Published in final edited form as:

J Am Stat Assoc. 1999 March 1; 94(445): 146–153.

Nonparametric Estimation of a Recurrent Survival Function

Mei-Cheng Wang [Professor] and

Department of Biostatistics, School of Hygiene and Public Health, Johns Hopkins University,
Baltimore, MD 21205

Shu-Hui Chang [Associate Professor]

School of Public Health, National Taiwan University, Taipei, Taiwan

Mei-Cheng Wang: mcwang@jhsph.edu

Abstract

Recurrent event data are frequently encountered in studies with longitudinal designs. Let the recurrence time be the time between two successive recurrent events. Recurrence times can be treated as a type of correlated survival data in statistical analysis. In general, because of the ordinal nature of recurrence times, statistical methods that are appropriate for standard correlated survival data in marginal models may not be applicable to recurrence time data. Specifically, for estimating the marginal survival function, the Kaplan-Meier estimator derived from the pooled recurrence times serves as a consistent estimator for standard correlated survival data but not for recurrence time data. In this article we consider the problem of how to estimate the marginal survival function in nonparametric models. A class of nonparametric estimators is introduced. The appropriateness of the estimators is confirmed by statistical theory and simulations. Simulation and analysis from schizophrenia data are presented to illustrate the estimators' performance.

Keywords

Correlated survival data; Frailty; Kaplan-Meier estimate; Longitudinal designs; Recurrent event

1. Introduction

Recurrent event data are frequently encountered in studies with longitudinal designs. To analyze recurrent event data, the focus can be placed on the time between events, the transition probability from the current event to the next, or the rate of occurrence of events over time. The appropriateness of a selected model depends on the nature of recurrent events as well as on the research interest of the study. In the statistical literature, various regression models have been considered to address different types of questions, including conditional models by Prentice, Williams, and Peterson (1981), semiconditional models by Pepe and Cai (1993), Poisson intensity models by Andersen and Gill (1982), and recurrence rate models by Lawless and Nadeau (1995).

In longitudinal studies, two sampling designs are commonly used for the collection of recurrent event data: The first design uses the initial occurrence of an event as the enrollment criterion, and repeated occurrences of the same event are observed within a time period. In this case the occurrence time of the initial event is defined as the time origin, 0. Many registry datasets can be identified as this type. In the second design, subjects are sampled from a target population, and recurrent events are observed during a follow-up period. In this case the occurrence of an initial event is not a requirement for recruitment, and thus it is possible that no events occurred during the follow-up period. In this article we develop statistical methods focusing on the first design.

Suppose that a sequence of events is to be observed in a longitudinal study, and the start time of follow-up is the time of the initial event. Let i be the index for an individual or subject and let j be the index for an event. Let $j = 0$ denote the index for the initial event and let T_{ij} denote the time from the $(j - 1)$ th to the j th event for the i th subject, $j = 1, 2, \dots$, and $i = 1, \dots, n$. Let C_i be the censoring time, be the time between the initial event and the end of follow-up, let G be the survival function of C_i and let $N_i = \{T_{ij} : j = 1, 2, \dots\}$. Assume that $(N_1, C_1), (N_2, C_2), \dots, (N_n, C_n)$ are independent and identically distributed. Let m_i denote the index satisfying

$$\sum_{j=1}^{m_i-1} T_{ij} \leq C_i$$

and

$$\sum_{j=1}^{m_i} T_{ij} > C_i.$$

Note that m_i is a random variable that takes the integer values $1, 2, \dots$. Throughout the article, for ease of notation, we use the lower case notation m_i to represent either a random variable or a realization value. The term “uncensored data” refers to the collection of $(t_{i1}, t_{i2}, \dots, t_{i, m_i})$ and “censored data” refers to $(t_{i1}, t_{i2}, \dots, t_{i, m_i-1}, t_{i, m_i}^+)$, Where t_{i, m_i}^+ is the time from event $m_i - 1$ to the end of follow-up.

The analyses of recurrence time data are frequently conducted using inappropriate methods in medical research. Aalen and Husebye (1991), for example, indicated statistical difficulties in the analyses of recurrence time data, with an emphasis on gastroenterology. They used a parametric approach to study their problem.

A particular feature in recurrence time data is that the last recurrence time T_{i, m_i} is always biased. We take the inverse sampling example (Cochran 1977) to illustrate the problem of biased estimation. Let T_{ij} be Bernoulli(p_i) and let the censoring time C_i be a fixed positive integer. The example can be thought of as a sequence of coin flips that stops when $C_i + 1$ heads are observed. In this case $\Pr(T_{mi} = 1) = 1$, and thus T_{mi} is clearly observed subject to biased sampling. Consider the case where $p_i = p, 0 < p < 1$. It is known that $C_i/(m_i - 1)$ is unbiased for p , and thus the expected value of

$$\frac{1}{m_i} \sum_{j=1}^{m_i} T_{ij} = \frac{C_i + 1}{m_i}$$

is larger than p . In contrast, the estimator $\{C_i + 1\}/m_i$ is asymptotically unbiased, because the combined coin flips form a sequence of independent Bernoulli trials with trial size $1/m_i$. Suppose now that the success probability p_i varies with different coins and that the parameter of interest, p , is the mean of p_i . In this case $\{C_i + 1\}/m_i$ estimates the mean success probability over coin flips, not the success probability over coins. Both the statistics $\{C_i + 1\}/m_i$ and $n^{-1} \sum \{C_i + 1\}/m_i$ in this case are biased for the estimation of p .

Our general one-sample model involves are two required assumptions:

A1: There exists a baseline random variable or vector $\mathbf{X}_j$, which may be unobservable or partially observable, so that the recurrence times $T_{j1}, T_{j2}, \dots$ are independent and identically (iid) distributed given $\mathbf{X}_j = x_j$.

A2: The censoring time C_j is independent of $(N_j, \mathbf{X}_j)$.

In this article, interests are centered on the marginal survival function of the time between two successive events, termed the *recurrent survival function*. The research topic is a fundamental one-sample problem when the recurrence times from different episodes have the same marginal distribution.

Let $F_{\mathbf{X}}$ denote the distribution function of the baseline random vector $\mathbf{X}_j$. Assume the existence of the conditional density of T_{ij} given x_j , denoted by $f(t|x_j)$. The recurrent survival function of T_{ij} can be expressed as

$$S(t) = \text{pr}(T_{ij} > t) = \int \int_t^\infty f(u|x) d\mu(u) dF_{\mathbf{X}}(x), \quad (1)$$

with μ denoting the σ -finite dominating measure of f .

Assumption A1 specifies a condition commonly used in frailty models to characterize the correlation of recurrence times from the same subject. Assumption A2 is essentially the usual *independent censoring condition*. The validity of assumptions A1 and A2 generally depends on the nature of recurrence times as well as on the specific censoring pattern in a study, and thus should be examined with care. In A1 the frailty $\mathbf{X}_j$ is used as a quantitative factor to characterize the cluster correlation, although the frailty assumption is stronger than necessary in the development of our methods. We discuss the possibility of using weaker conditions in Section 4.

The article is organized as follows. In Section 2 we derive a class of nonparametric estimators of S from uncensored data using hazard estimation techniques. In Section 3 we illustrate the proposed methods by simulation and analysis of the Denmark schizophrenia data. In Section 4 we provide a brief concluding discussion.

2. Estimation Of Survival Function

To understand how to estimate $S(t)$ from censored recurrence times, it is helpful to first consider estimation of $S(t)$ from uncensored recurrence times. Define

$$m_i^* = \begin{cases} 1 & \text{if } m_i = 1 \\ m_i - 1 & \text{if } m_i \geq 2 \end{cases}$$

and

$$U_i(t) = \frac{1}{m_i^*} \sum_{j=1}^{m_i^*} I(T_{ij} > t).$$

Note that the last recurrence time is not used for constructing U_j , unless $m_j = 1$. Construction of U_j is based on the idea of replacing $I(T_{j1} > t)$ by the average of m_j^* indicators. When m_j

2, the last recurrence time is not used in the average, to avoid the sampling bias. It can be shown that $T_{i1}, \dots, T_{i,m_i^*}$ are identically distributed conditional on $(m_i, \mathbf{X}_i, C_i)$, and this implies that $E[U_i(t)] = S(t)$. Define $a_i = a(C_i)$, with $a(\cdot)$ representing a positive-valued function subject to the constraint $E(a_i^2) < \infty$. The estimator

$$\hat{S}_n(t) = \left(\sum_{j=1}^n a_j \right)^{-1} \sum_{i=1}^n a_i U_i(t) \quad (2)$$

can be shown to be unbiased and nonincreasing in t , and to satisfy $0 \leq \hat{S}_n(t) \leq 1$. The random coefficient a_i in $\hat{S}_n(t)$ can be thought of as a weight given to U_i . This weight depends on the length of censoring time and thus has the potential to give more weight to those U_i 's with longer observation periods.

Next we consider the generalization of $\hat{S}_n(t)$ to censored recurrence times $\{t_{i1}, t_{i2}, \dots, t_{i,m_i-1}, t_{i,m_i}^+\}$. Define the observed recurrence times as

$$y_{ij} = \begin{cases} t_{ij} & \text{if } j=1, \dots, m_i-1 \\ t_{i,m_i}^+ & \text{if } j=m_i \end{cases}$$

The total mass of the risk set at t is calculated as

$$R^*(t) = \sum_{i=1}^n \left[\frac{a_i}{m_i^*} \sum_{j=1}^{m_i^*} I(y_{ij} \geq t) \right],$$

and the mass evaluated at t is

$$d^*(t) = \sum_{i=1}^n \left[\frac{a_i I(m_i \geq 2)}{m_i^*} \sum_{j=1}^{m_i^*} I(y_{ij} = t) \right].$$

Let $y_1^*, y_2^*, \dots, y_K^*$ be the ordered and distinct uncensored times. The estimator in (2) can be generalized to censored data using the techniques of hazard estimation. The generalized estimator takes the product-limit expression,

$$\hat{S}_n(t) = \prod_{y_i^* \leq t} \left\{ 1 - \frac{d^*(y_i^*)}{R^*(y_i^*)} \right\}, \quad (3)$$

which is nonincreasing in t , satisfies $0 \leq \hat{S}_n(t) \leq 1$, and reduces to (2) if the estimator is constructed based on uncensored data.

Define the functions $H_a(t) = E[a_i I(T_{i1} \leq t) I(C_i \geq t)]$ and $F_a(t) = E[a_i I(T_{i1} \leq t) I(T_{i1} < C_i)]$. When S is absolutely continuous, construction of $\hat{S}_n$ can also be motivated by the unbiased estimation of H_a and F_a using a representation of the cumulative hazard function,

$$\Lambda(t) = \int_0^t \frac{E[a_i I(C_i \geq u)] d\{1 - S(u)\}}{E[a_i I(C_i \geq u)] S(u)} = \int_0^t \frac{dF_a(u)}{H_a(u)}. \quad (4)$$

The advantage of using an expression such as (4) has been proven useful in the development of asymptotic properties for a variety of univariate and multivariate models (Andersen, Borgan, Gill, and Keiding 1993; Breslow and Crowley 1974; Gill 1980). Define the estimators

$$\hat{H}_a(t) = n^{-1} R^*(t) = n^{-1} \sum_{i=1}^n \left[\frac{a_i}{m_i^*} \sum_{j=1}^{m_i^*} I(y_{ij} \geq t) \right]$$

and

$$\hat{F}_a(t) = n^{-1} \sum_{i=1}^n \left[\frac{a_i I(m_i \geq 2)}{m_i^*} \sum_{j=1}^{m_i^*} I(y_{ij} \leq t) \right].$$

It can be proven that $\hat{H}_a(t)$ and $\hat{F}_a(t)$ are unbiased estimators for estimating $H_a(t)$ and $F_a(t)$. Replacing $H_a(u)$ and $F_a(u)$ in (4) by the corresponding unbiased estimators, a nonparametric estimator of $\Lambda(t)$, $\hat{\Lambda}(t)$, can be derived. The estimator $\hat{S}_n(t)$ can then be derived using the relation $S(t) = \exp\{-\Lambda(t)\}$ and further approximation $\hat{S}_n(t) \approx \exp\{-\hat{\Lambda}(t)\}$. As a result of the approximation stated in Theorem 1, the estimator $\hat{S}(t)$ can be shown to be asymptotically normal under regularity conditions.

Let t^* be a nonnegative constant satisfying $t^* < \sup\{t : S(t)G(t) > 0\}$. Define

$$\varphi_i(t) = \int_0^t \{H_a(u)\}^{-2} \left\{ \frac{a_i}{m_i^*} \sum_{j=1}^{m_i^*} I(Y_{ij} \geq u) dF_a(u) - \frac{a_i I(m_i \geq 2)}{m_i^*} \sum_{j=1}^{m_i^*} \frac{1}{H_a(Y_{ij})} I(Y_{ij} < t) \right\} \quad (5)$$

and $(t_1, t_2) = S(t_1)S(t_2)E[\hat{\Lambda}(t_1)\hat{\Lambda}(t_2)]$.

Theorem 1. Assume that $a_i = a(C_i)$ is a bounded function. As $n \rightarrow \infty$, the random process $n^{1/2}\{\hat{S}(t) - S(t)\}$ for $0 \leq t \leq t^*$ converges weakly to a mean 0 Gaussian process W^0 with the variance-covariance function (t_1, t_2) .

Proof. See the Appendix

With (F_a, H_a) replaced by $(\hat{F}_a, \hat{H}_a)$ in (5), let $\hat{\Lambda}(t)$ be the corresponding estimator of $\Lambda(t)$. The covariance function (t_1, t_2) can be consistently estimated by $n^{-1} \hat{S}_n(t_1) \hat{S}_n(t_2) E[\hat{\Lambda}(t_1) \hat{\Lambda}(t_2)]$. A question of interest is how to select a_i so that the asymptotic variance of $\hat{S}(t)$ is minimized. With the nonlinear structure of $\hat{S}(t)$, a closed-form solution does not seem to exist, and, undesirably, the choice of the optimal a_i could vary for different values of t . Practically, as suggested by an associate editor, the optimal a_i could be obtained by minimizing $n^{-1} \sum_i \hat{\varphi}_i^2(t)$, an estimate of $E[\varphi_i^2(t)]$. The procedure can also be extended to minimizing, say, the integrated mean square error (IMSE) if the IMSE is used as the efficiency criterion for the overall estimation. In the latter case, the choice of a_i is independent of t .

Although the Kaplan-Meier estimator (Kaplan and Meier 1958) based on the pooled recurrence times is generally biased, the Kaplan-Meier estimator derived from the first (possibly censored) recurrence times is appropriate for estimating S . This estimator has an asymptotic variance-covariance function $\sigma_{KM}(t_1, t_2)$, which can be formulated by the same technique as in Theorem 1 using

$$\psi_i(t) = \int_0^t \{H(u)\}^{-2} I(Y_{i1} \geq u) dF(u) - \frac{I(m_i \geq 2)}{H(Y_{i1})} I(Y_{i1} < t)$$

instead of $\lambda(t)$, with $H(t) = E[I(T_{i1} \leq t)I(C_i \leq t)]$ and $F(t) = E[I(T_{i1} \leq t)I(T_{i1} \leq C_i)]$. With appropriate choices of a_i , the proposed estimator $\hat{S}_H(t)$ is expected to outperform the Kaplan-Meier estimator, because subject-specific averages of two indicator functions in $\lambda(t)$ are used in $\lambda(t)$. In particular, for the convenient and important choice $a_i = 1$ (or, equivalently, a_i is a positive-valued constant), the pointwise asymptotic variance of $\hat{S}_H(t)$ is always an improvement. For each $0 \leq t \leq t^*$,

$$\sigma_{KM}(t, t) \geq \sigma(t, t). \quad (6)$$

The proof of (6) is given in the Appendix. In general, (6) may or may not be valid, depending on the choice of a_i ; see the proof of (6) for details. The improvement of efficiency is expected to be significant when recurrence times are weakly correlated. The gain in efficiency is typically less significant when recurrence times are highly correlated. Thus, for example, in a women's menstrual cycle study where the cycle time is of interest, the gain in efficiency from the proposed estimators may not be significant due to the high correlation among cycle times.

3. Examples

Here we illustrate the methods developed in the previous sections by a simulation study and an analysis of the Denmark schizophrenia data.

3.1 Simulation

In a simulation study a set of frailty values $x_1, \dots, x_n$, $n = 200$, is generated by the gamma distribution with parameters (a, b) , where $E(\mathbf{X}) = ab$ and $\text{var}(\mathbf{X}) = ab^2$. Four sets of parameter values of (a, b) are chosen: $(a, b) = (1/.75, .75)$, $(1, 1)$, $(1/1.5, 1.5)$, and $(1/2, 2)$. The gamma frailties under the four models have the same mean, 1, but different variances, (.75, 1, 1.5, 2). Given x_i , the iid recurrence times $t_{i1}, t_{i2}, \dots$ are generated by a Weibull distribution with the survival function $\exp\{-x_i(t/2)^{1.5}\}$. The true survival function can thus be expressed as in (1), with f , $d\mu(u)$, and F_X specified by the Weibull density, the Lebesgue measure du , and the gamma distribution function. The observation of the recurrence processes is terminated by a fixed censoring time $C_i = 2$.

One thousand samples are generated for calculating different estimates of S . In particular, we compare three estimators: the proposed estimate (with $a_i = 1$) from censored data, the Kaplan-Meier estimate from the pooled censored data (termed the *Kaplan-Meier estimate I*) and the Kaplan-Meier estimate based on the pooled data but with the last censored recurrence times t_{i, m_i}^+ excluded when $m_i \geq 2$ (termed the *Kaplan-Meier estimate II*). The average of 1,000 replicates for each estimate is computed and presented in Figures 1. As shown in the figure, the bias from Kaplan-Meier estimates I and II appear nonignorable, and the average of the proposed estimates cannot be distinguished from the true curve until $t = 2$.

The three estimates are not defined for $t > 2$, because $C_i = 2$ for all subjects. It is interesting to see that the bias from each of the two Kaplan-Meier estimates increases when the frailty variance increases. In general, the bias from Kaplan-Meier estimate I results from the estimate's overweighting on small uncensored recurrence times, where each uncensored recurrence time receives a unit weight, and the overweighting on large values from the last recurrence times. In our simulation, the overweighting on small uncensored recurrence times plays a more significant role in the estimation. Also, in this simulation, Kaplan-Meier estimate I is less biased than Kaplan-Meier estimate II, because the former estimate uses *all* of the censored recurrence times (t_{i,m_i}^+) , and thus it distributes more weight to larger values.

What is the gain in efficiency of the proposed estimate over the Kaplan-Meier estimate derived from the first (possibly censored) recurrence times? To address this question, we simulated the same type of recurrence time data with gamma(1, 1.5) frailty and different censoring times at 2, 4, and 6. One thousand replicates of the estimate for each censoring model are computed for the derivation of the relative efficiencies. The averages of m_i for the three censoring models are 2.069, 3.361, and 4.671. The relative efficiencies depend on both the censoring time and the recurrence time t ; see Figure 2. As expected, the efficiency of the estimate increases when the censoring time becomes large.

It is also interesting to compare the uses of different choices of weight, a_i , in the estimation. A further simulation is conducted by using the same recurrence time distribution with gamma(1, 1.5) frailty and uniform(0, 4) and uniform(0, 8) as the censoring distributions. As shown in Figure 3, when C_i is distributed as uniform(0, 8) (average $m_i = 3.379$), the estimate with $a_i = c_i$ outperforms the estimate with $a_i = 1$ for most portions of t . When C_i is distributed as uniform(0, 4) (average $m_i = 2.086$), the relative efficiency comparison suggests that the estimate with $a_i = c_i$ might be slightly better at earlier times but does not perform as well at later times. The simulation result seems to suggest that giving the weight proportional to the observation time, c_i , could be a good choice, especially when the probability structure of C_i allows for the possibility of observing three or more recurrent events. In the situations when recurrence times within subject are highly correlated, however, the unit weight $a_i = 1$ could be still preferred, as multiple recurrence times from the same subject do not contribute much more information than a single recurrence time.

3.2 Schizophrenia Data Example

The second example illustrates a recurrence time analysis from a subsample of a Denmark schizophrenia registry data. Systematic registration of patients with a diagnosis of schizophrenia was computerized in 1970 to include all cases from 86 psychiatric institutions in Denmark (Eaton et al. 1992a). The original anonymous data file recorded recurrences of hospitalizations and other associated information from patients who were admitted to the hospital due to schizophrenic symptoms for the first time in their lives during the period April 1, 1970–March 25, 1988. The data used in our analysis is a subsample of the original data and comprises 467 individuals admitted to the hospital for the first time during the 2-year period April 1, 1970–March 31, 1972. The recurrence and censoring times are measured in days. To avoid potential long-term pattern changes of recurrence times (Eaton et al. 1992b), the maximum follow-up time for each patient is set to be 3 years, although the original longitudinal data file provides information through March 25, 1988.

The data are categorized into two groups using the onset age of schizophrenia: onset at age 20 years or younger (113 individuals) and onset older than 20 years (354 individuals). For the group with onset age ≤ 20 , the number of recurrence times (m_i) ranges from 1 to 12. The numbers of patients with $m_i = 1, 2, 3, 4, 5, \dots, 8, 9, \dots, 12$ are 27, 25, 20, 15, 21, and 5. For the

group with onset age > 20 , the number of recurrence times (m_i) ranges from 1 to 21. The numbers of patients with $m_i = 1, 2, 3, 4, 5, \dots, 8, 9, \dots, 21$ are 169, 69, 52, 26, 28, and 10.

The estimate (3) is calculated for each group. The weight function is chosen to be $a_i = 1$, because the censoring times remain a constant (3×365 days) for the majority of the patients except for 15. A 95% pointwise confidence band for the recurrent survival function is calculated by the asymptotic variance estimate described in Section 2. These analytic results are presented in Figure 4. By contrasting the two survival function estimates, it is clearly shown that the onset age of schizophrenia plays an important role in determining the length of time between hospitalizations.

As an informal way of checking the validity of $S = S_k$ where S_k represents the marginal survival function for the k th recurrence time, $k = 1, 2, \dots$, we use different maximum censoring times at 1, 1.5, 2, 2.5, and 3 years to construct the nonparametric estimates. The resulting estimates are found to be close to each other in the estimable areas. We thus feel that the hypothesis for $S = S_k$ is approximately true at an early stage of the disease.

4. Concluding Remarks

We noted in Section 1 that assumption A1 states a condition that is stronger than necessary in the development of the proposed estimators. The existence of the frailty $\mathbf{X}_i$ in A1 is assumed to ensure that the iid structure of recurrence times within subject, but the distribution of such a quantitative factor is actually not used in the methodology. The proposed estimators and related properties would still hold even if $\mathbf{X}_i$ were an abstract variate with the probability measure $P_{\mathbf{X}}$. In this case, the marginal survival function would be defined as $S(t) = \int S(t|x_j) dP_{\mathbf{X}}(x_j)$.

In statistical literature, an abundance of work has been developed to analyze clustered survival data where the cluster size is independent of outcome measures, such as data collected from twins or families. Relatively little work has been developed that is appropriate when recurrence times are considered as the outcome variables. In particular, it is generally unclear how to estimate an average length of interevent times over the population. The marginal survival function of recurrence times is frequently used as a useful tool for evaluating the associated risks when the risk factors are categorical. The methods proposed in this article can be used to address questions of this type.

Acknowledgments

This research was supported by National Institute of Health grants R01 DA10184 and R01 MH56639. The authors thank Preben Bo Mortensen of the Danish Psychiatric Case Register and William W. Eaton at Johns Hopkins University for providing anonymous schizophrenia data.

Appendix: Proofs

Proof of Theorem 1

With correlated survival data, standard martingale theory (Fleming and Harrington 1991) may not be applicable, and thus we use an approach based on empirical processes. First, write

$$\begin{aligned}
& \hat{\Lambda}(t) - \Lambda(t) \\
&= \int_0^t \left\{ \frac{d\hat{F}_a(u)}{\hat{H}_a(u)} - \frac{dF_a(u)}{H_a(u)} \right\} = \int_0^t \left\{ \frac{1}{\hat{H}_a(u)} \right. \\
&\quad \left. - \frac{1}{H_a(u)} \right\} dF_a(u) \\
&+ \int_0^t \frac{1}{H_a(u)} d\left\{ \hat{F}_a(u) \right. \\
&\quad \left. - F_a(u) \right\} + \int_0^t \left\{ \frac{1}{\hat{H}_a(u)} \right. \\
&\quad \left. - \frac{1}{H_a(u)} \right\} d\left\{ \hat{F}_a(u) - F_a(u) \right\}.
\end{aligned} \tag{A.1}$$

Both $\hat{H}_a$ and F_a are sample mean-type estimators. The third term in (A.1) can be shown to be of order $o_p(n^{-1/2})$ and thus is asymptotically negligible. Approximating the first term using techniques similar to those of Breslow and Crowley (1974), one has

$$\begin{aligned}
& \hat{\Lambda}(t) - \Lambda(t) \\
&= \int_0^t \{H_a(u)\}^{-2} \{H_a(u) - \hat{H}_a(u)\} dF_a(u) \\
&+ \int_0^t \{H_a(u)\}^{-1} d\{\hat{F}_a(u) \\
&\quad - F_a(u)\} + O_p(n^{-1/2}) = - \int_0^t \{H_a(u)\}^{-2} \hat{H}_a(u) dF_a(u) + \int_0^t \{H_a(u)\}^{-1} d\{\hat{F}_a(u) \\
&\quad + O_p(n^{-1/2})\} \\
&= - \frac{1}{n} \sum_{i=1}^n \varphi_i(t) \\
&+ O_p(n^{-1/2}),
\end{aligned}$$

Where

$$\varphi_i(t) = \int_0^t \{H_a(u)\}^{-2} \left\{ \frac{a_i}{m_i^*} \sum_{j=1}^{m_i^*} I(Y_{ij} \geq u) \right\} dF_a(u) - \frac{a_i I(m_i \geq 2)}{m_i^*} \sum_{j=1}^{m_i^*} \frac{I(Y_{ij} < t)}{H_a(Y_{ij})}.$$

It can be proven that $E[\varphi_i(t)] = 0$. By applying the delta method to $\exp(-\hat{\Lambda}(t))$ and by further approximation using the formula $-\ln(1 - y^{-1}) - y^{-1} < y^{-1}(y - 1)^{-1}$ for $y > 1$, an iid representation of $\hat{S}_n - S$ can then be derived as

$$\hat{S}_n(t) - S(t) = n^{-1} \sum_{i=j}^n \{S(t)\varphi_i(t)\} + o_p(n^{-1/2}).$$

When a_j is bounded by a constant, the random variable $S(t) - \hat{S}(t)$ is uniformly bounded in $0 \leq t \leq t^*$. Thus, by the multivariate central limit theorem, the finite-dimensional distributions of the random process $n^{1/2}\{\hat{S}_n(\cdot) - S(\cdot)\}$ converge weakly to those of W^0 . Further, using the techniques of (13.17) of Billingsley (1968), we derive

$$E[\{\hat{H}_a(t) - H_a(t_1)\}\{\hat{H}_a(t_2) - H_a(t_1)\}] \leq \text{constant} \cdot (t - t_1)(t_2 - t_1)$$

and

$$E[\{\hat{F}_a(t) - F_a(t_1)\}\{\hat{F}_a(t_2) - F_a(t_1)\}] \leq \text{constant} \cdot (t - t_1)(t_2 - t_1)$$

for $t_1 \leq t \leq t_2$. By theorem 15.6 of Billingsley (1968) and arguments similar to those of Breslow and Crowley (1974), the tightness of the sequence $n^{1/2}\{\hat{S}_n(\cdot) - S(\cdot)\}$ follows. Hence we prove the result of Theorem 1.

Proof of (6)

Define

$$\varphi_{ij}(t) = a_i \left\{ \int_0^t \{H_a(u)\}^{-2} I(Y_{ij} \geq u) dF_a(u) - I(m_i \geq 2) \frac{I(Y_{ij} < t)}{H_a(Y_{ij})} \right\}.$$

Then

$$\varphi_i(t) = \frac{1}{m_i^*} \sum_{j=1}^{m_i^*} \varphi_{ij}(t).$$

Conditional on (m_i, x_i, C_i) , for each $0 \leq t \leq t^*$, $(\varphi_{i1}(t), \dots, \varphi_{i, m_i^*}(t))$ are identically distributed random variables. Thus, by the Minkowski's inequality, we derive

$$E \left[\left\{ \sum_{j=1}^{m_i^*} \varphi_{ij}(t) \right\}^2 \mid m_i, x_i, c_i \right] \leq (m_i^*)^{-2} E[\varphi_{i1}^2 \mid m_i, x_i, c_i].$$

This implies

$$E[\{\varphi_i(t)\}^2] = E \left[E \left[\left\{ \frac{1}{m_i^*} \sum_{j=1}^{m_i^*} \varphi_{ij}(t) \right\}^2 \mid m_i, X_i, C_i \right] \right] \leq E[E[\varphi_{i1}^2(t) \mid m_i, X_i, C_i]] = E[\varphi_{i1}^2(t)],$$

If $a_j = 1$, then one has $\hat{S}_n(t) = \hat{S}(t)$, and thus

$$\sigma(t, t) = S^2(t)E[\varphi_i^2(t)] \leq S^2(t)E[\psi_i^2(t)] = \sigma_{KM}(t, t)$$

for each $0 \leq t \leq t^*$, and hence (6). Note that, more generally, (6) is valid for other choices of a_i if $E[\varphi_{i1}^2(t)] \leq E[(\psi_i(t))^2]$.

References

- Aalen OO, Husebye E. Statistical Analysis of Repeated Events Forming Renewal Processes. *Statistics in Medicine*. 1991; 10:1227–1240. [PubMed: 1925154]
- Andersen, PK.; Borgan, O.; Gill, RD.; Keiding, N. *Statistical Models Based on Counting Processes*. New York: Springer-Verlag; 1993.
- Andersen PK, Gill RD. Cox's Regression Model for Counting Processes: A Large Sample Study. *The Annals of Statistics*. 1982; 10:1100–1120.
- Billingsley, P. *Convergence of Probability Measures*. New York: Wiley; 1968.
- Breslow NE, Crowley J. A Large Sample Study of the Life Table and Product Limit Estimates Under Random Censorship. *The Annals of Statistics*. 1974; 2:437–453.
- Cochran, WG. *Sampling Techniques*. New York: Wiley; 1977.
- Eaton WW, Mortensen PB, Herrman H, Freeman H, Bilker W, Burgess P, Wooff K. Long-Term Course of Hospitalization for Schizophrenia: Part I. Risk for Hospitalization. *Schizophrenia Bulletin*. 1992a; 18:217–228. [PubMed: 1621069]
- Eaton WW, Bilker W, Haro JM, Hermann H, Mortensen PB, Freeman H, Burgess P. Long-Term Course of Hospitalization for Schizophrenia: Part II. Change With Passage of Time. *Schizophrenia Bulletin*. 1992b; 18:229–241. [PubMed: 1621070]
- Fleming, TR.; Harrington, DP. *Counting Processes and Survival Analysis*. New York: Wiley; 1991.
- Gill, RD. *Censoring and Stochastic Integrals*, Mathematical Centre Tracts 124. Amsterdam: Mathematisch Centrum; 1980.
- Kaplan EL, Meier P. Nonparametric Estimation From Incomplete Observations. *Journal of the American Statistical Association*. 1958; 53:457–481.
- Lawless JF, Nadeau C. Some Simple Robust Methods for the Analysis of Recurrent Events. *Technometrics*. 1995; 37:158–168.
- Pepe MS, Cai J. Some Graphical Displays and Marginal Regression Analyses for Recurrent Failure Times and Time-Dependent Covariates. *Journal of the American Statistical Association*. 1993; 88:811–820.
- Prentice RL, Williams BJ, Peterson AV. On the Regression Analysis of Multivariate Failure Time Data. *Biometrika*. 1981; 68:373–379.

Figure 1. Survival Function Estimates From Simulated Data With Gamma(a, b) Frailty, (a) Frailty = gamma(1/.75, .75); (b) frailty = gamma(1, 1); (c) frailty = gamma(1/1.5, 1.5); (d) frailty = gamma(1/2, 2).—, true curve; (x025D6) (x025D6) (x025D6), proposed curve;— — —, Kaplan-Meier estimate I; &22EF;, Kaplan-Meier estimate II.

Figure 2. Relative Efficiency Plot. —, censoring time = 2; ---, censoring time = 4; ---, censoring time = 6.

Figure 3. Relative Efficiency Plot. —, $a = 1$ and C_1 is Uniform(0, 4); ---, $a = c$ and C_1 is Uniform(0, 4); ···, $a = 1$ and C_1 is Uniform(0, 8); - · -, $a = c$ and C_1 Uniform(0, 8).

Figure 4. Recurrent Survival Function Estimates From Schizophrenic Data. —, estimate for onset age ≤ 20 years; ---, 95% confidence interval; - - -, estimate for onset age > 20 years; - - -, 95% confidence interval.