BIOMETRICS 52, 137–151
March 1996

Nonparametric Survival Estimation Using Prognostic Longitudinal Covariates

Susan Murray and Anastasios A. Tsiatis

Department of Biostatistics, Harvard School of Public Health,
Boston, Massachusetts 02115, U.S.A

SUMMARY

One of the primary problems facing statisticians who work with survival data is the loss of information that occurs with right-censored data. This research considers trying to recover some of this endpoint information through the use of a prognostic covariate which is measured on each individual. We begin by defining a survival estimate which uses time-dependent covariates to more precisely get at the underlying survival curves in the presence of censoring. This estimate has a smaller asymptotic variance than the usual Kaplan–Meier in the presence of censoring and reduces to the Kaplan–Meier (1958, *Journal of the American Statistical Association* **53**, 457–481) in situations where the covariate is not prognostic or no censoring occurs. In addition, this estimate remains consistent when the incorporated covariate contains information about the censoring process as well as survival information. Because the Kaplan–Meier estimate is known to be biased in this situation due to informative censoring, we recommend use of our estimate.

1. Introduction

The collection of survival type data is often accompanied by right censoring. Sometimes this censoring is the result of random dropout or loss to follow up. More often subjects are accrued and followed over specified periods of time, and at the scheduled end of a trial some failures have not occurred. Ideally we would like to be able to predict precisely when the failures would have occurred if they had been observable. Although predicting precise failure times of censored individuals is not possible, it may be possible to recover some of the lost failure information by studying the observable survival behavior of subjects in the study with similar characteristics. Prognostic covariate information which is collected on all individuals may somehow be incorporated into survival estimates to improve the efficiency of estimation. Cox (1983) suggested this type of approach in estimating parameters from parametric survival models. However, parametric assumptions are often too restrictive and may give unreliable estimates when assumptions are violated. Recently much research has focused on adjusting survival estimates for a predictive covariate or other information with fewer parametric assumptions. Malani (1995) suggests a modification of the redistribution to the right algorithm originally suggested by Efron (1967). Robins and Rotnitzky (1992) have also done similar work in this area. Several papers have focused on incorporating information from disease progression by modeling the relationships between progression and survival. Gray (1993) considers a three-state model in which the distribution of survival following progression might be influenced by the time of progression. Finkelstein and Schoenfeld (1994) use a similar three-state model suggesting various ways of estimating the conditional distribution of surviving given progression at some point t. The estimate we shall propose is a nonparametric estimate which incorporates a predictive covariate without modeling assumptions as to how the covariate interrelates specifically with survival. Our estimate is more efficient than the Kaplan–Meier (1958) estimate when the covariate incorporated is prognostic. This estimate also adjusts for the effects of informative censoring when the censoring information is captured by the prognostic covariate information. In Section 2, we look at a simplified version of the problem where the covariate information incorporated is time-independent. Motivation for the estimate is demonstrated using conditional probability argu-

Key words: Censoring; Covariate; Nonparametric; Survival; Kaplan–Meier estimate.

ments. The variance of the estimate is derived and compared to the variance of the Kaplan–Meier estimate in various scenarios. In Section 3, the estimate is extended to incorporate longitudinal covariate information. In Section 4, simulation and closed form results regarding the performance of the estimate are presented. An example using data from an AIDS trial is presented in Section 5. A discussion follows in Section 6.

2. Estimation with a Time-independent Covariate

Let T denote the time an individual would fail if the failure were observable. Let U denote the time an individual would become censored if the censoring process were observable. In this section Z will denote a categorical covariate taking on values $0, 1, \ldots, k$. We assume that T and U are conditionally independent given Z. Let $X = \min(T, U)$ be the observable event time and let $\Delta = I(T \leq U)$ be the failure indicator. Let S and H denote the survival distributions corresponding to T and U, respectively.

We define the weighted Kaplan–Meier (WKM) statistic as follows. Using conditional probability, we can rewrite the survival function.

$$S(t) = P(T > t) = \sum_{i=0}^{k} P(T > t \mid Z = i) P(Z = i) = \sum_{i=0}^{k} \theta_i S_i(t),$$

where θ_i is the probability a subject has covariate value i, $(i = 0, 1, \ldots, k)$ and $S_i(t)$ is the probability of survival conditional on having covariate value i. So we can estimate the survival from the right-hand side of the above equation.

$$WKM(t) = \sum_{i=0}^{k} \frac{n_i}{n} \hat{S}_i(t),$$

where $\hat{S}_i(t)$ is the Kaplan–Meier (KM) estimate among those with covariate value i, n_i is the number of subjects with covariate value i, and $n = \sum_{i=0}^{k} n_i$.

The variance of the WKM survival can be derived with a simple application of the conditional variance formula conditioning on the number in each of the covariate strata.

$$\begin{aligned} \operatorname{var}(\sqrt{n} WKM(t)) &= n \operatorname{var}\left(\sum_{i=0}^{k} \frac{n_i}{n} \hat{S}_i(t) \right) \\ &= nE\left(\operatorname{var}\left(\sum_{i=0}^{k} \frac{n_i}{n} \hat{S}_i(t) \mid \boldsymbol{n} \right) \right) + n \operatorname{var}\left(E\left(\sum_{i=0}^{k} \frac{n_i}{n} \hat{S}_i(t) \mid \boldsymbol{n} \right) \right) \\ &= nE\left\{ \sum_{i=0}^{k} \frac{n_i^2}{n^2} \frac{S_i^2(t)}{n_i} \int_0^t \frac{\lambda_i(u) du}{H_i(u) S_i(u)} \right\} + \sum_{i=0}^{k} \theta_i S_i^2(t) - \left(\sum_{i=0}^{k} \theta_i S_i(t) \right)^2 \\ &= \sum_{i=0}^{k} \theta_i S_i^2(t) \int_0^t \frac{\lambda_i(u) du}{H_i(u) S_i(u)} + \sum_{i=0}^{k} \theta_i (S_i(t) - \bar{S}(t))^2, \end{aligned}$$

where $\bar{S}(t) = \sum_{i=0}^{k} \theta_i S_i(t)$ and $\boldsymbol{n}$ is the vector of n_i's. This variance can be easily estimated. Let $\hat{\theta}_i = \frac{n_i}{n}, \hat{S}_i(t) = KM(t)$. Let $\tilde{N}(t)$ be the observed number of deaths at time t and $\tilde{Y}(t)$ be the observed number of individuals still at risk at time t. Estimate the variance with

$$\hat{\sigma}^2 = \sum_{i=0}^{k} \hat{\theta}_i \hat{S}_i^2(t) \int_0^t \frac{d\tilde{N}_i(u)}{\tilde{Y}_i(u)(\tilde{Y}_i(u) - \Delta \tilde{N}_i(u))} + \sum_{i=0}^{k} \hat{\theta}_i (\hat{S}_i(t) - \hat{S}(t))^2.$$

The first term in the variance is simply a weighted function of Greenwood's formula for the variance of the KM estimate for subjects with covariate value i.

In order to understand how our statistic behaves in comparison to the KM statistic we shall study the relationship between the estimates in the following situations:

(2.1) $S_i(t) = S(t)$ and $H_i(u) = H(u)$ for all i. This is a case in which either the KM or the WKM could be applied. Here the covariate that is conditioned upon has no prognostic value for survival so it should not provide additional information about a censored individual's survival status. Use of the WKM in this situation would be equivalent to using the KM estimate in terms of precision.

(2.2) $S_i(t)$ are not all equal and $H_i(u) = H(u)$ for all i. This is another case in which either estimate would be appropriate for use since informative censoring is not an issue. Here it becomes interesting to compare the performances of the estimates to one another. When the covariate being conditioned upon is predictive of survival it may be possible to recover information about a censored subject's survival status through the extraneous covariate information collected. In fact we shall show that in this case the variance of the WKM estimate is always smaller than or equal to the variance of the KM estimate.

(2.3) $S_i(t) = S(t)$ and $H_i(u)$ are not all equal. This is another case where both estimates are consistent. This is the only case in which the WKM estimate loses efficiency in comparison to the KM estimate and is a convincing argument against arbitrarily applying the WKM method with covariates that have no prognostic value.

(2.4) Neither the $S_i(t)$'s nor the $H_i(u)$'s are equal for all i. In this case the KM estimate is subject to bias from informative censoring. By conditioning on the covariate causing the informative censoring the WKM estimate remains consistent and therefore is the recommended estimate for this situation.

2.1 $S_i(t) = S(t)$ *and* $H_i(u) = H(u)$ *for all* i

Here the failure time distribution, T, is completely independent of the covariate. In this case it is easy to show that the variances of the two estimates are equal. Although both estimates serve equally well in this situation, one might choose in favor of the KM because of the slight reduction in computational calculations.

2.2 $S_i(t)$ *are not all equal and* $H_i(u) = H(u)$ *for all* i

Define $G(u) = 1/H(u)$. Notice that $G(u)$ is an increasing function of u. A term that comes into play in various ways in both the variance of the KM and the WKM estimates is $\int G(u)f(u)/S^2(u)$. Rewriting this term via partial integration we find that

$$\int G(u)\frac{f(u)}{S^2(u)} = \frac{G(u)}{S(u)}\Big|_0^t - \int_0^t \frac{G'(u)du}{S(u)} = \left\{\frac{G(t)}{S(t)} - 1\right\} - \int_0^t \frac{G'(u)du}{S(u)}.$$

Using this fact we can rewrite both the variance of the KM and the variance of the WKM accordingly as

$$\begin{aligned}\operatorname{var}(\sqrt{n}KM) = S^2(t)\int_0^t G(u)\frac{f(u)du}{S^2(u)} &= S^2(t)\left\{\frac{G(t)}{S(t)} - 1 - \int_0^t \frac{G'(u)du}{S(u)}\right\}\\ &= G(t)\bar{S}(t) - \bar{S}^2(t) - \bar{S}^2(t)\int_0^t \frac{G'(u)du}{\bar{S}(u)}\end{aligned}$$

and

$$\begin{aligned}\operatorname{var}(\sqrt{n}WKM) &= \sum_{i=0}^{k}\theta_i S_i^2(t)\int_0^t G(u)\frac{f_i(u)du}{S_i^2(u)} + \sum_{i=0}^{k}\theta_i S_i^2(t) - \bar{S}^2(t)\\ &= \sum_{i=0}^{k}\theta_i S_i^2(t)\left[\frac{G(t)}{S_i(t)} - 1 - \int_0^t \frac{G'(u)du}{S_i(u)}\right] + \sum_{i=0}^{k}\theta_i S_i^2(t) - \bar{S}^2(t)\\ &= G(t)\bar{S}(t) - \bar{S}^2(t) - \sum_{i=0}^{k}\theta_i S_i^2(t)\int_0^t \frac{G'(u)du}{S_i(u)}.\end{aligned}$$

So

$$\begin{aligned}\operatorname{var}(\sqrt{n}KM) - \operatorname{var}(\sqrt{n}WKM) &= \sum_{i=0}^{k}\theta_i S_i^2(t)\int_0^t \frac{G'(u)du}{S_i(u)} - \bar{S}^2(t)\int_0^t \frac{G'(u)du}{\bar{S}(u)}\\ &= \int_0^t G'(u)\bar{S}(u)\left[\sum_{i=0}^{k}\left(\frac{\theta_i S_i(u)}{\bar{S}(u)}\right)\left[\frac{S_i(t)}{S_i(u)}\right]^2 - \left[\sum_{j=0}^{k}\left(\frac{\theta_j S_j(u)}{\bar{S}(u)}\right)\frac{S_j(t)}{S_j(u)}\right]^2\right] du\end{aligned}$$

$$= \int_0^t G'(u)\bar{S}(u) \sum_{i=0}^{k} \left(\frac{\theta_i S_i(u)}{\bar{S}(u)}\right) \left[\frac{S_i(t)}{S_i(u)} - \sum_{j=0}^{k} \left(\frac{\theta_j S_j(u)}{\bar{S}(u)}\right) \frac{S_j(t)}{S_j(u)}\right]^2 du.$$

Notice that all terms in the right side of the above statement are positive. Therefore the difference, $\text{var}(\sqrt{n}KM) - \text{var}(\sqrt{n}WKM)$, is also positive implying that the variance of the WKM estimate is less than or equal to the variance of the KM estimate.

2.3 *$S_i(t) = S(t)$ and $H_i(u)$ are not all equal*
In this case

$$\text{var}(\sqrt{n}KM) - \text{var}(\sqrt{n}WKM) = S^2(t) \int_0^t \frac{\lambda(u)}{S(u)} \left[\frac{1}{\sum_{i=0}^{k} \theta_i H_i(u)} - \sum_{i=0}^{k} \frac{\theta_i}{H_i(u)}\right] du.$$

Note that the function $1/H(t)$ is a convex and increasing function of time. Hence by applying Jensen's inequality we can see that the above expression must be less than or equal to zero. In this case the WKM estimate is not the most efficient estimate available. From this and (2.1) we see that use of the WKM method with an arbitrary covariate would not be recommended. Substantial gains in efficiency come about only when censoring occurs and the covariate used in the WKM method is related to survival.

2.4 *Neither the $S_i(t)$'s nor the $H_i(u)$'s are equal for all i*
The KM estimate is subject to bias in this case due to informative censoring. Bias of this nature can often lead to very misleading results and should be avoided. If the covariate is predictive of survival and captures the source of bias, the WKM estimate provides consistent results and should always be used in preference to the KM estimate. This sort of situation comes up in many situations where a clinical marker is available. Recently in many AIDS trials, CD4 levels have been suggested as markers for survival. These CD4 counts are also highly related to an individual's censoring status. Hence this situation would lend itself quite favorably to the WKM method of estimation. Consistent results would not depend on the performance of CD4 as a marker, but only on its predictive value.

3. Estimation with a Stratified Time-dependent Covariate

Now suppose that our covariate is longitudinal in nature. Then we might observe the value of the covariate at a finite number of prespecified times $T_0^*, T_1^*, \ldots, T_s^*$. The choice of these times will be discussed later. In this section let Z_1 represent the covariate value at time T_0^*, Z_2 represent the covariate value at time $T_1^*, \ldots$, and Z_{s+1} represent the covariate value at time T_s^*. To simplify notation we shall assume that k categories are observable each time a covariate is recorded. The theory put forward will still hold true in situations where the number of categories varies across covariate observation times. For instance, in the following sections one may define k_1 categories at time T_0^*, k_2 categories at time T_1^*, and so on, substituting these numbers for k in the appropriate places. In fact, there is no theoretical restriction on the way categories are defined across time. Categories may be grouped together at later time points or may be defined in relation to several covariates. Variables can also be defined completely differently at each occasion. For instance Z_1 could be based on a subject's hemoglobin count and Z_2 could be based on a person's CD4 count.

Figure 1 indicates the possible covariate paths that may occur in our notationally simplified setting. At each point in time we would like to have an estimate of survival which takes advantage of all possible information. Note that seeing future covariate information for a subject is dependent on his survival. Similarly, survival depends on the past covariate path of a subject in the case a covariate is prognostic. To understand how best to incorporate all of this information we shall describe in detail how the estimate comes about for the situation where the covariate is observed twice ($s = 1$). First note that when knowledge of the covariate measured at time T_0^* is the only covariate information available we have already described a statistic in the last section which takes all possible information into account. So the WKM statistic is defined the same way as it was in the last section until time T_1^*. For $t > T_1^*$ we shall define the WKM statistic based on the following use of conditional probability.

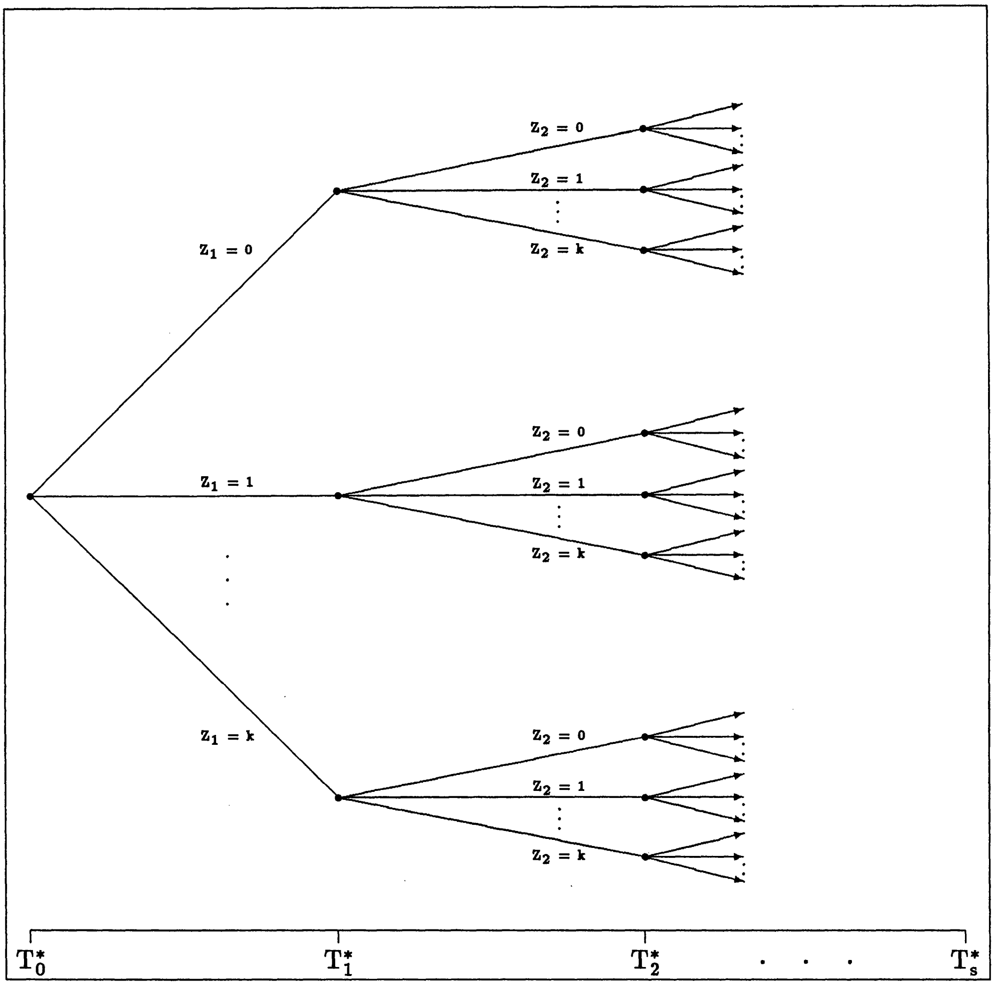

Figure 1. Possible paths of the longitudinal covariate, Z. Note that knowledge gained at each covariate look time $T_0^* \dots T_s^*$ causes the path to divide.

$$\begin{aligned}P(T > t) &= \sum_{i_1=0}^{k}\sum_{i_2=0}^{k} P(T > t, Z_1 = i_1, Z_2 = i_2)\\ &= \sum_{i_1=0}^{k}\sum_{i_2=0}^{k} P(T > t \mid T > T_1^*, Z_1 = i_1, Z_2 = i_2)P(Z_2 = i_2 \mid T > T_1^*, Z_1 = i_1)\\ &\qquad \times P(T > T_1^* \mid Z_1 = i_1)P(Z_1 = i_1)\\ &= \sum_{i_1=0}^{k}\sum_{i_2=0}^{k} S_{i_1 i_2}(t)\theta_{i_1 i_2} S_{i_1}(T_1^*)\theta_{i_1},\end{aligned}$$

where $\theta_{i_1 i_2}$ is the probability that a subject has covariate value i_2 measured at time T_1^*, conditional on the subject surviving at least to time T_1^* and previously having covariate value i_1 at time T_0^*, and $S_{i_1 i_2}(t)$ is the probability that a subject survives past time t, conditional on the subject surviving past time T_1^* and having covariate values i_1 at time T_0^* and i_2 at time T_1^*. This same type of conditioning argument extends easily to situations where more covariate information is observed. Let $\theta_{i_1 i_2 \dots i_m}$ represent the probability that a subject has covariate value i_m measured at time T_{m-1}^*, conditional on the subject surviving at least to time T_{m-1}^* and previously having covariate values i_1 at time T_0^*, i_2 at time $T_1^*, \ldots$, and i_{m-1} at time T_{m-2}^*, and let $S_{i_1 i_2 \ldots i_m}(t)$ represent the probability that a subject survives past time t, conditional on the subject surviving past time T_{m-1}^* and having covariate values i_1 at time T_0^*, i_2 at time $T_1^*, \ldots$, and i_m at time $T_{m-1}^*, m = 2, 3, \ldots, s+1$. Then we can rewrite $P(T > t)$ as

$$S(t) = \begin{cases} \sum_{i_1=0}^{k} S_{i_1}(t)\theta_{i_1} & 0 < t \le T_1^* \\ \sum_{i_1=0}^{k}\sum_{i_2=0}^{k} S_{i_1 i_2}(t)\theta_{i_1 i_2} S_{i_1}(T_1^*)\theta_{i_1} & T_1^* < t \le T_2^* \\ \sum_{i_1=0}^{k}\sum_{i_2=0}^{k}\sum_{i_3=0}^{k} S_{i_1 i_2 i_3}(t)\theta_{i_1 i_2 i_3} S_{i_1 i_2}(T_2^*)\theta_{i_1 i_2} S_{i_1}(T_1^*)\theta_{i_1} & T_2^* < t \le T_3^* \\ \vdots & \\ \sum_{i_1=0}^{k}\sum_{i_2=0}^{k}\cdots\sum_{i_{s+1}=0}^{k} S_{i_1 i_2 \ldots i_{s+1}}(t)\theta_{i_1 i_2 \ldots i_{s+1}} & \\ \qquad \times S_{i_1 i_2 \ldots i_s}(T_s^*)\theta_{i_1 i_2 \ldots i_s} \cdots S_{i_1 i_2}(T_2^*)\theta_{i_1 i_2} S_{i_1}(T_1^*)\theta_{i_1} & T_s^* < t. \end{cases}$$

Our Weighted Kaplan–Meier estimate then becomes

$$WKM(t) = \begin{cases} \sum_{i_1=0}^{k} \hat{S}_{i_1}(t)\frac{n_{i_1}}{n} & 0 < t \le T_1^* \\ \sum_{i_1=0}^{k}\sum_{i_2=0}^{k} \hat{S}_{i_1 i_2}(t)\frac{n_{i_1 i_2}}{n_{i_1 \cdot}} \hat{S}_{i_1}(T_1^*)\frac{n_{i_1}}{n} & T_1^* < t \le T_2^* \\ \sum_{i_1=0}^{k}\sum_{i_2=0}^{k}\sum_{i_3=0}^{k} \hat{S}_{i_1 i_2 i_3}(t)\frac{n_{i_1 i_2 i_3}}{n_{i_1 i_2 \cdot}} \hat{S}_{i_1 i_2}(T_2^*)\frac{n_{i_1 i_2}}{n_{i_1 \cdot}} \hat{S}_{i_1}(T_1^*)\frac{n_{i_1}}{n} & T_2^* < t \le T_3^* \\ \vdots & \\ \sum_{i_1=0}^{k}\sum_{i_2=0}^{k}\cdots\sum_{i_{s+1}=0}^{k} \hat{S}_{i_1 i_2 \ldots i_{s+1}}(t)\frac{n_{i_1 i_2 \ldots i_{s+1}}}{n_{i_1 i_2 \ldots i_s \cdot}} & \\ \qquad \times \hat{S}_{i_1 i_2 \ldots i_s}(T_s^*)\frac{n_{i_1 i_2 \ldots i_s}}{n_{i_1 i_2 \ldots i_{s-1} \cdot}} \cdots \hat{S}_{i_1 i_2}(T_2^*)\frac{n_{i_1 i_2}}{n_{i_1 \cdot}} \hat{S}_{i_1}(T_1^*)\frac{n_{i_1}}{n} & T_s^* < t, \end{cases}$$

where $n_{i_1 i_2 \ldots i_m}$ is the number of people having covariate values i_1 at time T_0^*, i_2 at time $T_1^*, \ldots$, and i_m at time $T_{m-1}^*, n_{i_1 i_2 \ldots i_{m-1} \cdot} = \sum_{i_m=0}^{k} n_{i_1 i_2 \ldots i_m}$, and $\hat{S}_{i_1 i_2 \ldots i_m}(t)$ is the conditional Kaplan–Meier survival estimate at time t given survival at T_{m-1}^* among those with past covariate values corresponding to $i_1, i_2, \ldots i_m, m = 2, 3, \ldots, s+1$. Note that we have used estimates of $\theta_{i_1 i_2 \ldots i_m}$ and $S_{i_1 i_2 \ldots i_m}(t)$ which are conditional on $X = \min(T, U) > T_{m-1}^*$ instead of $T > T_{m-1}^*$ since we can only observe the values of X. Therefore, we must assume that

$$P(Z_m = i_m \mid X > T_{m-1}^*, Z_1, \ldots, Z_{m-1}) = P(Z_m = i_m \mid Z_1, \ldots, Z_{m-1}, T > T_{m-1}^*) = \theta_{i_1 i_2 \ldots i_m}$$

and

$$P(T > t \mid X > T_{m-1}^*, Z_1, \ldots, Z_m) = P(T > t \mid T > T_{m-1}^*, Z_1, \ldots, Z_m) = S_{i_1 i_2 \ldots i_m}(t).$$

These two assumptions can be interpreted as uninformative censoring conditional on the covariate history along with past failure and censoring information. Note that these assumptions allow the censoring distribution to depend on the covariate history. In the literature of missing data this is often referred to as missing at random. This type of missingness occurs in many clinical trial situations. For instance censoring might occur more frequently among those with a steadily decreasing biological marker of one type or another. As patients become increasingly ill, they may be more likely to drop out of the study for personal reasons. In order to use the KM estimate, one must assume that the censoring mechanism, U, is completely independent of (T, Z). When censoring is related to past measurements of some biological marker the KM estimate becomes biased. Hence here is another indication of how weaker assumptions allow the WKM estimate to remain consistent in the presence of informative censoring.

The method for deriving the variance of the WKM estimate is inductive in nature. As is indicated by the following proof, one may describe the form for the variance of the estimate for $T_1^* < t \leq T_2^*$ through the form of the variance in the previous time interval $0 \leq t \leq T_1^*$. Similarly the variance of the estimate for $T_2^* < t \leq T_3^*$ can be described using the form of the variance in the previous time interval $T_1^* < t \leq T_2^*$. This pattern continues, always relating the current interval's variance to the previous interval's variance. The key to the proof below hinges upon using the conditional variance formula efficiently. We condition upon all of the failure, censoring, and covariate information accrued up until and including T_1^* with the exception of the value of Z_2. With this knowledge $\hat{S}_{i_1}(T_1^*)$ and n_{i_1} become deterministic functions. At this point the relationship between the methods for computing variances between intervals becomes clear. Let the notation $\mathcal{F}_{T_1^*-\mathcal{Z}_2}$ represent the survival, censoring, and covariate information up until and including T_1^* with the exception of the value of Z_2. Consider the following derivation of the variance for a particular interval in time, $T_{m-1}^* < t \leq T_m^*$.

$$\begin{aligned}
&\operatorname{var}(\sqrt{n}WKM(t)) \\
&\quad = \operatorname{var}\left(\sqrt{n}\sum_{i_1=0}^{k}\hat{S}_{i_1}(T_1^*)\frac{n_{i_1}}{n}\sum_{i_2=0}^{k}\hat{S}_{i_1 i_2}(T_2^*)\frac{n_{i_1 i_2}}{n_{i_1 \cdot}}\cdots\sum_{i_m=0}^{k}\hat{S}_{i_1\ldots i_m}(t)\frac{n_{i_1\ldots i_m}}{n_{i_1\ldots i_{m-1}\cdot}}\right) \\
&\quad = \operatorname{var}\left(\sqrt{n}\sum_{i_1=0}^{k}\hat{S}_{i_1}(T_1^*)\frac{n_{i_1}}{n}E_{i_1}\left[\sum_{i_2=0}^{k}\hat{S}_{i_1 i_2}(T_2^*)\frac{n_{i_1 i_2}}{n_{i_1 \cdot}}\cdots\sum_{i_m=0}^{k}\hat{S}_{i_1\ldots i_m}(t)\frac{n_{i_1\ldots i_m}}{n_{i_1\ldots i_{m-1}\cdot}} \mid \mathcal{F}_{T_1^*-\mathcal{Z}_2}\right]\right) \qquad (1) \\
&\qquad + E\left(\sum_{i_1=0}^{k}\hat{S}_{i_1}^2(T_1^*)\frac{n_{i_1}^2}{n}\operatorname{var}_{i_1}\left[\sum_{i_2=0}^{k}\hat{S}_{i_1 i_2}(T_2^*)\frac{n_{i_1 i_2}}{n_{i_1 \cdot}}\cdots\sum_{i_m=0}^{k}\hat{S}_{i_1\ldots i_m}(t)\frac{n_{i_1\ldots i_m}}{n_{i_1\ldots i_{m-1}\cdot}} \mid \mathcal{F}_{T_1^*-\mathcal{Z}_2}\right]\right). \qquad (2)
\end{aligned}$$

Let $S_{i_1\cdot}^{(m)}(t) = \sum_{i_2=0}^{k} S_{i_1 i_2}(T_2^*)\theta_{i_1 i_2}\cdots\sum_{i_m=0}^{k} S_{i_1\ldots i_m}(t)\theta_{i_1\ldots i_m}$. The conditional expectation in (1) reduces to $S_{i_1\cdot}^{(m)}(t)$. So we can rewrite (1) as

$$\operatorname{var}\left(\sqrt{n}\sum_{i_1=0}^{k}\hat{S}_{i_1}(T_1^*)\frac{n_{i_1}}{n}S_{i_1\cdot}^{(m)}(t)\right).$$

Using the same calculations as were used in finding the variance for the WKM for one covariate look we find that this term becomes

$$\sum_{i_1=0}^{k}\theta_{i_1}S_{i_1}^2(T_1^*)[S_{i_1\cdot}^{(m)}(t)]^2\int_0^{T_1^*}\frac{\lambda_{i_1}(u)du}{H_{i_1}(u)S_{i_1}(u)} + \sum_{i_1=0}^{k}\theta_{i_1}(S_{i_1}(T_1^*)S_{i_1\cdot}^{(m)}(t) - S_{\cdot}^{(m)}(t))^2,$$

where $S_{\cdot}^{(m)}(t) = \sum_{i_1=0}^{k} S_{i_1}(T_1^*)\theta_{i_1}\cdots\sum_{i_m=0}^{k} S_{i_1\ldots i_m}(t)\theta_{i_1\ldots i_m}$. If we rewrite (2) as

$$+\frac{1}{n}E\left(\sum_{i_1=0}^{k}\hat{S}_{i_1}^2(T_1^*)\frac{n_{i_1}^2}{n_{i_1\cdot}}\operatorname{var}_{i_1}\left[n_{i_1\cdot}^{\frac{1}{2}}\sum_{i_2=0}^{k}\hat{S}_{i_1 i_2}(T_2^*)\frac{n_{i_1 i_2}}{n_{i_1 \cdot}}\cdots\sum_{i_m=0}^{k}\hat{S}_{i_1\ldots i_m}(t)\frac{n_{i_1\ldots i_m}}{n_{i_1\ldots i_{m-1}\cdot}} \mid \mathcal{F}_{T_1^*-\mathcal{Z}_2}\right]\right),$$

then the conditional variance in this term also becomes a deterministic function which only depends on i_1. Since the variance is conditional on Z_1 and all failure time information up until time T_1^*, this conditional variance term is identical to the problem where baseline is set to time T_1^* and we have $m-1$ covariate looks. In other words, this term has the same form as the variance of the estimate in the last time interval, $T^*_{m-2} < t \le T^*_{m-1}$. Only the notation has been translated to include information specifying Z_1. Hence term (2) becomes

$$
\begin{aligned}
&\frac{1}{n}E\Bigg(\sum_{i_1=0}^{k} n_{i_1} E\left[\hat{S}^2_{i_1}(T^*_1)\frac{n_{i_1}}{n_{i_1\cdot}} \mid n\right] \\
&\qquad \times \operatorname{var}_{i_1}\left[n^{\frac{1}{2}}_{i_1\cdot}\sum_{i_2=0}^{k}\hat{S}_{i_1 i_2}(T^*_2)\frac{n_{i_1 i_2}}{n_{i_1\cdot}}\cdots\sum_{i_m=0}^{k}\hat{S}_{i_1\ldots i_m}(t)\frac{n_{i_1\ldots i_m}}{n_{i_1\ldots i_{m-1}\cdot}} \mid \mathcal{F}_{T^*_1}-\mathcal{Z}_2\right]\Bigg) \\
&\approx \frac{1}{n}E\Bigg(\sum_{i_1=0}^{k} n_{i_1}\left[\frac{S_{i_1}(T^*_1)}{H_{i_1}(T^*_1)}\right] \\
&\qquad \times \operatorname{var}_{i_1}\left[n^{\frac{1}{2}}_{i_1\cdot}\sum_{i_2=0}^{k}\hat{S}_{i_1 i_2}(T^*_2)\frac{n_{i_1 i_2}}{n_{i_1\cdot}}\cdots\sum_{i_m=0}^{k}\hat{S}_{i_1\ldots i_m}(t)\frac{n_{i_1\ldots i_m}}{n_{i_1\ldots i_{m-1}\cdot}} \mid \mathcal{F}_{T^*_1}-\mathcal{Z}_2\right]\Bigg) \\
&= \sum_{i_1=0}^{k}\frac{S_{i_1}(T^*_1)}{H_{i_1}(T^*_1)}\theta_{i_1}\operatorname{var}_{i_1}\left[n^{\frac{1}{2}}_{i_1\cdot}\sum_{i_2=0}^{k}\hat{S}_{i_1 i_2}(T^*_2)\frac{n_{i_1 i_2}}{n_{i_1\cdot}}\cdots\sum_{i_m=0}^{k}\hat{S}_{i_1\ldots i_m}(t)\frac{n_{i_1\ldots i_m}}{n_{i_1\ldots i_{m-1}\cdot}} \mid \mathcal{F}_{T^*_1}-\mathcal{Z}_2\right].
\end{aligned}
$$

So for $T^*_{m-1} < t \le T^*_m$,

$$
\begin{aligned}
&\operatorname{var}(\sqrt{n}WKM(t)) \\
&= \sum_{i_1=0}^{k}\theta_{i_1}S^2_{i_1}(T^*_1)[S^{(m)}_{i_1\cdot}(t)]^2\int_0^{T^*_1}\frac{\lambda_{i_1}(u)du}{H_{i_1}(u)S_{i_1}(u)} + \sum_{i_1=0}^{k}\theta_{i_1}(S_{i_1}(T^*_1)S^{(m)}_{i_1\cdot}(t) - S^{(m)}_{\cdot}(t))^2 \\
&\quad + \sum_{i_1=0}^{k}\frac{S_{i_1}(T^*_1)}{H_{i_1}(T^*_1)}\theta_{i_1}\operatorname{var}_{i_1}\left[n^{\frac{1}{2}}_{i_1\cdot}\sum_{i_2=0}^{k}\hat{S}_{i_1 i_2}(T^*_2)\frac{n_{i_1 i_2}}{n_{i_1\cdot}}\cdots\sum_{i_m=0}^{k}\hat{S}_{i_1\ldots i_m}(t)\frac{n_{i_1\ldots i_m}}{n_{i_1\ldots i_{m-1}\cdot}} \mid \mathcal{F}_{T^*_1}-\mathcal{Z}_2\right].
\end{aligned}
$$

From this last equation we can see that the same algorithm for finding the variance for $T^*_{m-2} < t \le T^*_{m-1}$ can be incorporated in finding the variance in the interval $T^*_{m-1} < t \le T^*_m$. To be more clear on how to calculate the variance for $T^*_{m-1} < t \le T^*_m$, we have included an algorithm in the appendix.

This estimate which incorporates longitudinal covariate information has the same features that were proven previously when the covariate was time-independent. As before, when the longitudinal covariate is not prognostic with respect to either the survival or censoring distributions the variance reduces to the KM variance. When the longitudinal covariate is not predictive with respect to the censoring distribution but is predictive with respect to survival, the variance of the WKM estimate is smaller than the variance of the KM estimate. In fact, the variance decreases with each additional prognostic covariate look incorporated. Proofs of these relationships can be found in Murray's thesis (1994). Other special cases arise when the covariate is longitudinal. For instance, in some cases it would not be unreasonable to suppose that a longitudinal covariate is predictive only up to some point in time. Since the WKM estimate should predict equally well without defining the later unprognostic strata, we would expect the estimate's variance to reflect this behavior. In fact, the variance of our WKM estimate at the later time points reduces to the form of the variance at the last point in time the covariate was predictive. This proof may also be located in Murray's thesis.

4. Simulation Studies and Other Results

Because a closed form of the variance is available, it is possible to derive asymptotic relative efficiencies (AREs). A simple example using exponential data is displayed in Table 1. In this example we created three categories of covariate values measured at baseline and two categories of covariate values measured at time T^*_1. This leads to six possible covariate paths. We assume for the purpose of this example that the paths are made up of piecewise exponentials with a hazard change at time T^*_1. Hence $S_{i_1}(t) = e^{-\lambda_{i_1}t}$ and $S_{i_1 i_2}(t) = e^{-\lambda_{i_1 i_2}(t-T^*_1)}I(T \ge T^*_1)$ for $i_1 = 0, 1, 2$ and $i_2 = 0, 1$. Specific values for the λ_{i_1}'s and the $\lambda_{i_1 i_2}$'s are displayed in Figure 2. We also assume an exponential censoring distribution with hazard λ_h that is independent of the covariate across time. So $H(t) = e^{-\lambda_h t}$. The subjects appeared roughly equally in each of the six categories.

Table 1

AREs for data with prognostic values of Z_1 and Z_2 at selected percentiles

% Censoring	$t : F(t) =$	ARE (WKM1:KM)[1]	ARE (WKM2:KM)[2]	ARE (WKM2:WKM1)
42.09	.3	1.02	1.02	1
42.09	.5	1.04	1.05	1
42.09	.7	1.07	1.12	1.05
54.79	.3	1.03	1.03	1
54.79	.5	1.08	1.09	1.01
54.79	.7	1.11	1.22	1.09
62.41	.3	1.05	1.05	1
62.41	.5	1.11	1.12	1.01
62.41	.7	1.12	1.26	1.12
71.56	.3	1.07	1.07	1
71.56	.5	1.16	1.18	1.02
71.56	.7	1.10	1.27	1.15

[1] WKM1 refers to the WKM estimate if using the time independent version of the statistic.

[2] WKM2 refers to the WKM estimate when covariate information Z_1 and Z_2 are both used.

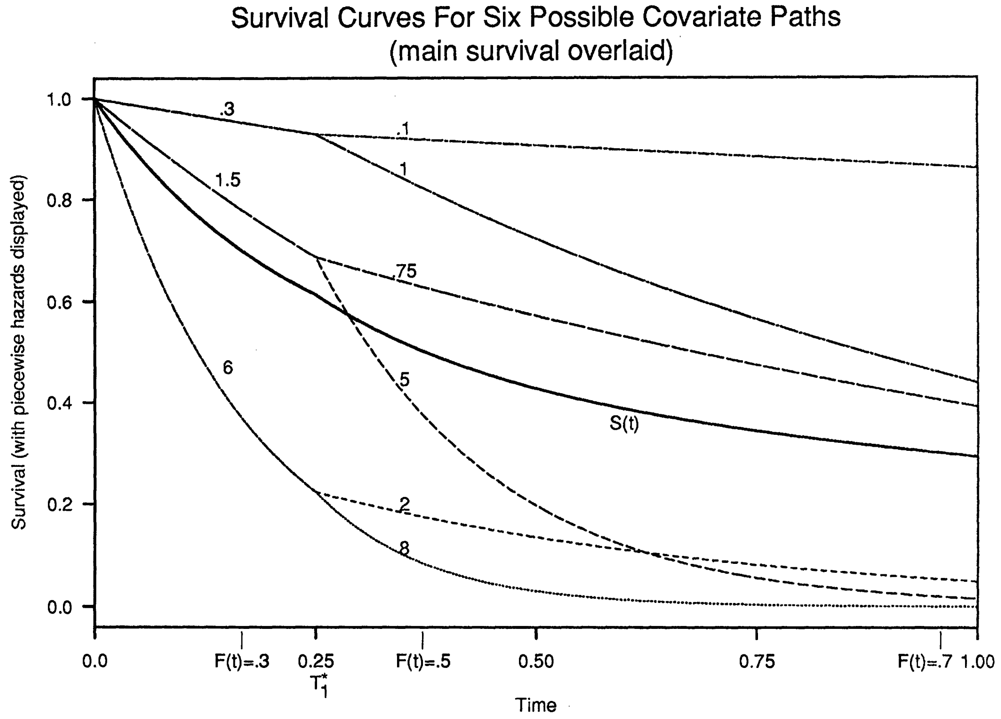

Figure 2. The piecewise exponential survival curves corresponding to the six possible covariate paths which are used to get the results in Tables 1 and 2. Hazards are displayed above the curves. Covariates were assumed to be measured at baseline and again at time .25. The overall survival probability is also displayed.

Depending on the amount of covariate information incorporated in the survival estimate, the formula for the variance changes according to the derivations in Section 3. Table 1 uses the appropriate variance formulas to derive AREs relating the KM estimate, the WKM estimate that incorporates baseline information only, and the WKM estimate that incorporates covariate information both at baseline and time T_1^*. Various degrees of censoring are explored in the various rows for $\lambda_h = 1$, 2, 3 and 5. Higher values of λ_h yield higher censoring percentages. Corresponding to each level of censoring we explore the AREs at three quantiles of the underlying failure time distribution. Notice that at the thirty-percent quantile the two WKM survival estimates are equivalent since covariate information collected at time T_1^* is not yet available. In the third column of Table 1 are AREs comparing the WKM estimate, which uses only covariate information collected at baseline, to the KM estimate. In the fourth column are AREs comparing the WKM estimate, which uses all available covariate information, to the KM estimate. The fifth column displays the AREs comparing the WKM estimate which uses both measured covariate times to the WKM estimate which uses only baseline covariate information. Several important and instructive patterns emerge in this table. As expected, all AREs are greater than one, implying that a gain is to be made by using as much prognostic longitudinal covariate information as possible. Notice that for the fifty- and seventy-percent quantiles the WKM estimate using information collected at baseline and time T_1^* improves over both the KM and the WKM that utilizes baseline covariate information alone. Also notice that the AREs tend to increase in comparison to the KM for data that has more censored observations. This is due to the supplementation of failure information that is provided by the use of covariates when censoring occurs. Hence, higher degrees of censoring affect the variance of the KM estimate much more than they affect the variance of the WKM estimates.

In addition to these closed form ARE results, simulations were run on data generated from the previously described piecewise exponential distributions. These simulations used a sample size of 150 subjects, with 25 subjects per possible covariate path. For each censoring percentage considered in Table 2, 5,000 simulations were done. Results of these simulations confirmed the ARE results of Table 1. Table 2 contains the MSE results from these simulations.

Table 2

MSEs $\times 10^3$ *from simulation on data with prognostic values of* Z_1 *and* Z_2 *at selected percentiles*

% Censoring	$t : F(t) =$	MSE(KM)	MSE(WKM1)[1]	MSE(WKM2)[2]
42.09	.3	1.15	1.13	1.13
42.09	.5	1.31	1.24	1.23
42.09	.7	1.67	1.55	1.44
54.79	.3	1.28	1.25	1.25
54.79	.5	1.79	1.68	1.67
54.79	.7	3.50	3.17	2.84
62.41	.3	1.41	1.37	1.37
62.41	.5	2.51	2.32	2.35
62.41	.7	7.57	6.62	5.07
71.56	.3	1.78	1.64	1.64
71.56	.5	4.52	4.19	4.00
71.56	.7	21.77	21.00	9.12

[1] WKM1 refers to the WKM estimate if using the time independent version of the statistic.

[2] WKM2 refers to the WKM estimate when covariate information Z_1 and Z_2 are both used.

Another important issue in applying this research is how best to incorporate a continuous time-dependent covariate. The proposed methodology requires the covariate to be incorporated in categorical form. Hence the investigator must decide how many categories to subdivide the continuous covariate into at each time T_i^* and how many times T_i^* to measure the time-dependent covariate. Although it is tempting to create many categories measured at many times, this is not advisable in moderate-sized data sets. The WKM estimate is consistent only for times t smaller than the minimum event time from any strata. Hence defining too many strata across time in moderate-sized data sets will restrict the range of times for which the WKM estimate exists. However, gains in efficiency will occur even when the number of times T_i and covariate levels are kept very small. In the following simulation we have studied the gains in efficiency associated with incorporating a continuous covariate in categorical form. Using the Cox proportional hazards model we simulated a failure time distribution which depends on a time-varying covariate. Hence for individual i, $Z_i(t) = \alpha_{1i} + \alpha_{2i}t$. The random variables $(\alpha_{1i}, \alpha_{2i})$, $i = 1, \ldots n$, were taken to be

Table 3
AREs[1] *for continuous covariate data by amount of defined strata*

		No. of baseline covariate strata				
Percentile	No. of covariate strata at year 1	1	2	3	4	5
30th	1	1.000	1.077	1.092	1.101	1.103
30th	2	1.068	1.121	1.133	1.132	1.140
30th	3	1.084	1.141	1.151	1.157	1.167
30th	4	1.088	1.147	1.160	1.158	1.170
30th	5	1.091	1.152	1.168	1.170	1.181
50th	1	1.000	1.078	1.094	1.103	1.105
50th	2	1.091	1.143	1.154	1.157	1.160
50th	3	1.105	1.160	1.169	1.180	1.191
50th	4	1.110	1.169	1.182	1.183	1.194
50th	5	1.113	1.174	1.189	1.197	1.205
70th	1	1.000	1.074	1.091	1.101	1.102
70th	2	1.101	1.150	1.160	1.166	1.165
70th	3	1.121	1.171	1.182	1.192	1.201
70th	4	1.126	1.185	1.200	1.203	1.206
70th	5	1.131	1.197	1.206	1.231	1.228

[1] AREs based on 500 simulations with sample size 500.

normally distributed and independent with means $(180, -100)$ and standard deviations (40,20). The hazard function for the failure time was of the form

$$\lambda(t \mid Z(t)) = \exp\{-0.046 Z(t)\},$$

where the coefficient was chosen to be reasonably prognostic. The censoring distribution was chosen to be Uniform(0,4). To include the continuous time-varying covariate information we considered the baseline covariate information and also the covariate information available at 1 year. The continuous covariate information was then stratified evenly by its quantiles. For instance, to create five categories of Z_1 we divided $Z(0)$ according to its 20th, 40th, 60th, and 80th quantiles. Similarly, to create five categories of Z_2 we divided $Z(1)$ according to these same quantiles. For each set of data simulated we created strata definitions with one to five baseline categories for Z_1 and one to five categories for Z_2 at year 1. These simulations used a large sample size of 500. Asymptotic relative efficiencies comparing the WKM estimate to the KM estimate were calculated at three different time points corresponding roughly to the 30th, 50th, and 70th percentiles of the failure time distribution. These are displayed in Table 3 for all strata definitions considered. We see that more covariate information used corresponds to greater gains in efficiency. Earlier we proved that asymptotic variances decrease when more longitudinal covariate looks are incorporated into the estimates. This simulation reaffirms that result. However, we also find that finer stratifications of incorporated continuous covariates result in higher efficiency gains. In any particular analysis the degree to which a continuous covariate should be stratified will depend to a large degree on the sample size available. In this simulation it appears that most efficiency gains for each time T_i^* occur with relatively few covariate strata, so even moderate-sized data sets could benefit from this method of estimation. This simulation also suggests that the gains in efficiency associated with additional covariate looks at times T_i^* surpass the gains made by finer stratification of each covariate Z_i.

5. Example

As an illustrative example in how to use this methodology we have looked at survival data in 524 AIDS patients who had had a first episode of *Pneumocystis carinii* pneumonia. This data comes from a randomized clinical trial previously examined by Fischl et al. (1990) in which patients were assigned either low dose ($n = 262$) or high dose ($n = 262$) zidovudine regimens. This trial is coming to a close so the data is fairly complete. Hence, to illustrate our methods we decided to recreate the data that would have been available at an earlier time in the trial. We specifically chose the date January 31, 1988, as our analysis time since at this particular time all of the participants in the study had been registered and substantial censoring existed in the data (87%). This is precisely the situation in which our methods for estimation become desirable since our estimate recovers some of the lost information caused by censoring. To use our methods it is also necessary to identify predictive categorical covariates from which the recovered information is gathered by our estimate. The usual precautions associated with identifying predictive covariates apply here. Excessive data snooping should not be employed. Covariates artificially constructed to appear predictive would falsely deflate the variance of the WKM estimate. In this trial CD4 count and hemoglobin level were known to be modestly predictive of survival, so a categorical combination of these variables would be of interest. Interestingly, treatment arm was not particularly predictive at this early stage in the study. Hemoglobin level and CD4 count were collected as continuous covariates so categorical versions of these covariates were constructed using the quantiles of these variables. The number of categories and covariate looks were kept minimal due to the sample size of this data. After minor exploratory analysis, categorical covariates were constructed at baseline and 200 days. At baseline two categories were formed from the baseline CD4 and hemoglobin measurements. For each of these two baseline categories, three categories were formed at 200 days based on CD4 and hemoglobin measurements taken at this time. That is, $Z_1 = 1, 2$ and $Z_2 = 1, 2, 3$ for a total of six possible covariate paths. The categorical covariate, Z_1, has virtually no predictive value until day 225 or so in this study. Hence in this range the use of covariate Z_1 would not improve the survival estimate. For the purpose of illustration we will present a few different analyses.

(5.1) The first naive analysis will use Z_1 alone to construct a survival estimate. As indicated before, we expect this estimate to have a larger variance during the period in which the two underlying survival curves are close. After day 225, the covariate Z_1 becomes modestly predictive and we expect the variance of the WKM estimate to be smaller than the variance of the KM estimate.

(5.2) Another naive analysis will use Z_1 and Z_2 to construct a survival estimate. Since the estimate before time T_1^* will be the same as in (5.1), we expect the same problems with the WKM variance in this range. However, since more information is being incorporated at 200 days, we expect this estimate to improve upon the estimate calculated in (5.1) following day 200.

(5.3) The preferred way to analyze this data involves utilizing the flexibility in defining covariates which our method allows. We would like to gain the efficiency that is possible to gain after day 225 from prognostic covariate information without paying the penalty of defining unprognostic covariates early on. Define

$$Z_1^* = \{1$$

and

$$Z_2^* = \begin{cases} 1, & \text{if } (Z_1 = 1 \text{ and } Z_2 = 1) \\ 2, & \text{if } (Z_1 = 1 \text{ and } Z_2 = 2) \\ 3, & \text{if } (Z_1 = 1 \text{ and } Z_2 = 3) \\ 4, & \text{if } (Z_1 = 2 \text{ and } Z_2 = 1) \\ 5, & \text{if } (Z_1 = 2 \text{ and } Z_2 = 2) \\ 6, & \text{if } (Z_1 = 2 \text{ and } Z_2 = 3). \end{cases}$$

Notice that before 200 days the WKM estimate using Z_1^* and Z_2^* will be identical to the KM estimate. However since we've included all prognostic information at baseline and 200 days in Z_2^*, our estimate will perform similarly to the estimate in (5.2) from 200 days on. Hence we've maximized our efficiency as much as possible across time.

Table 4 reveals in more detail the average efficiency gains for discretized time intervals. The asymptotic relative efficiencies of the various WKM estimates to the KM estimate were calculated at all event times and then averaged within several mutually exclusive time intervals. Notice how the AREs tend to increase at the later time intervals. Two factors are contributing to this effect. Most of the censoring in this data occurs at the later time intervals. Hence most of the failure time information that is recovered using covariate information improves survival estimation in these regions. Another factor that increases the efficiency of the WKM survival estimate in the later time regions is the increased prognostic value of the various covariates at these later time points.

6. Discussion

One can view this problem as a nonhomogeneous Markov chain with finitely many states. The sample proportions estimating the $\theta_{i_1 i_2 \ldots i_m}$'s and the conditional KM estimates for $S_{i_1 i_2 \ldots i_m}(t)$ have

Table 4
Average AREs across discrete sections of time

Days	ARE(WKM1(5.1):KM)	ARE(WKM2(5.2):KM)	ARE(WKM2(5.3):KM)
0–200	.9597	.9597	1.0000
201–300	1.0026	1.0087	1.0094
301–350	1.0483	1.0498	1.0424
351–390	1.0690	1.0912	1.0806
391–425	1.2550	—[1]	—[1]

[1] Survival estimates (5.2) and (5.3) were not available in the last time interval of this table.

been shown to be maximum likelihood estimates by Aalen and Johansen (1978). Because our estimation of the overall survival probability is a linear combination of these terms, this implies that our estimator is maximum likelihood. It turns out that the WKM estimate reduces to Malani's redistribution to the right estimate for the categorical variable case.

There are many useful applications of this method for survival estimation. For instance in many clinical trials, potentially prognostic laboratory measurements are being collected over time in ancillary companion protocols to the main therapeutic protocols. This additional information might be very useful in providing us with more precise estimates in studies where there are many censored observations. This estimate would also be useful in incorporating marker information in covariate form. No assumptions about the effectiveness of the marker would be necessary so long as the marker is in some way predictive. This has recently been the subject of much discussion in AIDS research.

As has been previously mentioned, the variance of the WKM estimate is always at least as small as the variance of the KM estimate in settings where censoring in uninformative. Therefore use of the WKM estimate would be a welcome change in these situations. The WKM estimate is also recommended when censoring is informative between strata of a prognostic categorical covariate. In this situation the KM estimate is subject to bias and should not be used. We would not recommend this method with arbitrary covariates which are unrelated to future survival. Although in theory the estimate would reduce to the KM estimate when censoring in uninformative, there might be a price to pay in terms of efficiency if censoring is informative in some way.

All of the results discussed here are asymptotic in nature. We have also run simulations to investigate properties of this estimate in small samples. The asymptotic results are closely approximated with occasional minor perturbations. One potential drawback to using the WKM estimate in very small samples is the range over which the estimate can be properly defined. For instance, if the last individual at risk in a particular covariate strata becomes censored, the KM for that strata cannot be consistently estimated past the censoring time. Because the WKM averages these KM estimates, the WKM estimate cannot be defined past that particular censoring time either. In large samples this problem does not tend to come up often since the KM usually is very close to zero at this point and can be comfortably labeled as such without affecting the consistency of the estimate. There are ways to get around this problem in practice since the data analyst has control over how to create covariate strata. For instance, if the range of the WKM estimate seems too small it is likely that there are too many strata and that some of them may be collapsed. There is no restriction in this theory to having different definitions of Z_i over time. It may be advisable in certain situations to reduce the number of categories in Z_i at later measurement times T^*_{i-1} to avoid a premature end to the WKM estimate. It may also be advisable to incorporate strategies of covariate definition similar to that of (5.3) to avoid defining unprognostic strata in finite time intervals. This tactic may also prove beneficial if most censoring occurs towards the tail area of the survival estimate.

Acknowledgements

This work was supported by Grant CA-51962 from the National Cancer Institute, Grant CA-09337 from the National Cancer Institute, and Grant AI-31789 from the National Institute of Allergy and Infectious Diseases. This work was also supported in part by grant CA-39640 from the National Institutes of Health. The authors would like to thank Nan Laird for her helpful comments and suggestions.

Résumé

L'un des principaux problèmes que rencontrent les statisticiens qui travaillent sur des données de survie, est la perte d'information associée à la censure à droite. Ce travail s'intéresse à une tentative de récupération d'une partie de cette information à travers l'utilisation d'un facteur pronostique mesuré sur chacun des sujets. Nous commençons par définir une estimation de la survie qui utilise des variables dépendantes du temps pour obtenir plus précisément la distribution de survie sous jacente en présence de censure. Cette estimation a une variance asymptotique plus faible que celle de l'estimateur habituel de Kaplan–Meïer en présence de censure, elle revient à celle de Kaplan–Meïer dans le cas où la covariable n'est pas pronostique ou si aucune censure n'intervient. De plus, cette estimation est consistante si la covariable prise en compte contient de l'information tant sur le processus de censure que sur la survie. Du fait que l'estimateur de Kaplan–Meïer est connu pour son biais dans cette situation, nous recommandons d'utiliser notre procédure.

References

Aalen, O. and Johansen, S. (1978). An empirical transition matrix for non-homogeneous Markov chains based on censored observations. *Scandinavian Journal of Statistics* **5,** 141–150.

Cox, D. R. (1983). A remark on censoring and surrogate response variables. *Journal of the Royal Statistical Society, Series B* **45,** 391–393.

Efron, B. (1967). The two sample problem with censored data. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability,* Vol. IV, 831–853. Berkeley, California: University of California Press.

Finkelstein, D. M. and Schoenfeld, D. A. (1994). Analysing survival in the presence of an auxiliary variable. *Statistics in Medicine* **13,** 1747–1754.

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

Gray, R. J. (1993). A kernel method for incorporating information on disease progression in the analysis of survival. *Biometrika* **81,** 527–539.

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

Malani, H. M. (1995). A modification of the re-distribution to the right algorithm using disease markers. *Biometrika,* **82,** 515–526.

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, Massachusetts.

Robins, J. M. and Rotnitzky, A. (1992). Recovery of information and adjustment for dependent censoring using surrogate markers. In *AIDS Epidemiology: Methodological Issues*, N. Jewell, K. Dietz, and V. Farewell (eds), 297–331. Boston: Birkhäuser–Boston.

Received November 1993; *received January* 1995; *accepted February* 1995.

Appendix

Algorithm for Calculating the Variance of the WKM Estimate for $T^*_{m-1} < t \leq T^*_m$.

Step 1. Calculate

$$V_{i_1 \ldots i_{m-1}}(t) = \sum_{i_m=0}^{k} \theta_{i_1 \ldots i_m} S^2_{i_1 \ldots i_m}(t) \int_{T^*_{m-1}}^{t} \frac{\lambda_{i_1 \ldots i_m}(u)du}{H_{i_1 \ldots i_m}(u) S_{i_1 \ldots i_m}(u)} + \sum_{i_m=0}^{k} \theta_{i_1 \ldots i_m} (S_{i_1 \ldots i_m}(t) - S^{(m)}_{i_1 \ldots i_{m-1}.}(t))^2$$

for all $i_1, \ldots, i_{m-1}$. This is simply the conditional variance from the outermost branches of the covariate path tree formed from $Z_1, \ldots, Z_m$ when all information up until time T^*_{m-1} is known except for the value of Z_m. In the outermost tail of the covariate path the problem reduces to the time independent covariate case.

Step 2. Calculate

$$\begin{aligned} &V_{i_1 \ldots i_{m-2}}(t) \\ &= \sum_{i_{m-1}=0}^{k} \theta_{i_1 \ldots i_{m-1}} S^2_{i_1 \ldots i_{m-1}}(T^*_{m-1})[S^{(m)}_{i_1 \ldots i_{m-1}.}(t)]^2 \int_{T^*_{m-2}}^{T^*_{m-1}} \frac{\lambda_{i_1 \ldots i_{m-1}}(u)du}{H_{i_1 \ldots i_{m-1}}(u) S_{i_1 \ldots i_{m-1}}(u)} \\ &\qquad + \sum_{i_{m-1}=0}^{k} \theta_{i_1 \ldots i_{m-1}} (S_{i_1 \ldots i_{m-1}}(T^*_{m-1}) S^{(m)}_{i_1 \ldots i_{m-1}.}(t) - S^{(m)}_{i_1 \ldots i_{m-2}.}(t))^2 V_{i_1 \ldots i_{m-1}}(t) \end{aligned}$$

for all $i_1, \ldots, i_{m-2}$.

Step j, $j = 3, \ldots, m-1$. Calculate

$$\begin{aligned} V_{i_1 \ldots i_{m-j}}(t) = &\sum_{i_{m-j+1}=0}^{k} \theta_{i_1 \ldots i_{m-j+1}} S^2_{i_1 \ldots i_{m-j+1}}(T^*_{m-j+1})[S^{(m)}_{i_1 \ldots i_{m-j+1}.}(t)]^2 \\ &\times \int_{T^*_{m-j}}^{T^*_{m-j+1}} \frac{\lambda_{i_1 \ldots i_{m-j+1}}(u)du}{H_{i_1 \ldots i_{m-j+1}}(u) S_{i_1 \ldots i_{m-j+1}}(u)} \\ &+ \sum_{i_{m-j+1}=0}^{k} \theta_{i_1 \ldots i_{m-j+1}} (S_{i_1 \ldots i_{m-j+1}}(T^*_{m-j+1}) S^{(m)}_{i_1 \ldots i_{m-j+1}.}(t) \\ &- S^{(m)}_{i_1 \ldots i_{m-j}.}(t))^2 V_{i_1 \ldots i_{m-j+1}}(t) \end{aligned}$$

for all $i_1, \ldots, i_{m-j}$.

$$\vdots$$

Step m. Calculate

$$\begin{aligned} &V(t) = \\ &\sum_{i_1=0}^{k} \theta_{i_1} S^2_{i_1}(T^*_1)[S^{(m)}_{i_1.}(t)]^2 \int_0^{T^*_1} \frac{\lambda_{i_1}(u)du}{H_{i_1}(u) S_{i_1}(u)} + \sum_{i_1=0}^{k} \theta_{i_1} (S_{i_1}(T^*_1) S^{(m)}_{i_1}(t) - S^{(m)}_{i_1.}(t))^2 V_{i_1}(t). \end{aligned}$$

The last step of the algorithm provides the variance, $V(t)$, for $T^*_{m-1} < t \leq T^*_m$.