Statistica Sinica **16**(2006), 287-302

SEMIPARAMETRIC ADDITIVE RISKS MODEL FOR INTERVAL-CENSORED DATA

Donglin Zeng[1], Jianwen Cai[1] and Yu Shen[2]

[1] *University of North Carolina at Chapel Hill and* [2] *The University of Texas*

Abstract: Interval-censored event time data often arise in medical and public health studies. In such a setting, the exact time of the event of interest cannot be observed and is only known to fall between two monitoring times. Our interest focuses on the estimation of the effect of risk factors on interval-censored data under the semiparametric additive hazards model. A nonparametric step-function is used to characterize the baseline hazard function. The covariate coefficients are estimated by maximizing the observed likelihood function, and their variances are obtained using the profile likelihood approach. We show that the proposed estimates are consistent and have asymptotic normal distributions. We also show that the estimator obtained for the covariate coefficient is the most efficient estimator. Simulation studies are conducted to assess the performance of the estimate. The method is illustrated through application to a data set from an HIV study.

Key words and phrases: Additive hazards regression, interval-censored data, nonparametric maximum likelihood estimates, profile likelihood, semiparametric efficiency.

1. Introduction

Incomplete follow-up data are often encountered in medical and public health studies. In particular, the exact onset of the event of interest cannot be observed directly, and is only known to fall in some interval. For example, although the seroconversion time of an HIV patient is unlikely to be observed directly, it is feasible to periodically monitor the patient's status to determine the time interval in which seroconversion may have occurred. A second example is the patient's age at onset of preclinical breast cancer; current technologies are unable to observe the exact time of tumor onset, but cancer screening histories can yield an interval of time during which it may have first occurred.

This type of survival data, consisting of time intervals monitored to assess the onset of the event of interest when the actual event cannot be observed, is called interval-censored data. When only one monitoring time is applied and each patient is known to experience the onset of the event either before or after this monitoring time, the data are called current status data, or interval-censored data, case I. When more than one monitoring time is applied and each patient is known to experience the onset of the event (also known as "failure") either

before the first monitoring time, between the two monitoring times, or after the last monitoring time, such data are called interval-censored data, case II.

Although various statistical methods have been proposed to study the effects of covariates for current status data, e.g., Huang (1996), Satten, Datta and Williamson (1998), Rossini and Tsiatis (1996), Lin, Oakes and Ying (1998) and Ghosh (2001), studies of interval-censored data, case II in the literature have been relatively limited. Among those available, Finkelstein (1986) studied the proportional hazards model and Rabinowitz and Tsiatis (1995) considered the accelerated failure time models. Some investigators have also looked into the nonparametric test with interval-censored data, case II (Sun (1996) and Zhang, Liu and Zhan (2001)).

For survival data, the additive and multiplicative risk models provide the two principal frameworks for studying the association between risk factors and event time, although the choice of these two models is more an empirical issue. Compared with the multiplicative model (Cox (1972)), the additive risk model is particularly useful for estimating the difference in hazards with the following form: for a k-vector of possibly time-varying covariates $z(\cdot)$, the hazard rate function of the event time is

$$h(t|z(s), 0 \le s \le t) = \lambda(t) + \beta^T z(t), \tag{1}$$

where $\lambda(t)$ is the baseline hazard function and β is a k-vector regression coefficient for covariate $z(t)$. Other forms of the additive risk model besides (1) have been eloquently advocated and successfully utilized for right-censored survival data by numerous authors, e.g., Andersen, Borgan, Gill, and Keiding (1993, pp.563-566), Lin and Ying (1994), McKeague and Sasieni (1994), and Shen and Cheng (1999). Furthermore, Lin, Oakes and Ying (1998) and Ghosh (2001) have considered model (1) for current status data. However, such a model has not been well studied for general interval-censored data; part of the reason is that the martingale-based estimation, which was successfully used by Lin et al. (1998) for current status data, cannot be generalized to interval-censored data, case II.

We consider estimation for a semiparametric additive hazards model with interval-censored data, case II. In the next section, we propose an efficient estimator of the covariate effects by using the maximum likelihood estimation approach. Furthermore, we obtain an estimator for the cumulative hazard function under the additive risk model. In Section 4, we provide the results from simulation studies and from the application of our model to a data example.

2. Maximum Likelihood Estimation

Typical observations of interval-censored data include a pair of non-negative random variables U and V, where U is called the left examination time and V the

right examination time. We assume that the examination times are independent of the event time, T, and (U, V) are random variables from a distribution with support

$$\mathcal{X} = \{(u, v) : 0 < \tau_u \leq u, v \leq \tau_v < \infty, v \geq u + \xi\},$$

where ξ is a positive constant. Note that the support for this bivariate distribution requires that the right examination time is not the same as the left examination time, a reasonable assumption when the exact failure times cannot be observed in many biomedical studies.

The observation regarding the true event time T falls into one of the following three exclusive categories: T is between U and V (interval-censored); T is larger than V (right-censored); or T is less than U (left-censored). Let $z(t)$ denote the covariate information at time t. We define two indicator variables as $\delta_1 = I(T \leq U)$ and $\delta_2 = I(U < T \leq V)$, then data from the ith subject can be expressed as

$$\begin{aligned}
&\{\delta_{1i} = 1, U_i, z_i(t), t \in [0, \tau_v]\}, && \text{if subject } i \text{ is left-censored,}\\
&\{\delta_{2i} = 1, U_i, V_i, z_i(t), t \in [0, \tau_v]\}, && \text{if subject } i \text{ is interval-censored,}\\
&\{\delta_{1i} = \delta_{2i} = 0, V_i, z_i(t), t \in [0, \tau_v]\}, && \text{if subject } i \text{ is right-censored.}
\end{aligned}$$

For n i.i.d. subjects, the data can be equivalently summarized as

$$\Big\{\delta_{1i}, \delta_{2i}, (\delta_{1i} + \delta_{2i})U_i, (1 - \delta_{1i})V_i, z_i(t), t \in [\tau_u, \tau_v], i = 1, \ldots, n\Big\}.$$

The additive risk model (1) can be also expressed in terms of the cumulative hazard function, $H(t|Z(s), 0 \leq s \leq t) = \Lambda(t) + \beta^T Z(t)$, where $\Lambda(t)$ is the baseline cumulative hazard function, and $Z(t) = \int_0^t z(s)ds$. Let $G(t) = e^{-\Lambda(t)}$ define the baseline survival function. The observed full likelihood can be expressed, by forming the product over n i.i.d. subjects, as

$$\begin{aligned}
\prod_{i=1}^{n} &[1 - G(U_i)e^{-\beta^T Z_i(U_i)}]^{\delta_{1i}}[G(U_i)e^{-\beta^T Z_i(U_i)} - G(V_i)e^{-\beta^T Z_i(V_i)}]^{\delta_{2i}}\\
&\times [G(V_i)e^{-\beta^T Z_i(V_i)}]^{1-\delta_{1i}-\delta_{2i}}.
\end{aligned}$$

Let $l_n(\beta, G)$ denote the logarithm of the observed likelihood function in terms of the parameters, β and $G(\cdot)$. The maximum likelihood estimates for (β, G) can be obtained by maximizing the observed log-likelihood function $l_n(\beta, G)$ over the parameter space,

$$\begin{aligned}
\Theta = \{(\beta, G) : &\beta \text{ is in a compact set } K \text{ of } R^k,\ \|\beta\| \leq B,\\
&G(t) \text{ is a non-increasing function with } G(0) = 1,\ G(t) > 0\}.
\end{aligned}$$

Since the observed likelihood function is clearly bounded by 1 and Θ is weakly compact, the maximum likelihood estimate for (β, G) exists. We can specifically choose the nonparametric estimate for G to be a non-increasing step-function with jumps only at the observed examination times.

Computationally, the above optimization can be carried out as follows. Let $y_{(1)} > \cdots > y_{(m)}$ denote the unique examination times from the largest to the smallest, where $m \leq n + \sum_{i=1}^{n} \delta_{2i}$. The maximum likelihood estimate can be derived by maximizing

$$l_n(\beta, G) = \sum_{i=1}^{n} \Big\{ \delta_{1i} \ln[1 - G(U_i)e^{-\beta^T Z_i(U_i)}] + \delta_{2i} \ln[G(U_i)e^{-\beta^T Z_i(U_i)} - G(V_i)e^{-\beta^T Z_i(V_i)}] + (1 - \delta_{1i} - \delta_{2i}) \ln[G(V_i)e^{-\beta^T Z_i(V_i)}] \Big\}$$

under the constraint that $0 \leq x_{(1)} < \cdots \leq x_{(m)}$, where $G(y_{(j)}) = x_{(j)}$ for $j = 1, \ldots, m$. To ensure the positivity and the monotonicity of $\{x_{(j)}, j = 1, \ldots, m\}$, we use the following transformed parameters in the optimization: $w_1 = \log x_{(1)}$, $w_j = \log(x_{(j)} - x_{(j-1)})$, $j = 2, \ldots, m$. The gradient and the Hessian matrix of the log-likelihood function with respect to these parameters can be evaluated and utilized in calculating the maximum likelihood estimates. When m is not large, Newton-Raphson iteration can be used to solve the score equations for β and $\{w_i, i = 1, \ldots, m\}$. When m is large, the Nelder-Mead simplex method is used to search for the optimum. Particularly, at each step of the search, a new point in or near the current simplex is generated and the function value at the new point is compared with the function's values at the vertices of the simplex. The algorithm is run until the diameter of the simplex is less than the specified tolerance. Such an algorithm is implicated in a built-in function "fminsearch" or "fminunc" in MATLAB software.

3. Asymptotic Properties

With the described model and method of estimation, we can derive large sample properties for the proposed estimators under the following assumptions.
(A.1) The true value for β, denoted as β_0, is in the interior of a compact set K, $\|\beta_0\| \leq B$ for a constant $B > 0$, and $\sup_{t \in [\tau_u, \tau_v]} \|z(t)\| \leq M/(\tau_v - \tau_u), a.s.$, for a given constant $M > 0$.
(A.2) The interval censoring times (U, V) have a positive bivariate density, $f_{U,V}(u, v)$, in the support $\mathcal{X}$ and the density has a bounded second order derivative.
(A.3) The underlying parameter function (β_0, G_0) satisfies $-[dG_0(t)/dt][1/G_0(t)] + \beta_0^T z(t) > 0$, a.s., where G_0 has bounded second order derivatives in $[\tau_u, \tau_v]$.

(A.4) If there exist a constant c_0 and a k-vector γ such that $\gamma^T z(t) = c_0$ for any $t \in [\tau_u, \tau_v]$ with probability 1, then $\gamma \equiv 0$ and $c_0 \equiv 0$.
(A.5) $P(T < \tau_u | z(t), t > 0)$ and $P(T > \tau_v | z(t), t > 0)$ have a positive lower boundary with probability 1.

Assumption (A.3) ensures that the underlying hazard rate function is a positive function; (A.4) is simply equivalent to the identification condition in a linear model; (A.5) stipulates that event times may occur outside of the support of the censoring times. Under the above conditions, we can obtain asymptotic properties of the proposed maximum likelihood estimator $(\hat{\beta}_n, \hat{G}_n)$.

Theorem 1. *Under* (A.1)–(A.5), $\|\hat{\beta}_n - \beta_0\| \to 0$ *and* $\sup_{t\in(\tau_u,\tau_v)} |\hat{G}_n(t) - G_0(t)| \to 0$ *with probability one. Furthermore,* $\sqrt{n}(\hat{\beta}_n - \beta_0)$ *has an asymptotic normal distribution with mean zero and a variance that attains the semiparametric efficiency bound for* β_0.

For a definition of the semiparametric efficiency bound, see Chapter 3 of Bickel, Klaassen, Ritov and Wellner (1993). We sketch a proof for Theorem 1 in the appendix.

From Theorem 1, we conclude that $\hat{\beta}_n$ is the most efficient estimator for β_0. We wish to estimate the asymptotic covariance for $\sqrt{n}(\hat{\beta}_n - \beta_0)$, denote by Σ. In the appendix, we find $\Sigma = E[l^*_\beta l^{*T}_\beta]^{-1}$, where l^*_β is the efficient score function for β_0. Particularly, $l^*_\beta = l_\beta - l_G[g^*]$, where l_β is the score for β_0 and $l_G[g^*]$ is the score function for G_0 along the direction of g^* given by

$$l_G[g^*] = \frac{-\delta_1 g^*(U) e^{-\beta_0{}^T Z(U)}}{1 - G_0(U) e^{-\beta_0{}^T Z(U)}} + \frac{\delta_2 (g^*(U) e^{-\beta_0{}^T Z(U)} - g^*(V) e^{-\beta_0{}^T Z(V)})}{G_0(U) e^{-\beta_0{}^T Z(U)} - G_0(V) e^{-\beta_0{}^T Z(V)}} + \frac{(1 - \delta_1 - \delta_2) g^*(V)}{G_0(V)}.$$

Furthermore, if $S_Z(t) = G_0(t) \exp\{-\beta^T Z(t)\}$, then g^* is the unique solution to the integral equation

$$-a(x)g(x) + \int_y b(x,y) g(y) dy = c(x), \tag{2}$$

where

$$a(x) = f_U(x) E\Big[\Big(\frac{1}{1 - S_Z(U)} + \frac{1}{S_Z(U) - S_Z(V)}\Big) e^{-2\beta_0{}^T Z(U)} | U = x\Big] + f_V(x) E\Big[\Big(\frac{1}{S_Z(U) - S_Z(V)} + \frac{1}{S_Z(V)}\Big) e^{-2\beta_0{}^T Z(V)} | V = x\Big],$$

$$b(x,y) = f_{U,V}(x,y)E\Big[\frac{e^{-\beta_0{}^T(Z(U)+Z(V))}}{S_Z(U)-S_Z(V)}|U=x,V=y\Big]$$
$$+f_{U,V}(y,x)E\Big[\frac{e^{-\beta_0{}^T(Z(U)+Z(V))}}{S_Z(U)-S_Z(V)}|U=y,V=x\Big],$$
$$c(x) = f_U(x)E\Big[\Big(\frac{S_Z(U)Z(U)}{1-S_Z(U)}+\frac{S_Z(U)Z(U)-S_Z(V)Z(V)}{S_Z(U)-S_Z(V)}\Big)e^{-\beta_0{}^TZ(U)}|U=x\Big]$$
$$+f_V(x)E\Big[\Big(Z(V)-\frac{S_Z(U)Z(U)-S_Z(V)Z(V)}{S_Z(U)-S_Z(V)}\Big)e^{-\beta_0{}^TZ(V)}|V=x\Big],$$

and f_U and f_V are the marginal densities for U and V, respectively. Equation (2) normally does not have an explicit solution, and a numerical solution may be complicated. Instead, we propose to estimate the asymptotic variance based on a difference involving the profile log-likelihood function, defined as

$$pl_n(\beta) = \frac{1}{n}\max_{G\in\Theta} l_n(\beta, G).$$

Let $\vec{e}_s$ be the vector in R^k with 1 at the sth position and 0 elsewhere. The asymptotic variance matrix of $\sqrt{n}(\hat{\beta}_n - \beta_0)$, Σ, is approximated by using

$$-\frac{pl_n(\hat{\beta}_n + h_n\vec{e}_s - h_n\vec{e}_l) - pl_n(\hat{\beta}_n + h_n\vec{e}_s) - pl_n(\hat{\beta}_n - h_n\vec{e}_l) + pl_n(\hat{\beta}_n)}{h_n^2} \approx \Sigma_{sl},$$

where Σ_{sl} is the (s,l)th element of Σ and h_n is a constant with order of $n^{-1/2}$. The theoretical justification of the profile likelihood estimation can be found in Murphy and van der Vaart (2000).

3. Numerical Studies and An Example

To assess the behavior of the proposed method with moderate sample sizes, we perform two simulation studies. Assume there are two independent covariates, $Z_1 \sim$ Bernoulli$(0.5, 0, 1)$ and $Z_2 \sim$ Uniform$(0, 1)$. In the first simulation study, conditional on these covariates, the hazard function for the underlying failure time T is additive to the baseline hazard, given by $0.2+\beta_{01}Z_1+\beta_{02}Z_2$ where $\beta_{01} = 0.5$ and $\beta_{02} = 0.2$. In the second simulation study, we allow a time-dependent covariate in the model, where the hazard function is $0.1+\beta_{01}Z_1+\beta_{02}Z_2t$, $\beta_{01} = 0.5$ and $\beta_{02} = 0.2$. For both studies, the left censoring time U and the right censoring time V are uniformly generated from the region $\{(u,v) : 0.1 \le u \le 2, u + 0.5 \le v \le 4\}$. With the above specifications, the proportion of left censoring, interval censoring, and right censoring is about 40%, 32%, and 28%, respectively.

For each choice of sample size $n = 100$ or $n = 200$, we use the optimization algorithm in MatLab 6.01 to calculate the maximum likelihood estimates of β_1 and

β_2. In the calculation, when the initial values are chosen to be close to the true values, the optimum search usually converges within 20 iterations. The profile likelihood approach, described in Section 3, is used to estimate their variances. To assess the robustness of the variance estimates to the choice of the oscillation parameter h_n in the profile likelihood approach, we let h_n be $\{2/(25n)\}^{-1/2}$ or $\{8/(25n)\}^{-1/2}$, where the scale 1/25 was chosen arbitrarily.

Table 1 summarizes the simulation results from 500 repetitions. Column "est" is the average of the estimates for β_{01} and β_{02} from 500 repetitions; column "est. se." is the mean of the estimated standard errors using the profile likelihood approach; column "emp. se." is the empirical standard deviation of the estimates; and column "cp" gives the coverage proportion of the 95% confidence intervals. From Table 1, we see that the maximum likelihood estimates perform reasonably well with moderate sample sizes: the bias of the estimates is small and the estimated standard errors based on the profile likelihood functions agree reasonably well with the empirical standard errors, especially for the time-independent covariate and the larger sample size. The 95% confidence intervals provide adequate coverage probabilities. The simulation studies also indicate that, compared with estimation with time-independent covariates, a larger sample is needed to ensure the correct inference for the cases with time-dependent covariates.

Table 1. Result from simulation studies with 500 repetitions.

					$h_n = 0.2\sqrt{2}n^{-\frac{1}{2}}$		$h_n = 0.4\sqrt{2}n^{-\frac{1}{2}}$	
n	par.	true value	est.	emp. se	est. se.	cp	est. se.	cp
			Study I: $\lambda(t\|Z) = 0.2 + 0.5Z_1 + 0.2Z_2$					
100	β_{01}	0.5	0.517	0.126	0.138	0.944	0.144	0.958
	β_{02}	0.2	0.183	0.125	0.123	0.920	0.124	0.926
200	β_{01}	0.5	0.504	0.092	0.095	0.946	0.099	0.954
	β_{02}	0.2	0.196	0.080	0.085	0.942	0.088	0.942
			Study II: $\lambda(t\|Z) = 0.1 + 0.5Z_1 + 0.2Z_2t$					
100	β_{01}	0.5	0.499	0.096	0.124	0.972	0.127	0.974
	β_{02}	0.2	0.207	0.058	0.076	0.974	0.077	0.976
200	β_{01}	0.5	0.497	0.073	0.086	0.966	0.087	0.954
	β_{02}	0.2	0.205	0.043	0.052	0.946	0.053	0.972

We illustrate the proposed methods through application to an HIV study recently conducted in the state of North Carolina. This study involved the recruitment of 183 HIV patients, all with CD4 cell counts less than $100/mm^3$ and serologic evidence of previous cytomegalovirus(CMV) infection. CMV infection is one of the most feared complications of HIV and may cause a persistent

sight-threatening retinitis. In an immunocompetent host, infection with CMV is followed by production of antibodies and an asymptomatic latent phase during which the virus remains quiescent in leukocytes. However, in an immunocompromised host, CMV replication can be activated and CMV dissemination and pathologic invasion of the retina can occur. Thus, it is important to know when such CMV activation occurs. To detect the activation of viremia from CMV during each patient visit, blood samples were tested with four different assays: CMV DNA polymerase chain reaction, hybrid capture CMV DNA, CMA antigen assay, and nucleic acid sequence based amplification of CMV pp67 mRNA. The test results were assumed to be accurate and sensitive in detecting active viremia from CMV. In our analysis, we focused on studying the relationship between the time of the first activation of viremia and the covariates, such as patient's sex, age, CD4 cell count, and HIV viral load measured at the baseline. Clearly, the exact time of the first active CMV infection could not be observed. However, if a patient tested positive for infection by at least one of these four assays in a visit, we concluded that the activation had occurred before that visit. In general, for each patient, his/her time to the first activation of the CMV infection was left-censored if he/she tested positive at the first visit after enrollment, right-censored if none of the test results were positive for infection during the follow-up, or the event time was interval-censored.

After excluding the individuals whose data had missing covariates (67 patients) and those who had developed the disease before their entry into the study (3 patients), there were 113 patients in the study. Of these, 10 patients were left-censored, 17 patients were interval-censored, and the remaining 86 patients were right-censored. Among patients with left-censored event times, the average length of time from entry to the left censoring time was 141 days; and among patients with interval-censored event times, the average length from entry to the left monitoring time was 301 days. The average length from entry to the right monitoring time was 420 days; and among patients with right-censored event times, the average length from entry to the right censoring time was 472 days. The data regarding covariates of this cohort included the following: 85 patients were male and 28 patients were female; the average age was 39 years with a standard deviation of 7.7; the mean baseline CD4 cell count with the log-10-scale was 4.37 (standard deviation 1.43); and the average HIV viral load level at the log-10-scale was 10.18 (standard deviation 2.56).

We fitted the semiparametric additive hazards model to this dataset, and incorporated into the model the four covariates, including age, sex, baseline CD4 count at the log-10-scale, and the baseline HIV viral load at the log-10-scale. When estimating the variances via the profile likelihood function, the oscillation parameter h_n was chosen to be $n^{-1/2}/20$. The results are given in Table 2. We also used $h_n = n^{-1/2}/100$, with similar results (not shown here due to limited

space). As presented in Table 2, although the results suggested that a high HIV viral load and male sex tended to increase the risk of developing the active CMV infection, none of the covariates appeared to be statistically significant. The non-significance of the results may have been due to a relatively small sample size and the limited number of the events in this study.

Table 2. Application to the CMV disease study in HIV patients.

covariate	coefficient	standard error	Z-value
Age	-0.0083	0.1247	-0.0669
Sex (male=1, female=0)	0.2328	0.7304	0.3187
CD4 count at log-scale	-0.0102	0.6579	-0.0155
HIV viral load at log-scale	0.0464	0.3716	0.1248

We define S as the linear score function $\hat{\beta}_n^T Z$. The median value of S is 0.325. We plotted the predicted survival curves based on the mean values of the linear scores in the group with S larger or less than 0.325 versus the corresponding nonparametric estimator, using the self-consistency algorithm of Turnbull (1976). Figure 1 shows that the predicted curves agreed reasonably well with the nonparametric estimates.

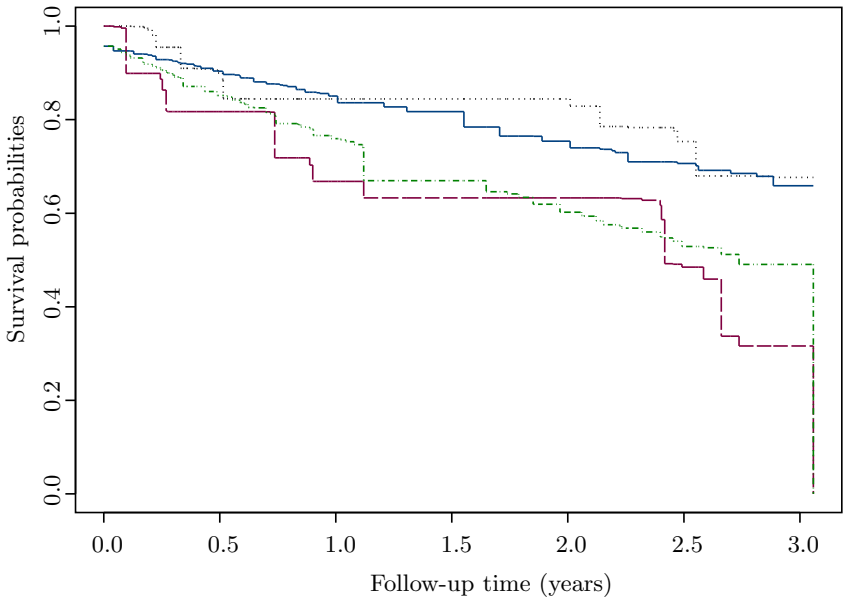

Figure 1. Predicted survival curves from CMV data (solid line: predicted curve for linear score less than 0.325; dotted line: nonparametric estimate for linear score less than 0.325; broken solid line: predicted curve for linear score larger than 0.325; dotted solid line: nonparametric estimate for linear score larger than 0.325).

5. Discussion

Although we have focused on interval-censored observations that correspond to only two examination times in this study, the methods can be easily generalized to multiple examination times if an ordered random examination time sequence is known for each subject, and if the failure time of interest is observed to belong to one of the partitioned intervals. As Huang (1996) points out, with multiple examination times, the effective observations for estimating the distribution of failure time given the covariates are the same as those with two examination times. These two examination times are the last examination time before the failure (even of interest) occurs, which can be 0, and the first examination time after the failure occurs, which can be infinity.

When one is interested in estimating the survival function for any given covariate process $Z(t)$, an intuitive estimate can be provided by $\hat{S}_n(t) = \hat{G}_n(t)e^{-Z(t)^T\hat{\beta}_n}$. However, when $Z(t)$ varies with time t, the estimated survival function might not be a monotone function. We suggest a modified estimator which is the maximal decreasing function below $\hat{S}_n(t)$. Denote this modified function by $\hat{S}_n^*(t)$. We can prove that $\hat{S}_n^*(t)$ is also a consistent estimator of the underlying survival function $S_0(t)$. According to the large sample result for $(\hat{\beta}_n, \hat{G}_n)$, we see that $\sup_{t\in(\tau_u,\tau_v)} |\hat{S}_n(t) - S_0(t)| = o_p(1)$. Therefore, by the definition of $\hat{S}_n^*(t)$, it holds that in probability, $S_0(t) - \epsilon \le \hat{S}_n^*(t) \le \hat{S}_n(t)$ for any small positive number ϵ. Consistency thus follows.

An area of future research is to check the goodness of fit for the additive model, and to compare it with other alternatives. One possible criteria is to examine the empirical difference between the model-based predicted survival function and the nonparametric maximum likelihood estimate of the survival function. Such estimate of survival function can be obtained using the approach of Wellner and Zhang (2000), within each category of the covariates.

Acknowledgement

This research was supported in part by the National Institutes of Health Grant HL 69720 (for Cai and Zeng) and the National Institutes of Health Grant CA 79466 (for Shen).

Appendix

A.1. Proof of Theorem 1

Consistency of $(\hat{\beta}_n, \hat{G}_n)$. In order to show that the estimators $(\hat{\beta}_n, \hat{G}_n)$ are consistent estimators for the true parameters (β_0, G_0), we consider a class of functions

with

$$\begin{aligned}\mathcal{F} = \{f_{\beta,G}(u,v,\delta_1,\delta_2) &= [(1-G(u)e^{-\beta^T Z(u)})I(1-G(u)e^{-\beta^T Z(u)}>0)]^{\delta_1}\\ &\times[(G(u)e^{-\beta^T Z(u)}-G(v)e^{-\beta^T Z(v)})I(G(u)e^{-\beta^T Z(u)}-G(v)e^{-\beta^T Z(v)}>0)]^{\delta_2}\\ &\times[G(v)e^{-\beta^T Z(v)}I(G(v)>0)]^{1-\delta_1-\delta_2} : (\beta,G)\in\Theta\}.\end{aligned}$$

First, we calculate the bracket covering number for this class. For any (β_1, G_1), $(\beta_2, G_2) \in \Theta$ such that $\sup_{t\in[\tau_u,\tau_v]}|G_1(t)-G_2(t)|<\epsilon, \|\beta_1-\beta_2\|<\epsilon$, we wish to set boundaries for the difference between $\sqrt{f_{\beta_1,G_1}}$ and $\sqrt{f_{\beta_2,G_2}}$. There are three scenarios.

1. $\delta_1 = 1$. Thus,

$$\begin{aligned}\sqrt{f_{\beta_1,G_1}}-\sqrt{f_{\beta_2,G_2}} = &\sqrt{(1-G_1(u)e^{-\beta_1^T Z(u)})I(1-G_1(u)e^{-\beta_1^T Z(u)}>0)}\\ &-\sqrt{(1-G_2(u)e^{-\beta_2^T Z(u)})I(1-G_2(u)e^{-\beta_2^T Z(u)}>0)}.\end{aligned}$$

There are two possibilities to be considered.
Case 1. $1-G_1(u)e^{-\beta_1^T Z(u)} > \delta$ for some positive constant δ such that $\delta - \epsilon e^{BM}(M+1) > 0$. Then $1-G_2(u)e^{-\beta_2^T Z(u)} > \delta-\epsilon e^{BM}(M+1) > 0$. Therefore,

$$|\sqrt{f_{\beta_1,G_1}}-\sqrt{f_{\beta_2,G_2}}| \le \frac{\epsilon e^{BM}+\epsilon M e^{BM}}{2\sqrt{\delta-\epsilon(M+1)e^{BM}}}.$$

Case 2. $1-G_1(u)e^{-\beta_1^T Z(u)} \le \delta$. Then $1-G_2(u)e^{-\beta_2^T Z(u)} \le \delta+\epsilon e^{BM}(1+M)$. Hence, $|\sqrt{f_{\beta_1,G_1}}-\sqrt{f_{\beta_2,G_2}}| \le \sqrt{\delta}+\sqrt{\delta+\epsilon e^{BM}(1+M)}$. We choose $\delta = 2\epsilon(M+1)e^{BM}$, then in either case, we obtain

$$|\sqrt{f_{\beta_1,G_1}}-\sqrt{f_{\beta_2,G_2}}| \le 4\sqrt{(M+1)e^{BM}\epsilon} = C_1\sqrt{\epsilon}.$$

2. $\delta_2 = 1$. For this situation,

$$\begin{aligned}&\sqrt{f_{\beta_1,G_1}}-\sqrt{f_{\beta_2,G_2}}\\ &=\sqrt{(G_1(u)e^{-\beta_1^T Z(u)}-G_1(v)e^{-\beta_1^T Z(v)})I(G_1(u)e^{-\beta_1^T Z(u)}-G_1(v)e^{-\beta_1^T Z(v)}>0)}\\ &-\sqrt{(G_2(u)e^{-\beta_2^T Z(u)}-G_2(v)e^{-\beta_2^T Z(v)})I(G_2(u)e^{-\beta_2^T Z(u)}-G_2(v)e^{-\beta_2^T Z(v)}>0)}.\end{aligned}$$

Essentially, we use the same arguments as above and divide it into two cases.
Case 1. $G_1(u)e^{-\beta_1^T Z(u)}-G_1(v)e^{-\beta_1^T Z(v)} > \delta$. Then $G_2(u)e^{-\beta_2^T Z(u)}-G_2(v)$ $e^{-\beta_2^T Z(v)} > \delta - 2\epsilon(M+1)e^{BM}$. So

$$|\sqrt{f_{\beta_1,G_1}}-\sqrt{f_{\beta_2,G_2}}| \le \frac{\epsilon e^{BM}(M+1)}{\sqrt{\delta-2\epsilon(M+1)e^{BM}}}.$$

Case 2. $G_1(u)e^{-\beta_1^T Z(u)} - G_1(v)e^{-\beta_1^T Z(v)} \leq \delta$. Then $G_2(u)e^{-\beta_2^T Z(u)} - G_2(v)e^{-\beta_2^T Z(v)} \leq \delta + 2\epsilon e^{BM}(M+1)$. So $|\sqrt{f_{\beta_1,G_1}} - \sqrt{f_{\beta_2,G_2}}| \leq \sqrt{\delta} + [\delta + 2\epsilon(M+1)e^{BM}]^{1/2}$. Let $\delta = 4\epsilon e^{BM}(M+1)$, then under both cases, $|\sqrt{f_{\beta_1,G_1}} - \sqrt{f_{\beta_2,G_2}}| \leq C_2\sqrt{\epsilon}$ for some constant C_2 depending on B, M.

3. $\delta_1 = \delta_2 = 0$. Similar arguments give $|\sqrt{f_{\beta_1,G_1}} - \sqrt{f_{\beta_2,G_2}}| \leq C_3\sqrt{\epsilon}$.

We thus obtain that $\ln N_{[]}(O(1)\sqrt{\epsilon}, \sqrt{\mathcal{F}}, \|\cdot\|_{l^\infty(\mathcal{X})}) \leq \ln N_{[]}(\epsilon, \Theta, \|\cdot\|_{l^\infty}) \leq O(1/\epsilon)$, where $N_{[]}(\cdot)$ denotes the bracket covering number. According to Theorem 2.4 and Lemma 1.1 in van der Geer (1993), and from the fact that $f_{\hat{\beta}_n,\hat{G}_n} \in \mathcal{F}$, we can prove that the Hellinger distance between $f_{\hat{\beta}_n,\hat{G}_n}$ and f_{β_0,G_0} converges to zero as $n \to \infty$. That is, $E[\sqrt{f_{\hat{\beta}_n,\hat{G}_n}} - \sqrt{f_{\beta_0,G_0}}]^2 \to 0$, a.s.. Second, since $\hat{\beta}_n$ is in a compact set K and $\hat{G}_n$ is a bounded non-increasing function, for any subsequence we can always find a sub-subsequence, still subscripted by n, such that $\hat{\beta}_n \to \beta^*$ and $\hat{G}_n(t) \to G^*(t)$ pointwise, with probability 1. Combining with the convergence of the Hellinger distance, and noticing that $f_{\hat{\beta}_n,\hat{G}_n}$ is bounded, we conclude that $f_{\beta^*,G^*} = f_{\beta_0,G_0}$ holds for (u,v) in the support $\mathcal{X}$ with probability 1. For any $(u,v) \in \mathcal{X}$, $G_0(v) > 0$ and under the assumption (A.5), for $\delta_1 = \delta_2 = 0$, we obtain that $G^*(v) > 0$ and $G^*(v)e^{-\beta^{*T} Z(v)} = G_0(v)e^{-\beta_0^T Z(v)}, a.s.$. However, since $[1, Z(v)]$ is linearly independent, we obtain that $\beta^* = \beta_0, G^*(v) = G_0(v)$, for $\delta_1 = \delta_2 = 0$. Now let $\delta_1 = 1$ (from the assumption (A.5)), then $G^*(u) = G_0(u)$. We finally conclude that $G^*(t) = G_0(t)$ for $t \in (\tau_u, \tau_v)$.

Furthermore, since G_0 is continuous, it holds that

$$\|\hat{\beta}_n - \beta_0\| \to 0, \sup_{t\in[\tau_u,\tau_v]} |\hat{G}_n(t) - G_0(t)| \to 0, a.s..$$

As a corollary, since $P(U < T \leq V|U, V, Z(t), t > 0), P(T \leq U|U, V, Z(t), t > 0)$, and $P(T > V|U, V, Z(t), t > 0)$ are larger than a positive constant with probability 1, we can assume that this is also true for each term in the $l_n(\beta, G)$ where the parameters are replaced by $(\hat{\beta}_n, \hat{G}_n)$.

Asymptotical normality of $\sqrt{n}(\hat{\beta}_n - \beta_0)$. First we need to derive the convergence rates of $(\hat{\beta}_n, \hat{G}_n)$. We use Theorem 3.2.1 of van der Vaart and Wellner (1996) for this purpose. The functionals $\mathbf{M}_n$ and $\mathbf{M}_0$ in Theorem 3.2.1 correspond to $\mathbf{P}_n \log f_{\theta,G}$ and $\mathbf{P} \log f_{\theta,G}$, respectively. The set Θ_0 in Theorem 3.2.1 is defined as

$$\Theta_0 = \{(\beta, G) : \|\beta - \beta_0\| < \epsilon_0, \sup_{t\in[\tau_u,\tau_v]} |G(t) - G_0(t)| < \epsilon_0,$$
$$G \text{ is a non-increasing step-function}\}$$

for a small constant $\epsilon_0 > 0$, and the distance $d((\beta_1, G_1), (\beta_2, G_2))$ stated in Theorem 3.2.1 is given by $\|\beta_1 - \beta_2\| + \|G_1 - G_2\|_{L_2([\tau_u, \tau_v])}$.

We check the conditions stated in Theorem 3.2.1. First, when (β, G) is in Θ_0 and $d((\beta, G), (\beta_0, G_0)) < \delta$, it is easy to see that $\log f_{\beta,G} - \log f_{\beta_0,G_0}$ is Lipschitz continuous with respect to β and G, and its $L_2(P)$ norm is bounded by $O(\delta)$. Thus, we have

$$\ln N_{[]}(\epsilon, \{\ln f_{\theta,G} : (\beta, G) \in \Theta_0\}, \|\cdot\|_{L_2(P)}) \leq O(1) \ln N_{[]}(\epsilon, \Theta_0, d) \leq O(\frac{1}{\epsilon}).$$

The maximal inequality for the empirical process $\mathbf{G}_n = \sqrt{n}(\mathbf{P}_n - \mathbf{P})$ gives that

$$E^* \sup_{d((\beta,G),(\beta_0,G_0)) \leq \delta, (\beta,G) \in \Theta_0} |\mathbf{G}_n(\ln f_{\beta,G} - \ln f_{\beta_0,G_0})| \leq O_p(1)\phi(\delta),$$

where $\phi(\delta) = \int_0^\delta \sqrt{1 + O(1/\epsilon)} d\epsilon (1 + \int_0^\delta \sqrt{1 + O(1/\epsilon)} d\epsilon / (\sqrt{n}\delta^2)) = O_p(1)(\sqrt{\delta} + 1/(\delta\sqrt{n}))$. Second, for any vector h and g, $E[l_\beta h + l_G[g]]^2 \geq 0$. Furthermore, if there exist a vector h and a function g such that $l_\beta h + l_G[g] = 0$, then let $\delta_1 = \delta_2 = 0$ to obtain $-Z(V)h + g(V)/G(V) = 0$. Hence, $h = 0$ and then $g(t) = 0, t \in [\tau_u, \tau_v]$. We conclude that the information operator at (β_0, G_0) is a positive bilinear operator in the Hilbert space $R^k \times L_2([\tau_u, \tau_v])$. Hence,

$$\begin{aligned} &\sup_{d((\beta,G),(\beta_0,G_0)) > \frac{\delta}{2}} E[(\beta - \beta_0)^T l_{\beta\beta}(\beta - \beta_0) + 2(\beta - \beta_0)^T l_G[G - G_0] \\ &+ l_{GG}[G - G_0, G - G_0]] \\ = &\sup_{d((\beta,G),(\beta_0,G_0)) > \frac{\delta}{2}} -E[(l_\beta(\beta - \beta_0), l_G[G - G_0])^{\otimes 2}] \leq -C\delta^2. \end{aligned}$$

As a result, by a Taylor expansion, it holds that

$$\sup_{\frac{\delta}{2} \leq d((\beta,G),(\beta_0,G_0)) \leq \delta, (\beta,G) \in \Theta_0} \mathbf{P}[\log f_{\beta,G} - \log f_{\beta_0,G_0}] \leq -O(1)\delta^2.$$

Finally, we note that $\phi(1/r_n) r_n^2 \leq \sqrt{n}$ for $r_n = O(n^{1/3})$. Thus, according to Theorem 3.2.1 of van der Vaart and Wellner (1996), we obtain $d((\hat{\theta}_n, \hat{G}_n), (\theta_0, G_0)) = O_p(n^{-1/3})$.

With the derived convergence rate and the proof of the existence of the efficient score function for β given in Section 3, the asymptotic normality for $\sqrt{n}(\hat{\beta}_n - \beta_0)$ follows from the same arguments which Huang (1996) used in the asymptotic proof for the current status data. We can easily check the conditions of Theorem 6.1 in Huang (1996). We skip the details, however, due to limited space. The asymptotic variance for $\sqrt{n}(\hat{\beta}_n - \beta_0)$ is equal to the generalized

Cramér-Rao bound, given by $E[l_\beta^* l_\beta^{*T}]^{-1}$. Here, l_β^* refers to the efficient score function for the parameter β.

A.2. Derivation of efficient score function

The score function for β_0 is given by

$$l_\beta = \frac{\delta_1 G_0(U) e^{-\beta_0{}^T Z(U)} Z(U)}{1 - G_0(U) e^{-\beta_0{}^T Z(U)}} - \frac{\delta_2 (G_0(U) e^{-\beta_0{}^T Z(U)} Z(U) - G_0(V) e^{-\beta_0{}^T Z(V)} Z(V))}{G_0(U) e^{-\beta_0{}^T Z(U)} - G_0(V) e^{-\beta_0{}^T Z(V)}} - (1 - \delta_1 - \delta_2) Z(V).$$

We consider G_0 to be a "parameter" with value in a metric space consisting of all non-increasing functions. The score function for G_0 then has the form

$$l_G[g] = \frac{-\delta_1 g(U) e^{-\beta_0{}^T Z(U)}}{1 - G_0(U) e^{-\beta_0{}^T Z(U)}} + \frac{\delta_2 (g(U) e^{-\beta_0{}^T Z(U)} - g(V) e^{-\beta_0{}^T Z(V)})}{G_0(U) e^{-\beta_0{}^T Z(U)} - G_0(V) e^{-\beta_0{}^T Z(V)}} + \frac{(1 - \delta_1 - \delta_2) g(V)}{G_0(V)},$$

where g is any function in $L_2([\tau_u, \tau_v])$. To calculate the efficient score function for $\beta_0 \in R^k$, we need to find k functions $g^* = (g_1, \ldots, g_k)^T$ in $L_2([\tau_u, \tau_v])$ that, based on semiparametric efficiency theory (Bickel et al. (1993)), should satisfy

$$E\left[(l_\beta - l_G[g^*])^T l_G[\tilde{g}]\right] = 0,$$

where $\tilde{g}$ is any k-vector function in $L_2([\tau_u, \tau_v])$.

Define $S_Z(t) = P(T > t|Z) = G_0(t) \exp\{-\beta_0^T Z(t)\}$. After some technical manipulations, we obtain $E[h_1(U,V)\tilde{g}(U) + h_2(U,V)\tilde{g}(V)] = 0$, where

$$\begin{aligned}
h_1(U,V) = E\Big[& \frac{-G_0(U) Z(U) e^{-2\beta_0{}^T Z(U)}}{1 - S_Z(U)} \\
& + \frac{-G_0(U) Z(U) e^{-\beta_0{}^T Z(U)} + G_0(V) Z(V) e^{-\beta_0{}^T Z(V)}}{S_Z(U) - S_Z(V)} e^{-\beta_0{}^T Z(U)} | U, V \Big] \\
& - E\Big[\frac{e^{-2\beta_0{}^T Z(U)}}{1 - S_Z(U)} + \frac{e^{-2\beta_0{}^T Z(U)}}{S_Z(U) - S_Z(V)} | U, V \Big] g^*(U) \\
& + E\Big[\frac{e^{-\beta_0{}^T Z(U) - \beta_0{}^T Z(V)}}{S_Z(U) - S_Z(V)} | U, V \Big] g^*(V), \\
h_2(U,V) = -E\Big[& \frac{-G_0(U) Z(U) e^{-\beta_0{}^T Z(U)} + G_0(V) Z(V) e^{-\beta_0{}^T Z(V)}}{S_Z(U) - S_Z(V)} e^{-\beta_0{}^T Z(V)} \\
& + \frac{G_0(V) Z(V) e^{-2\beta_0{}^T Z(V)}}{S_Z(V)} | U, V \Big] + E\Big[\frac{e^{-\beta_0{}^T Z(U) - \beta_0{}^T Z(V)}}{S_Z(U) - S_Z(V)} | U, V \Big] g^*(U) \\
& - E\Big[\frac{e^{-2\beta_0{}^T Z(V)}}{S_Z(U) - S_Z(V)} + \frac{e^{-2\beta_0{}^T Z(V)}}{S_Z(V)} | U, V \Big] g^*(V).
\end{aligned}$$

Thus, $f_U(x)E[h_1(U,V)|U=x] + f_V(x)E[h_2(U,V)|V=x] = 0$. Further simplification gives that $g^*(x)$ satisfies the integral equation

$$-a(x)g^*(x) + \int_y b(x,y)g^*(y)dy = c(x),$$

where $a(x), b(x,y)$ and $c(x)$ are given in Section 3. Clearly, $a(x) > 0$ for $x \in [\tau_u, \tau_v]$. Then the above equation is equivalent to $(\mathbf{I}+\mathbf{A})(g^*) = -c(x)/a(x)$, where $\mathbf{I}$ is the identity operator and $\mathbf{A}(g^*) = -\int_y b(x,y)g^*(y)dy/a(x)$.

Note that from (A.2) and (A.3), $\mathbf{A}$ maps any L_2-integrable function to a continuously differentiable function; thus it is a compact operator from $L_2([\tau_u, \tau_v])$ to $L_2[\tau_u, \tau_v]$. Therefore, to show the invertibility of the operator $(\mathbf{I}+\mathbf{A})$, by Theorem 4.25 in Rudin (1973), it suffices to show that the operator $(\mathbf{I}+\mathbf{A})$ is one-to-one. Now if $(\mathbf{I}+\mathbf{A})(g) = 0$, by reversing the above derivation for the situation $c(x) = 0$, we obtain that $E[l_G[g]^T l_G[\tilde{g}]] = 0$ for any $\tilde{g}$. Particularly, we let $\tilde{g} = g$ to obtain $l_G[g] = 0, a.s.$ Finally, by the expression of $l_G[g]$, (A.4) and (A.5), we conclude that $g(x) = 0$ for any $x \in [\tau_u, \tau_v]$. Therefore, the operator $(\mathbf{I}+\mathbf{A})$ is one-to-one. We denote g^* to be the solution. Then the efficient score function for β_0 is $l^*_\beta = l_\beta - l_G[g^*]$, and the efficiency boundary for β_0 is $E[l^*_\beta l^{*T}_\beta]^{-1}$.

References

Andersen, P. K., Borgan, Ø., Gill, R. D. and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer-Verlag, New York.

Bickel, P. J., Klaassen, C. A. J., Ritov, Y. and Wellner, J. A. (1993). *Efficient and Adaptive Inference in Semiparametric Models*. Johns Hopkins University Press, Baltimore.

Cox, D. R. (1972). Regression models and life tables (with discussion). *J. Roy. Statist. Soc. Ser. B* **34**, 187-220.

Finkelstein, D. M. (1986). A proportional hazards model for interval-censored failure time data. *Biometrics* **42**, 845-854.

Ghosh, D. (2001). Efficiency considerations in the additive hazards model with current status data. *Statist. Neerlandica* **55**, 367-376.

Huang, J. (1996). Efficient estimation for the proportional hazard model with interval censoring. *Ann. Statist.* **24**, 540-568.

Lin, D. Y., Oakes, D. and Ying, Z. L. (1998). Additive hazards regression with current status data. *Biometrika* **85**, 289-298.

Lin, D. Y. and Ying, Z. L. (1994) Semiparametric analysis of the additive risk model. *Biometrika* **81**, 61-71.

McKeague, I. W. and Sasieni, P. D. (1994). A partly parametric additive risk model. *Biometrika* **81**, 501-514.

Murphy, S. A. and van der Vaart, A. W. (2000). On profile likelihood. *J. Amer. Statist. Assoc.* **95**, 45-485.

Rabinowitz, D. and Tsiatis, A. A. (1995). Regression with interval-censored data. *Biometrika* **82**, 501-513.

Rossini, A. J. and Tsiatis, A. (1996). A semiparametric proportional odds regression model for the analysis of current status data. *J. Amer. Statist. Assoc.* **91**, 713-321.

Rudin, W. (1973). *Functional Analysis*. McGraw-Hill, New York.

Satten, G. A., Datta, S. and Williamson, J. M. (1998). Inference based on imputed failure times for the proportional hazards model with interval-censored data. *J. Amer. Statist. Assoc.* **93**, 318-327.

Shen, Y. and Cheng, S. C. (1999). Confidence bands for cumulative incidence curves under the additive risk model. *Biometrics* **55**, 1093-1100.

Sun, J. G. (1996). A nonparametric test for interval censored failure time data with application to AIDS studies. *Statist. Medicine* **15**, 1387-1395.

Turnbull, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *J. Roy. Statist. Soc. Ser. B* **38**, 290-295.

van de Geer, S. (1993). Hellinger-consistency of certain nonparametric maximum likelihood estimators. *Ann. Statist.* **21**, 14-44.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer, Berlin.

Wellner, J. A. and Zhang, Y. (2000). Two estimators of the mean of a counting process with panel count data. *Ann. Statist.* **28**, 779-814.

Zhang, Y., Liu, W., and Zhan, Y. H. (2001). A nonparametric two-sample test of the failure function with interval censoring case 2. *Biometrika* **88**, 677-686.

Department of Biostatistics, University of North Carolina at Chapel Hill, North Carolina 27599, U.S.A.

E-mail: dzeng@bios.unc.edu

Department of Biostatistics, University of North Carolina at Chapel Hill, North Carolina 27599, U.S.A.

E-mail: cai@bios.unc.edu

Department of Biostatistics, M. D. Anderson Cancer Center, The University of Texas, Houston, Texas 77030, U.S.A.

E-mail: yushen@mdanderson.org

(Received August 2004; accepted December 2004)