelson–Aalen estimator.

1033
only” group to be as high as 95%, which is the outcome observed from the first four protocols (see Figure 1). They plan to enroll about 200 patients over five years, and wish to examine the accumulating data on the two-year relapse-free survival at their annual meetings during years two to seven of the study. If interim results turn out worse than anticipated, it will be mandatory to stop further patient entry immediately and to commence chemotherapy for those already enrolled who may be in danger of relapse. Since the patients will enter the study sequentially over five years, some patients will have been followed for less than two years at the interim analyses, which entails that the two-year relapse-free survival probability be estimated by the Kaplan–Meier estimators even if the loss to follow-up is negligible. Furthermore, since no patient will have been in the study for two years at the first look, it will be more sensible to examine the one-year relapse-free survival probability at that time.

![Figure 1](image)

**Figure 1.** Relapse-free survival for Wilms tumor patients under two years of age at diagnosis with Stage I disease of “favorable” histology.

In order to devise a data monitoring scheme such as the one required for the NWTSG fifth protocol study, it is necessary to ascertain the joint distribution of the Kaplan–Meier estimators (or related quantities) calculated over interim analyses. This distributional result is derived in the Appendix and described in the next section. In Section 3 we show how to apply standard group sequential methods to a single survival probability. In Section 4 we develop and evaluate a group sequential design for the NWTSG which involves monitoring both the one-year and two-year survival probabilities. In Section 5 we discuss further applications of the proposed methodology.

### 2. Sequential Analysis of Survival Function

Suppose that \( n \) patients enter the study at times \( Y_1, \ldots, Y_n \): Associated with the patient entering at \( Y_i \) is a random pair \( (T_i, C_i) \), where \( T_i \) and \( C_i \) denote, respectively, the failure time and censoring time, both of which are measured from study entry. We assume that the failure time \( T_i \) is independent of the censoring time \( C_i \) and entry time \( Y_i \) and that the random triples \( (Y_i, T_i, C_i) \) \( (i = 1, \ldots, n) \) are independent and identically distributed.

When the data are examined at calendar time \( t \), for \( i = 1, \ldots, n \), we observe the time to failure or censoring \( X_i(t) \) and the failure indicator \( \Delta_i(t) \), where \( X_i(t) = T_i \wedge C_i \wedge (t - Y_i)^+ \) and \( \Delta_i(t) = I\{T_i \leq C_i \wedge (t - Y_i)^+\} \). Here and in the sequel we adopt the convention that \( a \wedge b = \min(a, b) \), \( a \vee b = \max(a, b) \), \( a^+ = \max(0, a) \), and \( I(A) \) is the indicator function for the event \( A \). Based on the data \( \{X_i(t), \Delta_i(t)\} \), \( i = 1, \ldots, n \), the Nelson–Aalen estimator for the cumulative hazard function \( \Lambda(x) = \int_0^x \lambda(u) \, du \) is

\[
\hat{\Lambda}(x; t) = \sum_{\{i: X_i(t) \leq x\}} \frac{\Delta_i(t)}{R_i(t)}.
\]
where \( R_i(t) = \sum_{j=1}^{n} I\{X_j(t) \geq X_i(t)\} \). For a fixed \( t \) it is well-known that the process \( n^{1/2}\{\hat{\Lambda}(x; t) - \Lambda(x)\}, x < t, \) converges weakly to a zero-mean Gaussian process with independent increments and with variance function

\[
\sigma^2(x; t) = \int_0^x \frac{\lambda(u) du}{\pi(u; t)}, \tag{2.1}
\]

where \( \pi(u; t) = \Pr\{X_1(t) \geq u\} \), and \( \sigma^2(x; t) \) can be consistently estimated by

\[
\hat{\sigma}^2(x; t) = \sum_{\{i: X_i(t) \leq x\}} \frac{\Delta_i(t)}{R_i^2(t)/n} \tag{2.2}
\]

(Breslow and Crowley, 1974, Theorem 4; Fleming and Harrington, 1991, Theorem 3.2.2).

In order to design and analyze sequential trials, we need to ascertain the joint distribution of \( \hat{\Lambda}(x; t) \) over the calendar time \( t \). If survival probabilities at different failure time points are tested at different interim looks, then it is also necessary to evaluate the joint distribution between \( \hat{\Lambda}(x; t) \) and \( \hat{\Lambda}(y; s), x \neq y, t \neq s \). Thus, it is desirable to establish the weak convergence of \( \hat{\Lambda}(x; t) \) as a two-parameter process indexed by the failure time \( x \) and calendar time \( t \). We show in the Appendix that the two-parameter process \( n^{1/2}\{\hat{\Lambda}(x; t) - \Lambda(x)\} \), \( x < t, \) converges weakly to a zero-mean Gaussian process \( W(x; t) \) with \( \text{cov}\{W(x; t), W(y; s)\} = \sigma^2(x \land y; t \lor s) \).

Let \( I(x; t) = 1/\sigma^2(x; t) \) and \( U(x; t) = I(x; t)n^{1/2}\hat{\Lambda}(x; t) - \Lambda(x) \). Note that \( U(x; t) \) and \( I(x; t) \) correspond to the score function and Fisher information in the likelihood setting. Asymptotically, \( \text{var}\{U(x; t)\} = I(x; t) \) and \( \text{cov}\{U(x; t), U(y; s)\} = I(x; t)I(y; s)/I(x; t \lor s) = I(x; t \land s) \). Thus, for a fixed failure time \( x, \) the “score process” \( U(x; t) \) (indexed by \( t \)) is asymptotically a (zero-drift) Brownian motion in the time scale of the “information” \( I(x; t) \). This “martingale” property has been known to hold for score processes of many commonly used parametric and semiparametric models, such as the cumulative sum of random variables and Cox’s partial-likelihood score process, the latter containing the (numerator of the) log-rank statistic as a special case (Siegmund, 1985, Chapter V). A number of methods are currently available to monitor Brownian motions.

If \( x \neq y, \) then asymptotically \( \text{cov}\{U(x; t), U(y; s)\} = I(x; t)I(y; s)/I(x \land y; t \lor s) \), which cannot be simplified. Because more than one time scale is involved here, the independent increment structure over the analysis time cannot hold. Thus, it is less straightforward to monitor multiple survival probabilities. Nevertheless, the general theory described previously enables one to evaluate explicitly the asymptotic joint distribution between \( U(x; t) \) and \( U(y; s), x \neq y, \) from which appropriate critical values can be obtained by the Slud–Wei method.

It is useful to consider certain transformations of \( \hat{\Lambda}(x; t) \). For example, a natural estimator for the survival function \( S(x) = e^{-\Lambda(x; t)} \), which is asymptotically equivalent to the Kaplan–Meier estimator. Let \( g(\cdot) \) be a known function whose derivative \( \hat{g}(\cdot) \) is nonzero and continuous. By the functional \( \delta \)-method, the transformed process \( n^{1/2}\{\hat{g}(\hat{\Lambda}(x; t)) - \hat{g}(\Lambda(x))\}, x < t, \) converges weakly to a zero-mean Gaussian process \( \hat{W}(x; t) \) with \( \text{cov}\{\hat{W}(x; t), \hat{W}(y; s)\} = \hat{g}(\hat{\Lambda}(x))\hat{g}(\hat{\Lambda}(y))\text{cov}\{\hat{W}(x; t), \hat{W}(y; s)\} \). Note that \( \text{corr}\{\hat{W}(x; t), \hat{W}(y; s)\} = \text{corr}\{W(x; t), W(y; s)\} \), which implies that the sequential boundary values (for standardized test statistics) derived from the asymptotic joint distribution of \( W(x; t) \) will be the same regardless of the choice of \( g(\cdot) \). Because \( \log \hat{\Lambda} \) takes the value over the entire real line whereas \( \hat{\Lambda} \) is restricted to be nonnegative, the asymptotic normal approximation turns out to be more accurate for \( \log \hat{\Lambda} \) than for \( \hat{\Lambda} \) and \( e^{-\hat{\Lambda}} \) (Kalbfleisch and Prentice, 1980, pp. 14–15; Bie, Borgan, and Liestøl, 1987; Lin, Fleming, and Wei, 1994). Therefore, at calendar time \( t, \) we shall test \( H_0: S(x) = S_0(x) \) or \( \Lambda(x) = \Lambda_0(x) \) using the standard-normal statistic

\[
Z(x; t) = \frac{n^{1/2}\{\log \hat{\Lambda}(x; t) - \log \Lambda_0(x)\}\hat{\Lambda}(x; t)}{\hat{\sigma}^2(x; t)^{1/2}}. \tag{2.3}
\]

Note that the covariance (correlation) between \( Z(x; t) \) and \( Z(y; s) \) can be consistently estimated by

\[
\frac{\hat{\sigma}^2(x \land y; t \lor s)}{\hat{\sigma}^2(x; t)\hat{\sigma}^2(y; s)^{1/2}}. \tag{2.3}
\]

3. Sequential Designs for a Single Survival Probability

As shown in the previous section, for a fixed failure time \( x, \) say two years, the “score statistic” \( U(t) \equiv U(2; t) \) is asymptotically a Brownian motion in the time scale of the “information” \( I(t) \equiv \)
$I(2; t)$. Thus, for monitoring $S(2)$ alone, the existing sequential designs (Pocock, 1977; O’Brien and Fleming, 1979; Lan and DeMets, 1983; Whitehead, 1983) can be utilized.

The null hypothesis of interest is $H_0$: $S(2) = S_0(2)$ or $\theta \equiv \log(S(2)/S_0(2)) = 0$. At the interim analysis times $t = t_1, t_2, \ldots$, we calculate $U(t)$ and $I(t)$, and reject $H_0$ at $t_k$ if $U(t_k)/I^{1/2}(t_k)$ or $Z(t_k) \equiv Z(2; t_k)$ exceeds $c_k$ for suitably chosen critical values $c_1, c_2, \ldots$. Since the increments $\{U(t_k) - U(t_{k-1})\}$ of $U(t)$ are asymptotically independent, commercially available software packages such as EaSt or PEST, or one of the many privately written sequential analysis programs, can be used to find the $c_k$.

It is easy and accurate to formulate a design as follows: perform the interim analyses when $I(t) = I_1, \ldots, I_K$ and reject $H_0$ if $Z_k > c_k$, where $Z_k$ is the value of $Z(t)$ when $I(t) = I_k$. Such a design will be chosen to satisfy a prespecified power requirement. The hard part is anticipating the values $(t_1, \ldots, t_K)$ which will yield $(I_1, \ldots, I_K)$. As is evident by (2.1), $I(t)$ depends on the distributions of entry time, failure time and censoring time.

At the design stage it is acceptable to make distributional assumptions in order to find out the consequences of looking at various choices of times $t_1, \ldots, t_K$. (The calculation of the $I_k$ given the distributional assumptions will be illustrated in the next section.) After some searching, a design with the appropriate power, subject to assumptions, will be found. At the interim analyses, however, the true values of $I(t_k)$ should be used rather than their anticipated values $I_k$. Thus, the actual critical value $c_k$ will become apparent only at the time of the $k$th interim look. The packages EaSt and PEST both operate in this way. In light of the rate of accumulating information, it may become necessary to adjust the timetable $(t_1, \ldots, t_K)$ to make the sequence $\{I(t_1), \ldots, I(t_K)\}$ adhere more closely to its intended form.

4. Design of the National Wilms Tumor Study

In the original design of the fifth NWTSG protocol study, an approximate sample size of 200 with accrual period of five years was determined to ensure that a fixed-sample one-sided binomial test with 20% significance level for testing the null hypothesis of 95% two-year relapse-free survival would have power of 95% to detect a 90% relapse-free survival. (To be specific, 184 or 203 patients would be required depending on whether or not a continuity correction is used in the normal approximation.) The rationale for setting the Type II error lower than the Type I error is that it would be highly unethical to withhold chemotherapy, which is a proven efficacious treatment with minimal acute toxicities, if the surgery-only strategy fails in its promise, although it would be desirable to adopt that strategy if it indeed improves long-term survival. (In a sense, one is weighing more heavily the short-term risk of failure due to withholding treatment than the long-term benefit of avoiding the chronic toxicities associated with that treatment.) Of course, the Type I error of 20% for testing $H_0$: $S(2) = 0.95$ with power of 95% at $H_1$: $S(2) = 0.90$ is equivalent to the Type I error of 5% for testing $H_0$: $S(2) = 0.90$ with power of 80% at $H_1$: $S(2) = 0.95$.

As mentioned in Section 1, the investigators plan to examine the one-year relapse-free survival at the end of the second year of the study and to examine the two-year relapse-free survival every year after that until the end of the seventh year, by which time all the 200 patients will have been followed for at least two years. This monitoring plan involves two null hypotheses, one concerning $S(1)$ and one concerning $S(2)$. Based on historical data, about two-thirds of the failures in the first two years occur in the first year. Thus, the value of $S(1)$ is determined by $S(1) = S(2) + \{1 - S(2)\}/3$, which is approximately 0.9667 if $S(2) = 0.95$. At the first look, we shall test the secondary null hypothesis $H'_0$: $S(1) = 0.9667$ using the statistic $Z'_1 \equiv Z(1; 2)$; at the remaining five looks we shall test the primary null hypothesis $H_0$: $S(2) = 0.95$ using the statistics $Z'_{k} \equiv Z(2; t_k)$, $t_k = 3, 4, 5, 6, 7$. If either $H'_0$ or $H_0$ is rejected, we shall reject the umbrella null hypothesis $H'_0$: $S(1) = 0.9667$ and $S(2) = 0.95$, and terminate the trial with a negative conclusion. Our task is to find appropriate critical values $(c'_1, \ldots, c'_6)$ such that the overall null probability of any form of rejection of $H'_0$ is 20% and the overall power is about 95% at $H'_1$: $S(1) = 0.9333$ and $S(2) = 0.90$.

Because we now deal with an umbrella null hypothesis rather than a single hypothesis, the methodology described in the previous section is not applicable. The key to any sequential test is knowing the joint distribution of the test statistics over the interim looks, from which the critical values can be obtained by numerical integration or simulation. In our case, the joint distribution of $(Z'_1, \ldots, Z'_6)$ is encompassed by the general theory presented in Section 2.

At the design stage it is necessary to make certain assumptions in order to predict how the trial is likely to go. We anticipate that the accrual rate will be roughly constant over the five-year period and the loss to follow-up negligible. For design purposes, assume that the disease-free survival time follows a Weibull distribution with survival function $S(x) = \exp(-ax^b)$, where the
parameters $a$ and $b$ are determined by the hypothesized values of $S(1)$ and $S(2)$. We expect the Weibull distribution to be a fairly good approximation, especially since we are only using the first two years of the distribution.

We now evaluate the variances and information. Recall the variance expression $\sigma^2(x; t)$ given in (2.1). In the absence of loss to follow-up, $\sigma^2(x; t) = \Pr(T_1 \leq t - Y_1^* \geq u)$, which equals $\Pr(T_1 \geq u)\Pr(Y_1 \leq t - u)$ due to the independence of $Y_1$ and $T_1$. Under the Weibull failure time distribution and constant accrual rate, $\Pr(T_1 \geq u) = \exp(-au^b)$ and $\Pr(Y_1 \leq t - u) = \min\{(t - u)/5, 1\}$, $t \geq u$. Thus,

$$\sigma^2(x; t) = \int_0^x \frac{abu^{b-1} du}{\exp(-au^b) \min\{(t - u)/5, 1\}}, \quad x < t.$$ 

Note that $a = 0.0339$ and $b = 0.5974$ under $H_0^*$, $S(1) = 0.9667$ and $S(2) = 0.95$. By numerical integration, $[\sigma^2(1; 2), \sigma^2(2; 3), \ldots, \sigma^2(2; 7)] \approx [0.1105, 0.1286, 0.0844, 0.0634, 0.0546, 0.0526]$, or $[I(1; 2), I(2; 3), \ldots, I(2; 7)] \approx [9.051, 7.774, 11.848, 15.773, 18.304, 19.001]$. Note that $I(1; 2) > I(2; 3)$, which indicates that the amount of information for estimating $S(1)$ at $t = 2$ is larger than that of estimating $S(2)$ at $t = 3$ even though there is no information for estimating $S(2)$ at $t = 2$.

It is customary to allocate the overall Type I error among the interim looks according to the amount of information (Lan and DeMets, 1983). The first look, however, requires special treatment since it involves $S(1)$ rather than $S(2)$. Because there will be very few failures by $t = 2$ even if $S(1) = 0.90$, we shall let the Type I error at the first look be $\alpha_1 = 0.01$. The remaining 0.19 Type I error will be distributed among the second–sixth looks according to the cumulative error function $\alpha^*(t) = 0.19\{I(2; t)/I(2; 7)\}^2$, $t = 3 \sim 7$ (Kim and DeMets, 1987), which yields the sequence of Type I errors ($\alpha_2, \ldots, \alpha_6$) = (0.032, 0.042, 0.057, 0.045, 0.014).

We also calculate $\sigma^2(1; t)$, $t = 3 \sim 7$, which together with the values of $\sigma^2(1; 2)$ and $\sigma^2(2; t)$ ($t = 3 \sim 7$) given above produce the following asymptotic covariance matrix for $(Z_1^*, \ldots, Z_6^*)$:

$$\begin{bmatrix}
1 & 0.559 & 1 \\
0.496 & 0.810 & 1 \\
0.447 & 0.702 & 0.867 & 1 \\
0.444 & 0.652 & 0.804 & 0.928 & 1 \\
0.452 & 0.640 & 0.790 & 0.911 & 0.982 & 1
\end{bmatrix}.$$ (3.1)

Then we obtain $(c_1^*, \ldots, c_6^*) = (2.326, 1.805, 1.498, 1.202, 1.043, 1.031)$ by using a multivariate normal integration algorithm (Schervish, 1984) to solve recursively the equations

$$\Pr\{G_1 < c_1^*, \ldots, G_{k-1} < c_{k-1}^*, G_k > c_k^*\} = \alpha_k, \quad k = 1, \ldots, 6,$$

where $(G_1, \ldots, G_6)$ is zero-mean normal with covariance matrix given in (3.1).

For power and sample size calculations we need to evaluate the distributions of the test statistics under alternative hypotheses. Under $S(\cdot) = S_1(\cdot)$ or $\Lambda(\cdot) = \Lambda_1(\cdot)$, the vector of test statistics $(Z(1; 2), Z(2; 3), \ldots, Z(2; 7))$ is approximately joint normal with unit variances and with means

$$\mu(x; t) = \frac{n^{1/2}\{\log \Lambda_1(x) - \log \Lambda_0(x)\} \Lambda_1(x)}{(\sigma^2(x; t))^{1/2}},$$

where $\sigma^2(x; t)$ are evaluated under $S_1(\cdot)$. Upon evaluations of $\sigma^2(1; t)$, $t = 2 \sim 7$, and $\sigma^2(2; t)$, $t = 3 \sim 7$, for $S_1(\cdot)$, we can obtain the values of $\mu(x; t)$ as well as the covariance matrix for the six test statistics. Power calculations can then be performed through multivariate normal integration. For example, under $H_1^*$: $S(1) = 0.9333$ and $S(2) = 0.90$ and with $n = 200$, the approximate powers at the six looks (conditional on not having stopped at an earlier look) are found to be 0.189, 0.429, 0.245, 0.100, 0.023, and 0.003 with a total of 0.989. This total is surprisingly high since the fixed-sample binomial test (without continuity correction) has approximate power of 0.95 for $n = 184$.

The latter power, however, is based on the direct normal approximation to the sample proportion, which is equivalent to the Kaplan–Meier estimator in the absence of censoring. It is therefore of interest to approximate the power of the sequential testing procedure based on the test statistics

$$\hat{Z}(x; t) = \frac{n^{1/2}\{S_0(x) - e^{-\bar{\Lambda}(x; t)}\}}{e^{-\bar{\Lambda}(x; t)}\{\hat{\sigma}^2(x; t)\}^{1/2}},$$

where $\hat{\sigma}^2(x; t)$ are evaluated under $S_1(\cdot)$.
The overall power is 0.927. The reason that the approximate power based on \( \hat{Z} \) is lower than that of \( Z \) is because \( \hat{Z} \) has a smaller approximate mean than \( Z \) under \( S_1(\cdot) \).

Simulation experiments were conducted to assess the performance of the above design. Two-hundred entry times were generated according to the uniform(0, 5) distribution; the failure times were generated from the Weibull distributions with \( a = 0.0339 \) and \( b = 0.5974 \) under \( H_0^* \); \( S(1) = 0.9667 \) and \( S(2) = 0.95 \), and with \( a = 0.0690 \) and \( b = 0.6108 \) under \( H_1^* \); \( S(1) = 0.9333 \) and \( S(2) = 0.90 \). The empirical rejection probabilities based on 1,000,000 simulation samples are displayed in Table 1. The attained overall Type I error is remarkably close to the target value. The attained Type I errors at the six looks differ from their respective nominal levels by no more than 0.008, though the relative differences are quite large at the first and last looks. The attained powers are not in complete agreement with the theoretical values, though the attained overall power of 0.961 is close to the original target of 0.95. (The theoretical powers are largely irrelevant as long as one is content with the attained levels.) Note that the attained (cumulative) probability of stopping by the fourth year is about 80% under \( H_1^* \); \( S(1) = 0.9333 \) and \( S(2) = 0.90 \). If \( S(2) = 0.80 \), then the attained cumulative powers are 0.853, 0.996, and 1.000 at \( t = 2, 3, \) and 4. The high probabilities of early stopping for a two-year relapse-free survival of 90% or less are very appealing since the main purpose of the sequential monitoring is to terminate the study as early as possible if the surgery-only treatment turns out worse than anticipated.

The results of Table 1 indicate that proper error structure will be maintained with the use of predetermined critical values if the entry time and failure time distributions are correctly specified. In the actual implementation of the sequential design, it is desirable and possible to avoid any distributional assumptions. We shall estimate the real covariance matrix of \( (Z_1^*, \ldots, Z_6^*) \) from the data collected so far at the interim looks by using formulae (2.2) and (2.3) and adjust the critical values accordingly (Slud and Wei, 1982). This will lead to an asymptotically nonparametric sequential test, which will preserve the specified Type I error even if the initial guesses about the distributions of failure time, entry time, and censoring are wildly wrong.

### Table 1

**Operating characteristics of the NWTSG design**

<table>
<thead>
<tr>
<th>Time of analysis (years)</th>
<th>Critical value</th>
<th>Type I error</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( t = 2 )</td>
<td>( t = 3 )</td>
<td>( t = 4 )</td>
</tr>
<tr>
<td>Critical value</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Type I error</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nominal</td>
<td>2.326</td>
<td>1.805</td>
<td>1.498</td>
</tr>
<tr>
<td>Attained</td>
<td>0.010</td>
<td>0.032</td>
<td>0.042</td>
</tr>
<tr>
<td>Power</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Theoretical</td>
<td>0.189</td>
<td>0.429</td>
<td>0.245</td>
</tr>
<tr>
<td>Attained</td>
<td>0.237</td>
<td>0.343</td>
<td>0.214</td>
</tr>
</tbody>
</table>

5. Discussion

It is straightforward to extend the one-sample results obtained in the previous sections to the two-sample case. The key is again to derive the joint distribution for the test statistics over the calendar time. Let \( \Lambda_1(\cdot) \) and \( \Lambda_2(\cdot) \) be the cumulative hazard functions for groups 1 and 2, respectively. The notation of \( \Lambda_1, \Lambda_2, \sigma^2_1, \) and \( \sigma^2_2 \) is similarly defined. In addition, let \( n \) be the maximal number of patients in the study and let \( \rho \) be the limiting proportion of patients in group 1. Then the process \( n^{1/2} \left[ (\Lambda_1(x; t) - \Lambda_2(x; t)) - (\Lambda_1(x) - \Lambda_2(x)) \right], \quad x < t, \) converges weakly to a zero-mean Gaussian process \( \mathcal{W}(x; t) \) with \( \text{cov}\{\mathcal{W}(x; t), \mathcal{W}(y; s)\} = \rho^{-1} \sigma^2_1(x \wedge y; t \vee s) + (1 - \rho)^{-1} \sigma^2_2(x \wedge y; t \vee s) \).

In the NWTSG study, the investigators are mainly interested in the two-year disease-free survival because it essentially constitutes cure. Recently, Lee and Sather (1995, p. 756) argued that for childhood cancer more generally “one might be more interested in testing the differences in cure rates rather than other types of differences in failure distributions.” In comparison with the two sequential methods discussed by Lee and Sather, our approach has the clear advantage that we do not parameterize the failure time distribution or impose the proportional hazards structure.

The results derived in our paper have potential applications in other disease areas as well. For example, Harris and Hellman (1986, p. 925) emphasize that for breast cancer treatment “effects seen
on the early portion of survival curves may not only be premature, but may also be misleading."
In a recent study on the efficacy of zidovudine in reducing the risk of maternal–infant transmission of
human immunodeficiency virus, the proportion infected at 18 months was used as the primary
endpoint because infection occurring after 18 months cannot be attributed to maternal transmission
and because the exact timing of the infection (within the 18 months period) is immaterial (Connor
et al., 1994).

In Section 4 we showed how to use two sequences of statistics to test an umbrella null hypothesis
concerning two survival probabilities. This type of problem also arises in other contexts of clinical
trials. For example, in long-term clinical trials it may be desirable to examine a quickly available
“surrogate endpoint” at early interim analyses. Another example is the simultaneous monitoring of
efficacy and safety measures. The key to developing appropriate sequential designs in these settings
is again to ascertain the joint distribution for the two sequences of test statistics. This can usually
be achieved by expressing each test statistic as a sum of i.i.d. random variables so that the joint
distribution will follow from the multivariate central limit theorem; see Lin (1991) and Su and
Lachin (1992) for some important examples. Although the overall probability of rejection based
on two sequences of test statistics can be evaluated by multivariate normal integration (Schervish,
1984), the effect of testing a secondary hypothesis during the trial on the post-trial analysis of the
primary endpoint requires future research.

ACKNOWLEDGEMENTS

This research was supported by the National Institutes of Health grants GM-47845, AI-291968,
and CA-42326. The authors thank the associate editor and referees for their helpful comments.

RÉSUMÉ

Dans cet article, nous développons le calcul de la distribution asymptotique de l’estimateur de
Nelson–Aalen pour la fonction donnant le taux cumulé d’événements (ou de l’estimateur de Kaplan
Meier pour la fonction de survie) calculée à différents instants (éventuellement aléatoires) d’une
étude de survie. Ce résultat est utilisé pour développer des plans d’analyses séquentiels groupées
pour des essais cliniques dans lesquels les malades sont inclus pendant une longue période et où le
paramètre de premier intérêt est la probabilité de survivre à un instant donné. Les plans d’analyse
proposés permettent de prévoir des analyses intermédiaires sur les taux de survie de telle sorte
que l’étude peut être interrompue dès qu’il y a une évidence suffisante en faveur d’un résultat
quelconque positif ou négatif. Une illustration détaillée est donnée à partir d’un problème, posé
par le National Wilms’ Tumor Study Group, et qui est à l’origine de cette recherche.

REFERENCES

Bie, O., Borga, Ø., and Liestøl, K. (1987). Confidence intervals and confidence bands for the
cumulative hazard rate function and their small sample properties. Scandinavian Journal of
Statistics 14, 221–233.
of human immunodeficiency virus type 1 with zidovudine treatment. New England Journal of
Medicine 331, 1173–1180.
Wiley.
applications to sequential testing in clinical trials. The Annals of Statistics 19, 1403–1433.
Harris, J. R. and Hellman, S. (1986). Observations on survival curve analysis with particular refer-
cence to breast cancer treatment. Cancer 57, 925–928.
Jones, D. and Whitehead, J. (1979). Sequential forms of the log rank and modified Wilcoxon tests
York: Wiley.


Received January 1995; revised December 1995; accepted February 1996.

**APPENDIX**

**Weak Convergence of Sequential Nelson–Aalen Estimator**

It is easy to see that

\[
\hat{\Lambda}(x; t) = \frac{\sum_{i=1}^{n} I\{T_i \leq C_i \wedge (t - Y_i)^+ \} I(T_i \leq x)}{\sum_{i=1}^{n} I\{T_i \wedge C_j \wedge (t - Y_j)^+ \geq T_i\}}
\]

\[
= \int_{0}^{x} \frac{\sum_{i=1}^{n} I\{C_i \wedge (t - Y_i)^+ \geq u\} dI(T_i \leq u)}{\sum_{j=1}^{n} \sum_{i=1}^{n} I\{T_j \wedge C_j \wedge (t - Y_j)^+ \geq u\}}.
\]

Letting \( D(x; t) = n^{1/2}\{\hat{\Lambda}(x; t) - \Lambda(x)\} \), we observe that, for \( x < t \),

\[
D(x; t) = n^{1/2} \int_{0}^{x} \frac{\sum_{i=1}^{n} I\{C_i \wedge (t - Y_i)^+ \geq u\} dM_i(u)}{\sum_{j=1}^{n} \sum_{i=1}^{n} I\{T_j \wedge C_j \wedge (t - Y_j)^+ \geq u\}}, \tag{A.1}
\]
where $M_t(x) = I(T_t \leq x) - \int_0^x I(T_t \geq u) \lambda(u) \, du$ is a martingale in $x$ (which is free of $t$). Thus, for fixed $t_1$ and $t_2 \geq t_1$, the predictable covariation process between $D(x; t_1)$ and $D(y; t_2)$ is

$$
\begin{align*}
\sum_{i=1}^{n} I\{C_i \wedge (t_1 - Y_i) \geq u\} I(T_i \geq u) \lambda(u) \, du \\
\left\{ \sum_{j=1}^{n} I\{T_j \wedge C_j \wedge (t_1 - Y_j)^+ \geq u\} \right\} \left\{ \sum_{i=1}^{n} I\{T_i \wedge C_i \wedge (t_2 - Y_i)^+ \geq u\} \right\}^{-1} \\
\int_0^{x \wedge y} \frac{\lambda(u) \, du}{n^{-1} \sum_{i=1}^{n} I\{T_i \wedge C_i \wedge (t_2 - Y_i)^+ \geq u\}},
\end{align*}
$$

which converges in probability to $\sigma^2(x \wedge y; t_2)$ defined by (2.1). Note that the right side of (A.1) is different from the familiar martingale representation for the fixed-sample Nelson–Aalen estimator, which is a stochastic integral with respect to the compensated counting process for the observed number of failures (Fleming and Harrington, 1991, Formula (2.1)). The use of the latter representation would yield stochastic integrals with respect to different martingales for different $t$, making it difficult to evaluate the covariation between $D(x; t_1)$ and $D(y; t_2)$, $t_1 \neq t_2$.

Now, by Lenglart’s inequality (Fleming and Harrington, 1991, Theorem 3.4.1),

$$D(x; t) = \frac{1}{2} \left( \sum_{i=1}^{n} \int_0^{x} I\{C_i \wedge (t - Y_i)^+ \geq u\} \pi(u; t) \, dM_t(u) + o_p(1) \right).$$

Therefore, the classical multivariate central limit theorem implies that the finite-dimensional distribution of the two-parameter process $D(x; t)$, $x < t$, is asymptotically the same as that of a zero-mean Gaussian process $W(x; t)$ with $\text{cov}(W(x; t), W(y; s)) = \sigma^2(x \wedge y; t \wedge s)$. To establish the weak convergence of $D(x; t)$ as a process, we need to prove its tightness. The tightness of a two-parameter process cannot be handled by the martingale theory. But it is not difficult to show that the two-parameter process $D(x; t)$ is tight by using the techniques of Bilić, Gu, and Ying (1996), which were derived from the modern empirical process theory of Pollard (1990).