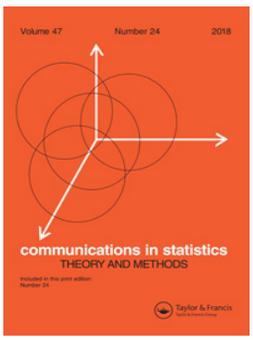

Taylor & Francis
Taylor & Francis Group

Communications in Statistics - Theory and Methods

ISSN: 0361-0926 (Print) 1532-415X (Online) Journal homepage: https://www.tandfonline.com/loi/lsta20

A generalized kaplan-meier estimator for heterogenous populations

David A. Amato

To cite this article: David A. Amato (1988) A generalized kaplan-meier estimator for heterogenous populations, Communications in Statistics - Theory and Methods, 17:1, 263-286, DOI: 10.1080/03610928808829621

To link to this article: https://doi.org/10.1080/03610928808829621

COMMUN. STATIST.-THEORY METH., 17(1), 263-286 (1988)

A GENERALIZED KAPLAN-MEIER ESTIMATOR FOR HETEROGENEOUS POPULATIONS

David A. Amato

Department of Biostatistics
Harvard School of Public Health and
Dana-Farber Cancer Institute
Boston, Massachusetts

Key Words and Phrases: censored data; survival curve estimation; covariate adjustment; strata; clinical trials

ABSTRACT

Current methods for adjusting estimated survival curves for the effects of important prognostic factors are reviewed. These methods often are not useful because of the censoring pattern or because the underlying assumptions are not met. A new method, which is a generalization of the product-limit estimator, is proposed and its properties are considered. It also has limitations, but is useful in several cases where the existing methods are not suitable.

1. INTRODUCTION

In many clinical trials, the comparison of two or more treatment groups is facilitated by a plot of the survival curves. Such estimates usually are generated by the method of Kaplan and Meier (1958), which assumes that the survival times for each treatment are i.i.d. and subject to independent random right-censoring. Frequently, however, the analysis must account for the effects of important covariates. This is particularly true when there is an evident imbalance in the distributions of one or more

covariates among the various treatment arms. The stratified log-rank test (Mantel, 1966) and the proportional hazards model (Cox, 1972) are commonly used to adjust test statistics for such effects. Because the Kaplan-Meier curves do not adjust for prognostic factors, however, the visual assessment of the survival curves may not be consistent with the significance levels generated by these tests. For example, the test may not indicate any significant treatment differences, yet the unadjusted survival curves may suggest important treatment effects. Conversely, the treatment effects may be detected only by an adjusted test, with the unadjusted survival curves showing no apparent treatment differences. Either situation may be confusing to medical investigators and might discourage the use of survival curves in presenting the results.

Several authors have proposed methods for adjusting the estimated survival curves for covariate effects when there are evident imbalances among the treatment arms. Hankey and Myers (1971) presented a method based on the log-rank test (Mantel, 1966). The drawbacks of this method are that it requires large sample sizes and that it uses survival information from other treatments to estimate survival for a particular treatment group. Chang, Gelman, and Pagano (1982) proposed a general approach, the "corrected group prognosis" method (CGP), which averages the estimated survival curves for each patient in the trial. These estimates depend on the patients' covariates and the model chosen. Chang et al. do not recommend any particular method, but present an example based on the Cox (1972) model. Makuch (1982) also illustrates how to compute adjusted survival curves using the Cox model. As noted in Section 2.2, however, these estimators may be seriously misleading if proportional hazards does not hold.

Another type of CGP estimator useful in the case of stratified data is a weighted average of the Kaplan-Meier estimators for each stratum. The principal problem with this technique, as discussed in Section 2.2, is that the time interval on which it is uniquely defined may be too short to usefully depict the survival experience of the treatment groups. The problem arises because the estimate for a given stratum becomes indeterminate when the patient with the longest observation time

in the stratum is censored. Various ad hoc extensions may be employed (e.g., extending the estimated survival curve for a stratum by adding an exponential tail), but are unappealing because they lose the nonparametric flavor of the original procedure and may also be misleading.

Both of these methods rely on assumptions about the forms of the distributions. The stratified estimator often must specify the forms of the tails of the distributions to be useful in practice, while the Cox model employs the proportional hazards assumption. The Kaplan-Meier estimator, on the other hand, does not adjust for covariates when there are serious imbalances. In this paper, we consider a new estimator, proposed independently by Amato (1982) and Campbell and Földes (1984), which adjusts for covariate imbalances without assumptions about the forms of the underlying distributions. It has an intuitively appealing justification, and is a natural generalization of the Kaplan-Meier estimator. It is uniquely defined over a larger time interval than the stratified estimator, which makes it more useful in practice. When the censoring pattern is the same for all strata, it is consistent, nearly unbiased, and has essentially the same asymptotic variance as the stratified estimator. Simulation results show that the estimator is robust against unequal censoring distributions.

2. METHODS

2.1 Formulation

Suppose there are I treatments and the patients are grouped into J strata. It is only necessary to consider the construction of an adjusted survival curve for a single treatment, as the method applies to each treatment separately. Thus, for a given treatment, let N_j be the number of patients in stratum j and let $N = \sum_{j=1}^{J} N_j$. For the k^{th} patient in stratum j, let X_{jk} and Y_{jk} be the survival and censoring times, respectively, $T_{jk} = \min(X_{jk}, Y_{jk})$ and $\delta_{jk} = I\{X_{jk} \leq Y_{jk}\}$. Further, let $S_j(t) = \Pr[X_{jk} > t]$, $\Lambda_j(t) = -\log S_j(t)$, $C_j(t) = \Pr[Y_{jk} > t]$, and $G_j(t)$

$= Pr[X_{jk} > t, Y_{jk} > t]$. Independent right-censoring within each stratum is assumed, so that $G_j(t) = S_j(t)C_j(t)$.

To adjust the estimated survival curve for covariate effects, we choose a (possibly) hypothetical population of patients whose characteristics constitute a reference set. We will use this population to standardize the survival curves for the various treatments. Let P_j be the proportion of patients in stratum j for this standard population. Then, the survival function for the treatment if given to a representative sample from the standard population is

$$S(t) = \sum_{j=1}^{J} P_j S_j(t) \qquad (1).$$

Specifically, S(t) is the probability that an individual drawn at random from the standard population will survive to time t if given the treatment, assuming no information regarding stratum membership is available. By plotting estimates of S for each treatment group, using the same standard population in each case, an adjusted visual comparison of the treatments is obtained. Such a procedure provides a fair comparison of the treatments under study which is not obscured by differences in stratum membership between the therapies. Although the comparison depends on the standard population selected, the particular application will often suggest a natural choice. This issue is discussed further in Section 4.

2.2 Existing Estimators

A straightforward approach to adjusting the survival curves is to replace $S_j(t)$ in (1) by an appropriate estimate. One such method estimates $S_j(t)$ using the Cox model (Makuch, 1982; Chang, Gelman, and Pagano, 1982). One difficulty associated with methods based on the Cox model is the proportional hazards assumption. If, in fact, the hazard functions for the treatment groups cross, it is important for this to be illustrated by the adjusted curves. The proportional hazards assumption, however, restricts the adjusted estimators so that this situation cannot be depicted. In particular, this method results in a distinct ordering of the

estimated survival curves, whereas the true survival curves may cross or have an initial separation which diminishes with time. These phenomena are common in studies of adjuvant chemotherapy following local therapy. The chemotherapy, for example, may delay but not prevent disease recurrence, resulting in an initial benefit from chemotherapy, but no long-term difference. Methods based on the Cox model, however, would depict a clear advantage for the chemotherapy arm. Therefore, if proportional hazards does not hold, the estimators are inconsistent and may be very misleading.

Makuch (1982) correctly points out that methods based on the Cox model should not be used unless the proportional hazards assumption has been verified. The methods of Schoenfeld (1980) or Lagakos (1981) may be used to test for proportional hazards, but may lack power in small data sets. Therefore, if proportional hazards does not hold, the Cox model method can present a seriously misleading picture, yet the data may be consistent with this assumption.

A second approach, which we refer to as the "stratified" Kaplan-Meier estimator (SKM), replaces $S_j(t)$ in (1) by its Kaplan-Meier (KM) estimate, $\tilde{S}_j(t)$. Although the SKM has many appealing properties, it can suffer from severe definitional problems, as it is uniquely defined only at times when each of these J estimators are uniquely defined. Thus, if there are many strata, each with a small number of observations, or if there is heavy censoring, $\tilde{S}$ may not be very useful for describing the survival experience of the treatment group. Equally important, failures which occur beyond the interval where $\tilde{S}$ is unique are not depicted, yielding a potentially misleading plot. For example, late failures may help determine whether or not an early observed treatment difference is indicative of a real cure or only a delay in the time of failure.

One ad hoc solution adds an exponential tail to the estimate for a stratum when its Kaplan-Meier estimate becomes indeterminate. As noted previously, this is intuitively unappealing and may also be seriously misleading. In many studies, particularly those involving adjuvant chemotherapy, the

failure rate decreases over time. The ad hoc procedure would then underestimate the true survival. If there is a cure rate (e.g. disease free at 5 years implies the patient is cured), this method gives the misleading impression that patients continue to fail beyond this time.

Thus, both the estimator based on the Cox model and the SKM may have serious drawbacks. In the next section, a new estimator is proposed which has a larger range of unique definition, but which does not rely on the proportional hazards assumption.

2.3 The "Weighted" Kaplan-Meier Estimator

Let $a_{j,N}$ be a known, positive weight assigned to each individual in stratum j, for a given treatment, such that $\sum_{j=1}^{J} a_{j,N} N_j = N$. Here, $a_{1,N}, \ldots, a_{J,N}$ may depend on $N_1, \ldots, N_J$, as well as on N. Define $N_j^D(t) = \sum_{k=1}^{N_j} I\{T_{jk} \leq t, \delta_{jk}=1\}$, $W_j^D(t) = a_{j,N} N_j^D(t)$, and $W^D(t) = \sum_{j=1}^{J} W_j^D(t)$. Thus, $W^D(t)$ is the sum of the weights of all individuals observed to fail by time t. Similarly, define

$$N_j^C(t) = \sum_{k=1}^{N_j} I\{T_{jk} \leq t, \delta_{jk}= 0\}, \quad W_j^C(t) = a_{j,N} N_j^C(t), \quad W^C(t) = \sum_{j=1}^{J} W_j^C(t),$$

$$N_j^R(t) = \sum_{k=1}^{N_j} I\{T_{jk} \geq t\}, \quad W_j^R(t) = a_{j,N} N_j^R(t), \text{ and } W^R(t) = \sum_{j=1}^{J} W_j^R(t), \text{ so}$$

that $W^C(t)$ and $W^R(t)$ are the sums of the weights of all individuals censored by time t and at risk at time t, respectively. We then define a class of "weighted" Kaplan-Meier estimators (WKM) as the set of all survival functions $\hat{S}$ such that

$$\hat{S}(t) = \prod_{u \leq t} \left(1 - \frac{\Delta W^D(u)}{W^R(u)}\right) \tag{2}$$

for $t \leq T = \sup\{u: N^R(u) > 0\}$. Note that in the case of unit weights, the WKM is simply the Kaplan-Meier estimator. Unless stated otherwise, the weights $a_{j,N} = NP_j/N_j$ will be assumed.

The intuitive justification of the WKM is that we wish to estimate the survival function for the standard population (1), but are sampling from a different population. To compensate, weights are assigned so that the total weight for a given stratum is equal to the expected number of observations which would have fallen in that stratum had N individuals been drawn at random from the standard population. This is equivalent to replicating each observation $c \cdot a_{j,N}$ times, for some suitably large c, so that the proportion of observations in stratum j is P_j, and then computing the KM estimator.

We conclude this section by noting that the WKM is identical to the SKM in the case of no censoring. Then, $W^R(u) = N\tilde{S}(u-)$ and $W^R(u) - \Delta W^D(u) = N\tilde{S}(u)$ so that

$$\hat{S}(t) = \prod_{u \le t} \left(1 - \frac{\Delta W^D(u)}{W^R(u)}\right) = \prod_{u \le t} \frac{\tilde{S}(u)}{\tilde{S}(u-)} = \tilde{S}(t).$$

2.4 Properties of the WKM

For simplicity, the properties of the WKM are derived only for the case of a discrete time scale. The corresponding results for the continuous case are indicated, but are not proven here, as they are similar. Thus, let $0 = t_0 < t_1 < \ldots$ be the time points at which events may occur. By convention, it is assumed that the censored observations occur just after each time point. Accordingly, we define two σ-algebras:

$$F_k = \sigma\{N_j^D(t_m),\ N_j^R(t_n);\ 0 \le m \le k,\ 0 \le n \le k,\ 1 \le j \le J\}$$

and $$F_k^- = \sigma\{N_j^D(t_m),\ N_j^R(t_n);\ 0 \le m \le k-1,\ 0 \le n \le k,\ 1 \le j \le J\}.$$

We begin by examining conditions for the consistency of $\hat{S}$:

Theorem 1: If $a_{j,N} \to \bar{a}_j$ and $N_j/N \to Q_j$ in probability, and $C_j(t_{k-1})S_j(t_{k-1}) > 0$ for all j, then

$$\hat{S}(t_K) \to \prod_{i=1}^{k} \frac{\sum_{j=1}^{J} \bar{a}_j Q_j C_j(t_{i-1}) S_j(t_i)}{\sum_{j=1}^{J} \bar{a}_j Q_j C_j(t_{i-1}) S(t_{i-1})} \quad \text{in probability, as } N \to +\infty.$$

Proof: $1 - \Delta W^D(t_i)/W^R(t_i)$

$$= 1 - \{\sum_{j=1}^{J} a_{j,N}(N_j/N)[\Delta N_j^D(t_i)/N_j]\} / \{\sum_{j=1}^{J} a_{j,N}(N_j/N)[N_j^R(t_i)/N_j]\}$$

$$-P\to 1 + \{\sum_{j=1}^{J} \bar{a}_j Q_j C_j(t_{i-1}) dS_j(t_i)\} / \{\sum_{j=1}^{J} \bar{a}_j Q_j C_j(t_{i-1}) S_j(t_{i-1})\}$$

$$= \{\sum_{j=1}^{J} \bar{a}_j Q_j C_j(t_{i-1}) S_j(t_i)\} / \{\sum_{j=1}^{J} \bar{a}_j Q_j C_j(t_{i-1}) S_j(t_{i-1})\}.$$

The theorem follows from (2).

In the continuous case, it can be shown (Amato, 1982) that

$$\hat{S}(t) \to \exp\left[\int_0^t \frac{\sum_{j=1}^{J} \bar{a}_j Q_j C_j(u) dS_j(u)}{\sum_{j=1}^{J} \bar{a}_j Q_j C_j(u) S_j(u)}\right], \text{ under the same}$$

conditions. The proof is similar to Breslow and Crowley (1974).

Corollary: Sufficient conditions for $\hat{S}$ to be consistent, in both the discrete and continuous cases, are (1) $C_1 = C_2 = \ldots = C_J$ and $\bar{a}_j = P_j/Q_j$ or (2) $S_1 = S_2 = \ldots = S_J$.

The condition that $C_1 = C_2 = \ldots = C_J$ is referred to as equal censoring throughout the paper. In this case, the subscript j is omitted. Note that the weights $a_{j,N} = NP_j/N_j$ guarantee the consistency of $\hat{S}$, provided equal censoring holds.

Campbell and Földes (1984) established uniform convergence of the estimator in the continuous case under equal censoring, but did not consider the case of unequal censoring. To see the importance of unequal censoring, recall that the KM estimator is the WKM with unit weights. In many applications, particularly randomized clinical trials, the sampled population is the one of interest, so no adjustment is required and $P_j = Q_j$ for all j. This is the case in which the KM estimator may be appropriate and is used routinely. Even in this case, however, if there is unequal censoring and the population is truly heterogeneous, the

KM is inconsistent, as noted recently by Slud and Rubinstein (1984). Therefore, the KM should be used with caution when no adjustment is required and the population is heterogeneous. The WKM is also inconsistent when there is unequal censoring. Under equal censoring, however, it does adjust for imbalances among the treatments in the distributions of covariates, whereas the KM does not. In addition, the WKM has a longer interval of unique definition then the SKM, making it potentially more useful in practice.

This result can be understood in two ways. First, the proportion of the risk set at time t that is in stratum j depends on the censoring distributions, unless they are equal. Hence, the overall estimate of the hazard at time t depends on the censoring unless the survival distributions are identical or equal censoring holds. Second, the KM and WKM ignore the strata once weights are assigned and treat the censoring time as being noninformative concerning survival time. This is the case under equal censoring. With unequal censoring, however, the censoring time contains information regarding stratum membership. This in turn contains information regarding survival time, provided the strata do not have the same survival distributions. Thus, ignoring the strata induces a dependence between the survival and censoring times, making the censoring informative.

Fortunately, this is usually not a problem in the setting of clinical trials. Censoring is mostly administrative (i.e. due to staggered entry) with relatively few patients lost to follow-up. It is much less likely that administrative censoring is due to stratum membership than for losses to follow-up. Thus, although the entry rate of patients over time may vary, it is only required that the proportions of patients entered to the different strata do not change substantially with time. When there is a time trend in the prognoses of patients, outcome is related to censoring time and both the KM and WKM are inconsistent. Simulation results, however, suggest that even moderate departures from the equal censoring assumption do not seriously affect the performance of these estimators (see Section 3).

Next, we examine the expectation and variance of $\hat{S}$ in the discrete case under the condition of equal censoring. Let α_k

$= S(t_k)/S(t_{k-1})$, $\beta = C(t_k)/C(t_{k-1})$, $\alpha_k^T = 1 - (1-\alpha_k)I\{t_k \leq T\}$, and $\hat{\alpha}_k^T = 1 - [\Delta W^D(t_k)/W^R(t_k)]I\{t_k \leq T\}$. Define $S^T(t_k) = \prod_{i=1}^{k} \alpha_i^T$, $\hat{S}^T(t_k) = \prod_{i=1}^{k} \hat{\alpha}_i^T$, and $R^T(t_k) = \hat{S}^T(t_k)/S^T(t_k)$. We define similar quantities for the strata by adding the subscript j. The next two theorems use the convention $0/0 = 1$ for convenience.

Theorem 2: If equal censoring holds, $a_{j,N} = NP_j/N_j$, and $C(t_{k-1})S_j(t_{k-1}) > 0$ for all j, then $E[R^T(t_k)] = 1 + O(N^{-1})$.

Proof: $E[R^T(t_k)] = E\{R^T(t_{k-1})E[\hat{\alpha}^T_k/\alpha^T_k|F_k^-]\}$. It is easily shown that $E[\hat{\alpha}_k^T/\alpha_k^T|F_k^-] = \sum_{j=1}^{J} W_j^R(t_k)\alpha_{kj}/W^R(t_k)\alpha_k + o(N^{-1})$. Hence,

$$E[R^T(t_k)] = E\{R^T(t_{k-1})\alpha_k^{-1}E[\sum_{j=1}^{J} W_j^R(t_k)\alpha_{kj}/W^R(t_k)|F_{k-1}]\} + o(N^{-1}).$$

$$E[\sum_{j=1}^{J} W_j^R(t_k)\alpha_{kj}/W^R(t_k)|F_{k-1}]$$

$$= E\left\{ \frac{\sum_{j=1}^{J} [W_j^R(t_{k-1}) - \Delta W_j^D(t_{k-1}) - \Delta W_j^C(t_{k-1})]\alpha_{kj}}{W^R(t_{k-1}) - \Delta W^D(t_{k-1}) - \Delta W^C(t_{k-1})} \mid F_{k-1}\right\}$$

Expanding in a Taylor's series about the conditional expectations of the numerator and denominator with respect to F_{k-1}, and applying the condition of equal censoring, this last term becomes $\sum_{j=1}^{J} [W_j^R(t_{k-1}) - \Delta W_j^D(t_{k-1})]\alpha_{kj}/[W^R(t_{k-1}) - \Delta W^D(t_{k-1})] + O(N^{-1})$. Thus,

$$E[R^T(t_k)] = E\left\{ R^T(t_{k-2}) \frac{\hat{\alpha}^T_{k-1} \sum_{j=1}^{J} [W_j^R(t_{k-1}) - \Delta W_j^D(t_{k-1})]\alpha_{kj}}{\alpha^T_{k-1}\alpha_k[W^R(t_{k-1}) - \Delta W^D(t_{k-1})]} \right\} + O(N^{-1})$$

$$= E\left\{R^T(t_{k-2})(\alpha^T_{k-1}\alpha_k)^{-1} \frac{\sum_{j=1}^{J} [W_j^R(t_{k-1}) - \Delta W_j^D(t_{k-1})]\alpha_{kj}}{W^R(t_{k-1})}\right\} + O(N^{-1}).$$

Repeating this conditioning argument, we eventually obtain

$$E[R^T(t_k)] = \prod_{i=1}^{k} \alpha_i^{-1} \cdot N^{-1} \sum_{j=1}^{J} a_{j,N} N_j \prod_{i=1}^{k} \alpha_{ij} + O(N^{-1})$$

$$= \sum_{j=1}^{J} P_j S_j(t_k)/S(t_k) + O(N^{-1}) = 1 + O(N^{-1}).$$

The exact variance of R^T is difficult to express. Instead, we derive an approximate variance expression. Proceeding as in Theorem 2, we have $E\{[R^T(t_k)]^2\} = E\{[R^T(t_{k-1})]^2 E[(\hat{\alpha}_k^T/\alpha_k^T)^2|F_k^-]\}$. It can be shown that $E[(\hat{\alpha}_k^T/\alpha_k^T)|F_k^-]$

$$= [\alpha_k W^R(t_k)]^{-2}\{\sum_{j=1}^{J} a_{j,N} W_j^R(t_k)\alpha_{kj}(1-\alpha_{kj}) + [\sum_{j=1}^{J} W_j^R(t_k)\alpha_{kj}]^2\} + O(N^{-2}).$$

Using Taylor series expansions, we obtain

$$E\{\sum_{j=1}^{J} a_{j,N} W_j^R(t_k)\alpha_{kj}(1-\alpha_{kj})/[W^R(t_k)]^2|F_{k-1}\} \tag{2}$$

$$= \beta_{k-1}^{-1}[W^R(t_{k-1}) - \Delta W^D(t_{k-1})]^{-2} \sum_{j=1}^{J} a_{j,N}[W_j^R(t_{k-1}) - \Delta W_j^D(t_{k-1})]\alpha_{kj}(1-\alpha_{kj})$$

$+ O(N^{-2})$ and $E\{[\sum_{j=1}^{J} W_j^R(t_k)\alpha_{kj}/W^R(t_k)]^2|F_{k-1}\}$ (3)

$$= \{\sum_{j=1}^{J} [W_j^R(t_{k-1}) - \Delta W_j^D(t_{k-1})]\alpha_{kj} / [W^R(t_{k-1}) - \Delta W^D(t_{k-1})]\}^2 + O(N^{-1}).$$

The terms of order N^{-1} in expression (3) will, in practice, be small compared with expression (2). Hence, they are ignored in deriving the approximate variance. Thus, we obtain $E\{[R^T(t_k)]^2\}$

$$\doteq E\{[R^T(t_{k-1})]^2 \alpha_k^{-2}[W^R(t_{k-1}) - \Delta W^D(t_{k-1})]^{-2}$$

$$\cdot [\beta_{k-1}^{-1} \sum_{j=1}^{J} a_{j,N}[W_j^R(t_{k-1}) - \Delta W_j^D(t_{k-1})\alpha_{kj}(1-\alpha_{kj})$$

$$+ (\sum_{j=1}^{J} [W_j^R(t_{k-1}) - \Delta W_j^D(t_{k-1})]\alpha_{kj})^2]\} = E\{[R^T(t_{k-2})]^2[\alpha_{k-1}^T \alpha_k W^R(t_{k-1})]^{-2}$$

$$\cdot [\beta_{k-1}^{-1} \sum_{j=1}^{J} a_{j,N}[W_j^R(t_{k-1}) - \Delta W_j^D(t_{k-1})]\ \alpha_{kj}(1-\alpha_{kj})$$

$+ \left(\sum_{j=1}^{J} [W_j^R(t_{k-1}) - \Delta W_j^D(t_{k-1})]\alpha_{kj} \right)^2 \}$. Repeating this conditioning argument, we obtain

$$E\{[R^T(t_k)]^2\} \doteq N^{-2} \prod_{i=1}^{k} \alpha_i^{-2} [(\sum_{j=1}^{J} a_{j,N} N_j \prod_{i=1}^{k} \alpha_{ij})^2$$

$$+ \sum_{i=1}^{k} \prod_{m=1}^{i-1} \beta_m^{-1} \sum_{j=1}^{J} a_{j,N}^2 N_j (\prod_{m=1}^{i} \alpha_{mj})(1-\alpha_{ij})(\prod_{m=i+1}^{k} \alpha_{mj}^2)]$$

$$= 1 + S^{-2}(t_k) \sum_{j=1}^{J} (P_j^2/N_j) S_j^2(t_k) \sum_{i=1}^{k} [C(t_{i-1})S_j(t_i)]^{-1} d\Lambda_j(t_i).$$

This suggests the approximation

$$\mathrm{Var}[R^T(t)] \doteq S^{-2}(t) \sum_{j=1}^{J} P_j^2 V_j(t) \qquad (4)$$

where $V_j(t) = N_j^{-1} S_j^2(t) \int_0^t [C(u-)S_j(u)]^{-1} d\Lambda_j(u)$. In practice, we estimate $\mathrm{Var}[\hat{S}(t)]$ by $\sum_{j=1}^{J} P_j^2 \hat{V}_j(t)$, where

$$\hat{V}_j(t) = \tilde{S}_j^2(t) \int_0^t \frac{dN_j^D(u)}{N_j^R(u)[N_j^R(u) - \Delta N_j^D(u)]} \qquad (5).$$

It can be shown (Amato, 1982) that the approximate formula (4) holds in the continuous case as well. The right-hand side of (4) is equal to the variance of $\tilde{S}(t)/S(t)$ under equal censoring, suggesting that $\hat{S}$ and $\tilde{S}$ have approximately the same variance in this case. Accordingly, the expression for $\hat{V}_j(t)$ is simply Greenwood's (1926) formula for the estimated variance of $\tilde{S}_j(t)$.

In deriving (4), some terms of order N^{-1} were ignored. These terms were very cumbersome to write out explicitly and, in applications, would tend to be small compared with other terms of order N^{-1}. Thus, formula (4) should be viewed with some caution. Nevertheless, simulation results, which will be presented in section 3, support this approximation.

We conclude this section with a discussion of the asymptotic distribution of $\hat{S}$. It is easily shown that $\hat{S}(t_k) = 1 + \sum_{i=1}^{k} \hat{S}(t_{i-1})(\hat{\alpha}_i - 1)$,

with $\hat{\alpha}_i = 1 - \Delta W^D(t_i)/W^R(t_i)$. Given F_i^-, $\hat{\alpha}_i$ is a linear combination of independent binomial random variables. Since $N_j^R(t_i)/N_j \to S_j(t_{i-1})C_j(t_i)$ and, $\hat{S}(t_i) \to S(t_i)$ in probability, and applying Slutsky's Theorem, it follows that $\hat{S}(t_i-1)(\hat{\alpha}_i - 1)$ is asymptotically normal, and hence, $\hat{S}(t_K)$ is asymptotically normal.

This suggests that approximate confidence intervals for $\hat{S}$ may be computed using equation (5) and, by Theorem 1, assuming $\hat{S}$ is unbiased. Note, however, that $\hat{V}_j(t)$ is not defined unless $t \leq T_j^*$, where T_j^* is the time at which $\tilde{S}_j$ becomes indeterminate. If $t > T_j^*$, the conservative approximation

$$V_j^*(t) = \tilde{S}_j^2(T_j^*) \int_0^{T_j^*} \frac{dN_j^*(u)}{N_j^R(u)[N_j^R(u) - \Delta N_j^*(u)]}$$

may be used, where $N_j^*(u)$ is $N_j^D(u)$ if $u < T_j^*$ and is $N_j^D(T_j^*-) + N_j^R(T_j^*)$ if $u = T_j^*$.

3. SIMULATION RESULTS

The small-sample properties of the WKM and SKM were compared using simulations. The model assumes patients are entered to a clinical trial uniformly for 3 years and are observed for 2 additional years. Hence, the censoring distribution is uniform on the interval [2,5]. The model for survival is $S_j(t) = \exp\{-\lambda_j t^\alpha\}$, with $\lambda_j = \lambda\exp\{\beta_j\}$, with $\lambda = -\log(0.2)/5^\alpha$. Thus, a patient with $\lambda_j=\lambda$ has a 5-year survival probability of 20%. Covariate effects are equally spaced between -1 and 1; i.e. $\beta_j=(2j-J-1)/J$ for $j=1,\ldots,J$. This choice allows some growth in the size of the effects as the number of strata increases.

For each N ε {20,40,80}, J ε {2,4,8}, and α ε {0.5,1.0,2.0}, 5000 runs were made. In each case, two situations were considered:

$$Q_j = \begin{cases} 1/2J & j \leq J/2 \\ 3/2J & j \geq J/2+1 \end{cases}$$ (more poor risk patients in the sample)

and $$Q_j = \begin{cases} 3/2J & j \leq J/2 \\ 1/2J & j \geq J/2+1 \end{cases}$$ (more good risk patients in the sample)

In each case, we took the standard population to have the other distribution; i.e. $P_j = 2/J - Q_j$.

The measurements of primary interest were the bias and root mean square error (RMSE), particularly at t = 3.5, the midpoint of the interval in which censoring could occur. Unfortunately, the estimators were not necessarily uniquely defined at this time point on each run. Two methods were employed to account for this problem. First, the results of a run were used for an estimator only if it was uniquely defined at t = 3.5. Since the WKM is uniquely defined whenever the SKM is, but not vice-versa, the data for some runs were used only for the WKM. The number of runs for which each estimator was uniquely defined were recorded. Second, the estimators were extended, if necessary, by adding (weighted) exponential tails, so that both were defined at t = 3.5. Only the results for the first case, with J = 2, are reported here, as the other results were similar.

Tables 1-3 display the results. The biases were very small: no more than 0.0100 for the SKM and no more than 0.0055 for the WKM, in all cases. The RMSE's varied according to sample size, naturally, but were very similar for the two estimators, supporting the approximate variance formula (4). As indicated in Table 3, however, the probabilities of unique definition were substantially different, showing the greater utility of the WKM.

The results for the WKM and SKM shown in Tables 1 and 2 are not directly comparable, because, as noted above, the WKM was uniquely defined on some runs for which the SKM was not. However, as the conditions under which this arises are unfavorable (e.g. heavy censoring), the comparisons are biased in favor of the SKM.

Table 1 - Bias at t = 3.5

Sample Size	α	More Poor Risk Patients SKM	More Poor Risk Patients WKM	More Good Risk Patients SKM	More Good Risk Patients WKM
20	0.5	0.0100	0.0033	-0.0044	0.0043
	1.0	0.0071	0.0012	-0.0041	0.0036
	2.0	0.0080	0.0024	-0.0018	0.0055
40	0.5	0.0026	-0.0016	-0.0006	0.0006
	1.0	0.0030	-0.0017	0.0008	0.0011
	2.0	0.0029	0.0003	0.0027	0.0014
80	0.5	0.0008	0.0000	0.0024	0.0003
	1.0	0.0011	0.0003	0.0019	0.0006
	2.0	0.0019	0.0007	0.0014	0.0017

Table 2 - Root Mean Square Error at t = 3.5

Sample Size	α	More Poor Risk Patients SKM	More Poor Risk Patients WKM	More Good Risk Patients SKM	More Good Risk Patients WKM
20	0.5	0.1306	0.1306	0.0720	0.0763
	1.0	0.1314	0.1325	0.0857	0.0882
	2.0	0.1304	0.1335	0.1080	0.1127
40	0.5	0.0875	0.0877	0.0521	0.0510
	1.0	0.0916	0.0911	0.0618	0.0612
	2.0	0.0907	0.0901	0.0782	0.0776
80	0.5	0.0612	0.0603	0.0367	0.0355
	1.0	0.0620	0.0618	0.0434	0.0430
	2.0	0.0632	0.0630	0.0529	0.0540

Table 3 - Estimated Probability of Unique Definition at t = 3.5

Sample Size	α	More Poor Risk Patients SKM	More Poor Risk Patients WKM	More Good Risk Patients SKM	More Good Risk Patients WKM
20	0.5	0.550	0.976	0.539	0.980
	1.0	0.547	0.983	0.538	0.986
	2.0	0.524	0.996	0.515	0.996
40	0.5	0.672	1.000	0.664	0.998
	1.0	0.680	1.000	0.675	1.000
	2.0	0.700	1.000	0.703	1.000
80	0.5	0.776	1.000	0.778	1.000
	1.0	0.778	1.000	0.767	1.000
	2.0	0.813	1.000	0.804	1.000

These results suggest that the WKM performs as well as the SKM under equal censoring, but that the WKM is considerably more useful in practice. The disparity in the probabilities of unique definition are even greater if there are more than two strata.

We also employed different censoring distributions for each stratum. The bias and RMSE of the WKM were similar to the SKM for even moderate deviations from the equal censoring assumption. The differences in the RMSE were less than 2% in all cases. As the comparison is biased in favor of the SKM, the SKM does not appear preferable to the WKM even with fairly unequal censoring.

4. INTERACTIONS AND THE CHOICE OF THE STANDARD POPULATION

A common objection to adjusting survival curves for prognostic factors is that if there are qualitative treatment-strata interactions (as defined by Peto, 1982), then an overall survival curve is an inappropriate summary of the data because these interactions are hidden. We agree in principle, but point out that the existence of such interactions is almost always in doubt, even after the data have been analyzed. As stressed by Peto, quantitative interactions are almostAs stressed by Peto, quantitative interactions are almost certain to exist, but qualitative interactions are a priori not very plausible. Even if no true interactions exist, apparent qualitative interactions are likely to occur, particularly if there are many strata. Peto argues that unless there are good prior reasons for expecting qualitative interactions, the appearance of such interactions should be reported but not believed. He notes that significant tests for interactions do not in themselves constitute evidence for qualitative interactions; there must be strong scientific reasons to support these findings. Peto argues convincingly that the inference about treatment effects should be guided primarily by the overall results, rather than by subgroup analyses, even if there are apparent qualitative interactions. Thus, the possibility of qualitative interactions should rarely be a deterrent to the use of an overall survival curve to summarize the findings.

If there are no qualitative treatment-strata interactions, the choice of $P_1,\ldots,P_J$ should have little effect on the relative

differences between the adjusted survival curves for the various treatment groups, but may affect the magnitudes of the estimates. If qualitative interactions are present, however, the relative treatment differences also can be affected by the choice of the P's. Thus, some care should be taken in selecting the standard population. For most applications, however, there is a natural choice for the P's. In a randomized clinical trial, for example, the eligible patients are the population of interest. Thus, it is natural to let P_j be the proportion of all entered patients who are in stratum j. In a historically controlled study, P_j would be the proportion of the controls (or cases) who are in stratum j. By using the "natural" standard population for the given application, the appearance of arbitrariness in the selection of the P's is removed.

5. EXAMPLE

The example involves the same data set used by Chang, Gelman, and Pagano (1982). Forty-two stage II ovarian patients on an Eastern Cooperative Oncology Group (ECOG) protocol were randomized between two chemotherapy regimens, PAM and CMF. Chang et al. noted that although the unadjusted survival curves did not show a treatment difference, once the survival curves were corrected for the most significant prognostic factors using the Cox model, the treatment groups had a marked difference in survival.

For simplicity, the patients were grouped into three strata based on the most important covariates: age and time from diagnosis to randomization. The (unadjusted) Kaplan-Meier estimates of the two survival distributions are indicated in Figure 1. Note that the treatments do not appear different with regard to survival. Figure 2 shows the adjusted estimates based on the Cox model. In agreement with Chang et al., the adjusted survival curves indicate that CMF is associated with longer survival. Figure 3 depicts the SKM estimates. In contrast to the Cox model estimates, the SKM reveals no treatment differences. As shown in Figure 4, the WKM agrees with the SKM, but is uniquely defined over a larger time interval.

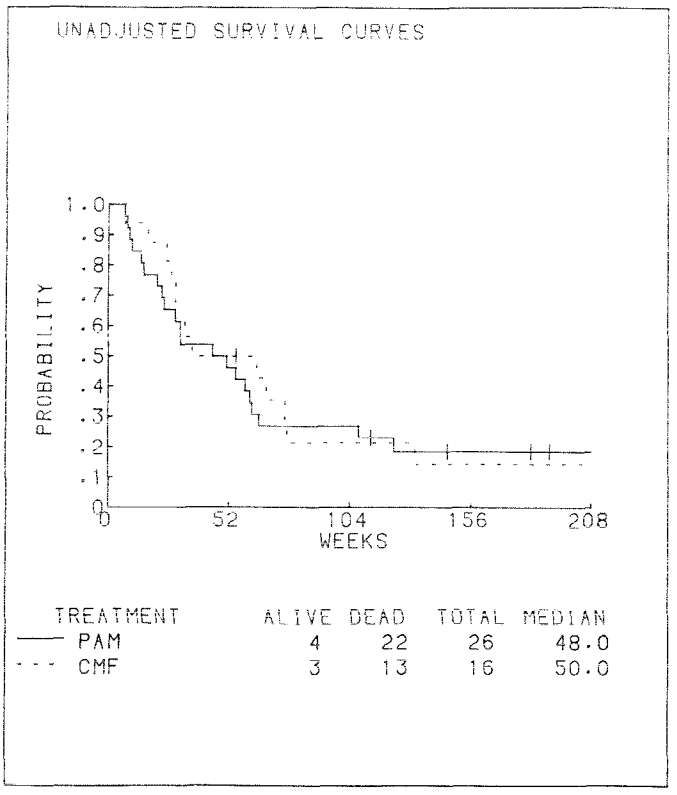

FIGURE 1.

Because the SKM makes no assumptions other than that censoring within strata is noninformative, it is consistent and nearly unbiased whatever the underlying survival distributions. The Cox model, on the other hand, employs proportional hazards which, if incorrect, may present a misleading picture. Given the opposing impressions generated by the two estimators, the graphical method of Lagakos (1981) was employed to check the proportional hazards assumption. A comparison of the adjusted ranks by treatment and by stratum did not provide evidence of lack of fit to the proportional hazards model. A comparison of the treatments within stratum 2, which contained 64% of the patients, suggested the possibility of crossing hazards, but due to the small sample sizes, was not conclusive. Thus, there is no strong

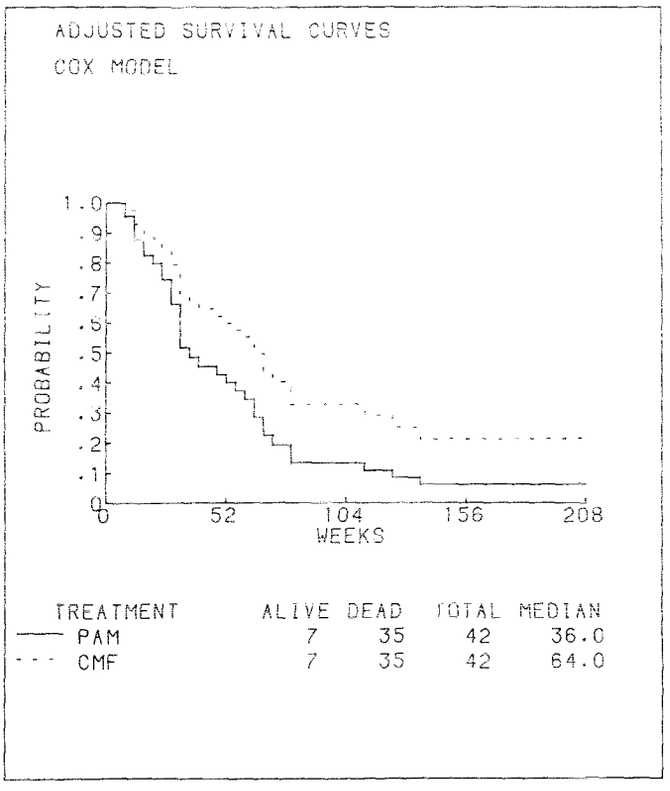

FIGURE 2.

evidence that the proportional hazards model is incorrect, yet the stratum 2 results are cause for concern.

Figures 3 and 4 more accurately summarize the data than does Figure 2 because they show the possibility of crossing hazards and the resulting uncertainty as to the treatment effect. Figure 2, however, suggests a strong benefit for CMF. It appears that by forcing the data to fit the proportional hazards framework the method based on the Cox model exaggerates the treatment effect. We emphasize that we do not contend that the proportional hazards assumption is wrong or that the Cox model is inappropriate for testing the treatment effect. Our concern is only that the Cox

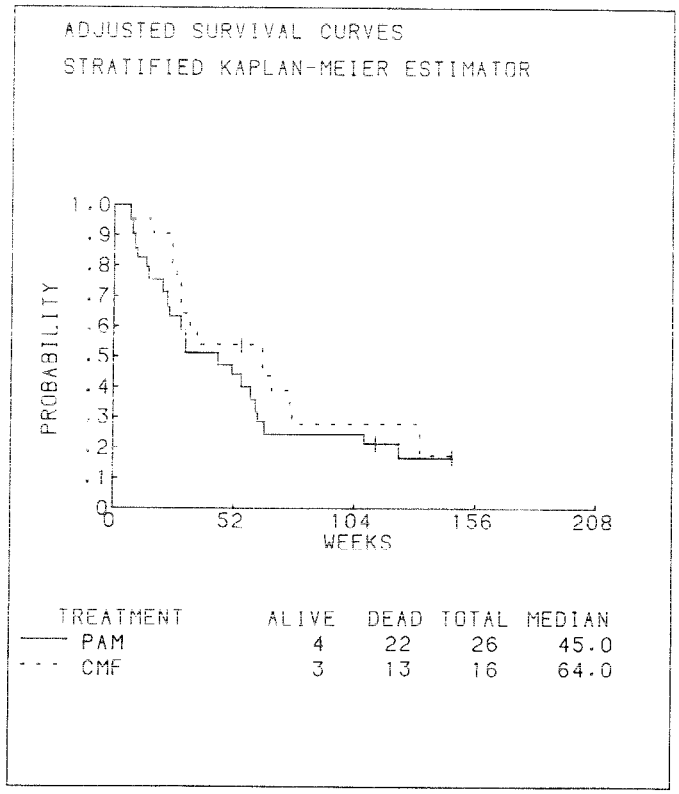

FIGURE 3.

model method may suggest greater treatment effects than are supported by the data and therefore may mislead medical investigators.

6. DISCUSSION

The main application of these methods will be in situations where the treatment groups are drawn from different populations and need to be standardized before comparisons are made. They may be particularly useful in nonrandomized or historically controlled studies, where the distribution of covariates are likely to differ. These methods, however, also are appropriate for analyzing randomized clinical trials, as in our example. They

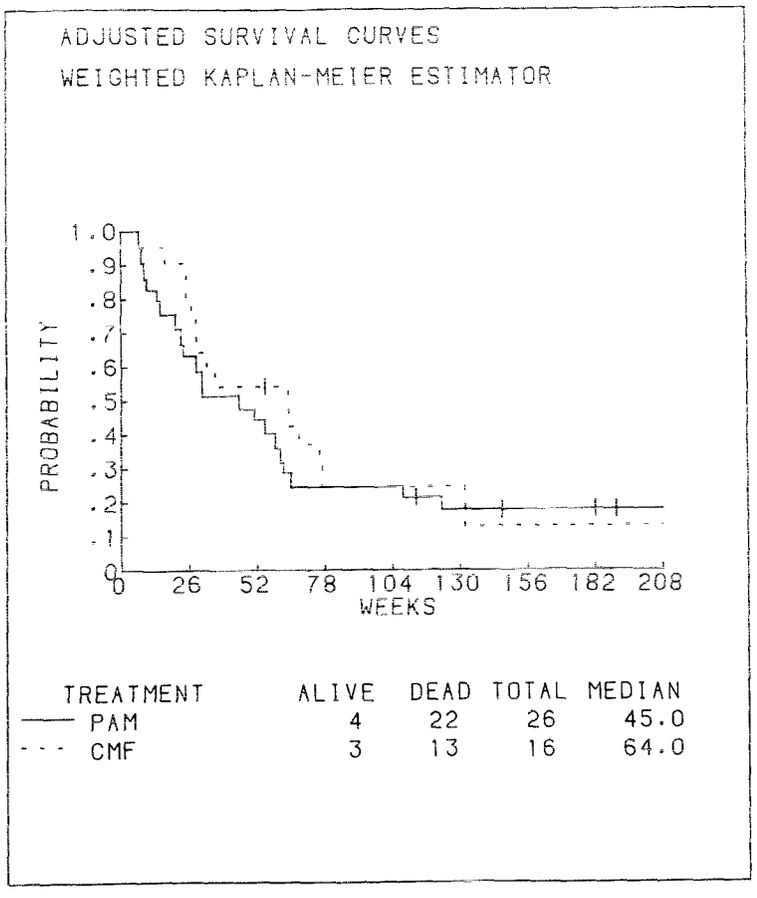

FIGURE 4.

allow one to depict the true effects of treatments or covariates which are obscured by covariate imbalances. In addition, a more sensitive comparison of the treatment groups is possible by adjusting for imbalances in important covariates.

The WKM is an appropriate method of adjustment for covariate effects in a wide variety of applications. It depicts all of the data and therefore is a useful estimator in many cases in which the SKM is not. It avoids semi-parametric assumptions, such as proportional hazards or exponential tails, which can be seriously misleading. Its bias and RMSE appear equivalent to the SKM in small and moderate samples, even when there is unequal censoring.

The equal censoring assumption is the only potential drawback to the WKM, but will not be a concern in most clinical trials applications. First, this assumption is testable (for a given treatment, simply reverse the censoring indicator and compare the strata using, say, the log-rank test). Second, censoring in most clinical trials is largely administrative, which conforms to the equal censoring assumption. In addition, recall that the Kaplan-Meier is a special case of the WKM. Hence, the WKM is appropriate in any application for which the KM is suitable, and conversely, any concerns over the use of the WKM apply equally to the KM. As the KM is used routinely to summarize the results of cancer clinical trials, in which the patient populations are very heterogeneous, the instances in which the WKM is not appropriate should be relatively infrequent.

There are situations, however, in which the WKM (and the KM) should be used with caution. Studies in which there are changes in the prognoses of entered patients with time are one example. Combining evidence from several studies with different lengths of follow-up is another. In each of these instances, the censoring distributions are known to differ among the strata. The simulation results suggest that the WKM is still very good if the degree of inequality is small or moderate. Note also that the SKM is most likely to have a short range of unique definition, and therefore not be very useful, when the censoring is unequal, unless the sample sizes are very large.

The method of adjustment based on the Cox model can be seriously misleading, since it forces the estimators for the various treatment groups to fit the proportional hazards framework. The case of crossing hazards, such as when treatment delays, but does not prevent, disease recurrence cannot be depicted by this method. Thus, it should be used with considerable caution, and only after testing or graphically examining the proportional hazards assumption.

The KM estimator is not a competitor to the WKM when adjustment for covariates is required, as it does not provide any adjustment. In randomized clinical trials, however, large imbalances tend to be infrequent, so it is not clear that such

techniques are necessary. An interesting question, which we have not investigated, is whether the WKM is more precise than the KM in randomized clinical trials.

ACKNOWLEDGEMENTS

The author would like to thank David Harrington and David Schoenfeld for their helpful comments, the Eastern Cooperative Group for the use of their data, and Jean Ryan for assistance in preparing the manuscript. This work was supported in part by a grant from the National Institutes of Health (CA 23415).

BIBLIOGRAPHY

Amato, D.A. (1982). The statistical design and analysis of animal cancer treatment studies. Unpublished Ph.D. Dissertation. Cornell University, Ithaca, New York.

Breslow, N. and Crowley, J. (1974). A large sample study of the life table and product limit estimates under random censorship. Annals of Statistics 3, 437-453.

Campbell, G. and Földes, A. (1984). A generalized product-limit estimator for weighted distribution functions based on censored data. Statistics & Decisions Supplement Issue 1, 87-109.

Chang, I.M., Gelman, R., and Pagano, M. (1982). Corrected group prognostic curves and summary statistics. Journal of Chronic Diseases 35, 669-674.

Cox, D.R. (1972). Regression models and life-tables (with discussion). Journal of the Royal Statistical Society, Series B 34, 187-220.

Greenwood, M. (1926). The natural duration of cancer. Reports on Public Health and Medical Subjects 33. His Majesty's Stationary Office.

Hankey, B.F. and Myers, M.H. (1971). Evaluating differences in survival between two groups of patients. Journal of Chronic Diseases 24, 523-531.

Kaplan, E.L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. Journal of the American Statistical Association 53, 457-481.

Lagakos, S.W. (1981). The graphical evaluation of explanatory variables in proportional hazards regression models. Biometrika 68, 93-98.

Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. Cancer Chemotherapy Reports 50, 163-170.

Makuch, R.W. (1982). Adjusted survival curve estimation using covariates. Journal of Chronic Diseases 35, 437-443.

Peto, R. (1982). Statistical aspects of cancer trials. In Treatment of Cancer, K.E Halnan ed., 867-871. Chapman and Hall, London.

Schoenfeld, D. (1980). Chi-squared goodness-of-fit tests for the proportional hazards regression model. Biometrika 67, 145-153.

Slud, E.V. and Rubinstein, L.V. (1983). Dependent competing risks and summary survival curves. Biometrika 70, 643-649.

Received by Editorial Board member March, 1986, Revised June, 1987.

Recommended by L.J. Wei, University of Michigan, Ann Arbor, MI.