Published in final edited form as:

Scand Stat Theory Appl. 2008 June 1; 35(2): 228–247. doi:10.1111/j.1467-9469.2007.00581.x.

Non-parametric Estimation of a Survival Function with Two-stage Design Studies

GANG LI and

Department of Biostatistics, University of California at Los Angeles

CHI-HONG TSENG

Department of Medicine, University of California at Los Angeles

Abstract

The two-stage design is popular in epidemiology studies and clinical trials due to its cost effectiveness. Typically, the first stage sample contains cheaper and possibly biased information, while the second stage validation sample consists of a subset of subjects with accurate and complete information. In this paper, we study estimation of a survival function with right-censored survival data from a two-stage design. A non-parametric estimator is derived by combining data from both stages. We also study its large sample properties and derive pointwise and simultaneous confidence intervals for the survival function. The proposed estimator effectively reduces the variance and finite-sample bias of the Kaplan–Meier estimator solely based on the second stage validation sample. Finally, we apply our method to a real data set from a medical device post-marketing surveillance study.

Keywords

censoring; Kaplan–Meier estimator; martingale; Nelson–Aalen estimator; truncation

1. Introduction

In this paper, we study non-parametric estimation of a survival function based on a two-stage design in which survival information could possibly be biased in the first stage data, while accurate information is collected only for a subset of individuals in the second stage. Specifically, the first stage data consist of right-censored survival data of N independent and identically distributed (i.i.d.) pairs, (X_{1i}, D_{1i}) , $i = 1, \dots, N$, where for subject i , $X_{1i} = \min(T_{1i}, C_{1i})$ is the observation time, $D_{1i} = I(T_{1i} \leq C_{1i})$ is the failure indicator, T_{1i} is a non-negative absolute continuous random variable with survival function $S_1(t) = P(T_{1i} \geq t)$ and C_{1i} denotes the censoring time. In the second stage, we are able to obtain accurate follow-up data reflecting the true survival time, say T_2 , for a subset $V(n)$ of n subjects. For subject $i \in V(n)$, in addition to (X_{1i}, D_{1i}) from the first stage, one also observes (X_{2i}, D_{2i}) , where $X_{2i} = \min(T_{2i}, C_{2i})$, $D_{2i} = I(T_{2i} \leq C_{2i})$ and T_{2i} and C_{2i} are the corresponding survival time and censoring time with respect to the second stage follow-up data. We further assume that the second stage sample is a simple random sample drawn from the first stage and that the censoring time is independent of the survival time in both stages. Of interest is to estimate the survival function $S_2(t) = P(T_{2i} \geq t)$ of the survival time T_{2i} .

The two-stage design arises frequently in epidemiology and other studies. One of the motivating applications for this research is the medical device postmarketing surveillance study. The US Congress first instituted postmarked surveillance of certain medical devices with the enactment of the Safe Medical Devices Act in 1990. In the Food and Drug Administration Act in 1997, Congress modified this authority, giving the agency more discretion in imposing postmarked surveillance and establishing a time limit. The primary purpose of the postmarked surveillance study is to evaluate the safety and efficacy of marketed devices, which usually involves survival analysis to estimate the device failure rate and/or incidence of adverse events. Medical device companies usually maintain a database registering all the market devices, related failures and associated adverse events. However, the company administrative data are often associated with the under-reporting problem that could underestimate the failure and adverse event rate. To calibrate the under-reporting bias, an active follow-up validation sample is necessary in conducting a postmarked surveillance study. Devices in this active sample are carefully followed to ensure comprehensive data collection that reflects the true adverse event/failure rate and device explant/out-of-service status. In this case, the company administrative data are the first stage sample and the active follow-up sample is the second stage sample.

To the best of our knowledge, no result is available in the literature on non-parametric estimation of the target survival function $S_2(t)$ based on the two-stage data. Note that the Kaplan–Meier estimator based on the first stage data is not a consistent estimator of $S_2(t)$ if $S_1 \neq S_2$. Although the Kaplan–Meier estimator based on the second stage sample is a consistent estimator of $S_2(t)$, it ignores all the information contained in the first stage data. The purpose of this paper is to develop a non-parametric method for estimating $S_2(t)$ by combining information from both stage samples. We are able to bridge the two-stage data together as both the first stage measurements and the second stage measurements are available for the subjects in the second stage validation sample $V(n)$. Our estimator corrects the bias in the first stage sample, and is more efficient and less biased in the tail than the Kaplan–Meier estimator based solely on the second stage sample. We also derive pointwise and simultaneous confidence bands for the survival function S_2 . Our results are derived using the counting process setting and therefore apply to other types of survival data including right-censored and left-truncated data. Finally, we note that our results do not follow directly from the standard martingale theory as the martingale property does not hold with respect to the joint filtration from the two-stage data. (See Remark 1 for more details.) A careful theoretical development is indeed needed.

We point out that the two-stage design has received much attention during the past two decades (cf. Prentice, 1982; White, 1982; Breslow & Cain, 1988; Carrol & Wand, 1991; Flanders & Greenland, 1991; Pepe & Fleming, 1991; Robins *et al.*, 1994,¹⁹⁹⁵; Zhou & Pepe, 1995; Breslow & Holubkov, 1997; Muller & Roeder, 1997; Scott & Wild, 1997; Wang *et al.*, 1997; Chen & Chen, 2000; Zhou & Wang, 2000; Chatterjee *et al.*, 2003). However, most research has exclusively focused on regression problems in which the covariates are surrogates or with errors in the first stage sample, and their primary research interest has been to estimate the regression parameters. Brookmeyer & Yasui (1995) discussed issues with passive surveillance disease registry data where there may be under-reporting of disease cases and incomplete vital status. The main interest was to estimate relative risks. They also considered the situation with available external validation sample, and estimated the relative risks and the extent of under-reporting with a proportional hazard model. Very few works address the issue of estimation of the survival function with the exception of Chernick *et al.* (2002). Chernick *et al.* addressed the same issue of estimating survival curve with the concept of capture–recapture. They divided the time range into several intervals and used the per cent of failures in active sample to estimate the number of failures in whole population in each interval independently. It is an *ad hoc* method that does not recognize that each data point represents a continuous time process. It is worth noting that Chen (2002) presented a method to combine data from both stages in the Cox (1972) regression analysis context. Chen (2002) studied only the estimation of the

regression parameters. We adapt the concept of Chen (2002) to derive our estimator based on the Nelson–Aalen estimator of the cumulative hazard function, which in turn leads to an estimator of the survival function.

The rest of the paper is organized as follows. In section 2, we derive a non-parametric estimator of S_2 based on the two-stage survival data, and state its large sample properties. In section 3, we apply our method to analyse a real-life example. Section 4 presents some simulation studies to evaluate the performance of our method. Section 5 gives some concluding remarks. The proofs are collected in the Appendix.

2. Main results

2.1. Notation and assumptions

We first give some notations and definition needed in the paper. Using the standard counting process notation, for the first stage data, define the following at-risk processes

$$\begin{aligned} Y_{1i}(t) &= I(X_{1i} \geq t) \\ &, \bar{Y}_1(t) \\ &= \sum_{i=1}^N Y_{1i}(t), \end{aligned}$$

and the counting processes

$$\begin{aligned} N_{1i}(t) &= I(X_{1i} \leq t, D_{1i}=1), \bar{N}_1(t) \\ &= \sum_{i=1}^N N_{1i}(t). \end{aligned}$$

Let

$$\begin{aligned} M_{1i}(t) &= N_{1i}(t) \\ &- \int_0^t Y_{1i}(u) d\Lambda_1(u) \\ &, \bar{M}_1(t) \\ &= \sum_{i=1}^N M_{1i}(t), \end{aligned}$$

where $\Lambda_1(t)$ is the cumulative hazard function of T_1 . Note that $M_{1i}(t)$ is a martingale with respect to the marginal filtration $\mathcal{F}_1(t) = \sigma\{N_{1i}(s), Y_{1i}(s): 0 \leq s \leq t\}$, the failure and censoring histories in the first stage data up to time t . Let

$$\widehat{\Lambda}_1(t) = \int_0^t \frac{d\bar{N}_1(u)}{\bar{Y}_1(u)} \quad (1)$$

be the Nelson–Aalen estimator of $\Lambda_1(t)$.

Similarly, define

$$\begin{aligned}\bar{Y}_1^V(t) &= \sum_{i \in V(n)} Y_{1i}(t) \\ &, \bar{N}_1^V(t) \\ &= \sum_{i \in V(n)} N_{1i}(t) \text{ and } \bar{M}_1^V(t) \\ &= \sum_{i \in V(n)} M_{1i}(t)\end{aligned}$$

to be the at-risk process, counting process and martingale process with respect to the marginal filtration $\mathcal{F}_1$, respectively, of the first stage observations for individuals in the active validation sample, and

$$\begin{aligned}Y_{2i}(t) &= I(X_{2i} \geq t), \quad \bar{Y}_2^V(t) = \sum_{i \in V(n)} Y_{2i}(t), \\ N_{2i}(t) &= I(X_{2i} \leq t, D_{2i} = 1), \quad \bar{N}_2^V(t) = \sum_{i \in V(n)} N_{2i}(t),\end{aligned}$$

and

$$\begin{aligned}M_{2i}(t) &= N_{2i}(t) \\ &- \int_0^t Y_{2i}(u) d\Lambda_2(u) \\ &, \bar{M}_2^V(t) \\ &= \sum_{i \in V(n)} M_{2i}(t),\end{aligned}$$

the at-risk processes, counting processes and martingale processes with respect to the marginal filtration $\mathcal{F}_2(t) = \sigma\{N_{2i}(s), Y_{2i}(s): 0 \leq s \leq t\}$, respectively, of the second stage accurate observations for individuals in the active validation sample, where $\Lambda_2(t)$ is the cumulative hazard function of T_2 . Let

$$\begin{aligned}\widehat{\Lambda}_1^V(t) &= \int_0^t \frac{d\bar{N}_1^V(u)}{\bar{Y}_1^V(u)} \text{ and } \widehat{\Lambda}_2^V(t) \\ &= \int_0^t \frac{d\bar{N}_2^V(u)}{\bar{Y}_2^V(u)}\end{aligned}\tag{2}$$

be the Nelson–Aalen estimators of the cumulative hazard functions $\Lambda_1(t)$ and $\Lambda_2(t)$, respectively, based on the active validation samples.

The main results of the paper are proved under the following set of regularity conditions.

Condition 1 (finite interval). $\Lambda_i(\tau) < \infty$, $i = 1, 2$, where τ is usually the time when data collection ends.

Condition 2 (asymptotic stability). There exist some functions $y_1(t)$ and $y_2(t)$ such that for $i = 1, 2$,

$$\sup_{t \in [0, \tau]} \left| \frac{\bar{Y}_i^V(t)}{n} - y_i(t) \right| \xrightarrow{P} 0 \quad \text{and} \quad \sup_{t \in [0, \tau]} \left| \frac{\bar{Y}_1(t)}{N} - y_1(t) \right| \xrightarrow{P} 0,$$

as $n, N \rightarrow \infty$

Condition 3 (asymptotic regularity conditions). The functions $y_1(t)$ and $y_2(t)$ are uniformly continuous and bounded away from zero on the interval $[0, \tau]$.

2.2. The estimators

We first derive an estimator for the cumulative hazard function $\Lambda_2(t)$ combining the two-stage design data. The survival function $S_2(t)$ is then estimated by a corresponding product-limit estimator.

It is well known that $\hat{\Lambda}_1(t)$ and $\hat{\Lambda}_1^V(t)$ are consistent estimators of $\Lambda_1(t)$ and that $\hat{\Lambda}_2^V(t)$ is a consistent estimator $\Lambda_2(t)$ (cf. Andersen *et al.*, 1993). To derive an estimator of $\Lambda_2(t)$ using both stage samples, we need to know the joint behaviour of $\hat{\Lambda}_1(t)$, $\hat{\Lambda}_1^V(t)$ and $\hat{\Lambda}_2^V(t)$. Let

$$\begin{aligned} W_N(t) = & N^{1/2} [\{ \hat{\Lambda}_1^V(t) \\ & - \Lambda_1(t) \} \\ & , \{ \hat{\Lambda}_2^V(t) \\ & - \Lambda_2(t) \} \\ & , \{ \hat{\Lambda}_1(t) - \Lambda_1(t) \}] \end{aligned}$$

and let $\mathcal{D}[0, \tau]^3$ be the metric space consisting of right-continuous functions with left limits $\{f_1(t), f_2(t), f_3(t)\}$, where $f_k: [0, \tau] \rightarrow \mathcal{R}$ for $k = 1, 2, 3$. The metric is defined as

$$\begin{aligned} d(f, g) = & \max_{t \in [0, \tau]} \{ \| f_k(t) \\ & - g_k(t) \| \\ & , 1 \leq k \leq 3 \} \quad \text{for } f, g \in \mathcal{D}[0, \tau]^3. \end{aligned}$$

Theorem 1 gives the weak convergence of $W_N(t)$.

Theorem 1—Under conditions 1, 2, 3, and $n/N \rightarrow \rho$ for some constant $0 < \rho < 1$ as $n, N \rightarrow \infty$, the random field $W_N(t)$ converges weakly to $\mathcal{W}(t)$ in $\mathcal{D}[0, \tau]^3$ as $N \rightarrow \infty$, where $\mathcal{W}(t) = \{\mathcal{W}_1^V(t), \mathcal{W}_2^V(t), \mathcal{W}_1(t)\}$ is a zero-mean Gaussian random field with the variance–covariance function among $\mathcal{W}_1^V(t_1), \mathcal{W}_2^V(t_2)$ and $\mathcal{W}_1(t_3)$ being

$$\begin{pmatrix} \rho^{-1}\sigma_{11}(t_1) & \rho^{-1}\sigma_{12}(t_1, t_2) & \rho^{-1/2}\sigma_{11}(t_1 \wedge t_3) \\ \rho^{-1}\sigma_{12}(t_1, t_2) & \rho^{-1}\sigma_{22}(t_2) & \rho^{-1/2}\sigma_{12}(t_3, t_2) \\ \rho^{-1/2}\sigma_{11}(t_1 \wedge t_3) & \rho^{-1/2}\sigma_{12}(t_3, t_2) & \sigma_{11}(t_3) \end{pmatrix},$$

with

$$\begin{aligned} \sigma_{11}(t) &= E \left[\int_0^t \frac{dM_{1i}(u)}{y_1(u)} \right]^2, \quad \sigma_{22}(t) = E \left[\int_0^t \frac{dM_{2i}(u)}{y_2(u)} \right]^2, \\ \sigma_{12}(t_1, t_2) &= E \left[\int_0^{t_1} \frac{dM_{1i}(u)}{y_1(u)} \int_0^{t_2} \frac{dM_{2i}(u)}{y_2(u)} \right]. \end{aligned}$$

Furthermore, $\sigma_{11}(t)$, $\sigma_{22}(t)$ and $\sigma_{12}(t_1, t_2)$ can be consistently estimated by

$$\widehat{\sigma}_{11}(t) = n \sum_{i \in V(n)} \left[\int_0^t \frac{d\widehat{M}_{1i}(u)}{\widehat{Y}_1^V(u)} \right]^2, \quad (3)$$

$$\widehat{\sigma}_{22}(t) = n \sum_{i \in V(n)} \left[\int_0^t \frac{d\widehat{M}_{2i}(u)}{\widehat{Y}_2^V(u)} \right]^2, \quad (4)$$

$$\widehat{\sigma}_{12}(t_1, t_2) = n \sum_{i \in V(n)} \int_0^{t_1} \frac{d\widehat{M}_{1i}(u)}{\widehat{Y}_1^V(u)} \int_0^{t_2} \frac{d\widehat{M}_{2i}(u)}{\widehat{Y}_2^V(u)}, \quad (5)$$

where

$$\begin{aligned} \widehat{M}_{1i}(t) &= N_{1i}(t) - \int_0^t Y_{1i}(u) d\widehat{\Lambda}_1^V(u) \quad \text{and} \quad \widehat{M}_{2i}(t) \\ &= N_{2i}(t) \\ &\quad - \int_0^t Y_{2i}(u) d\widehat{\Lambda}_2^V(u). \end{aligned}$$

Remark 1: It is now a well-known fact that the marginal asymptotic distribution can be established by a direct application of the standard martingale central limit theorem (cf. Andersen *et al.*, 1993, theorem 2.5.1). However, the joint asymptotic distribution cannot be established the same way. Due to the intraclass dependence, $M_{1i}(t)$, $M_{2i}(t)$ are not martingales with respect to the joint filtration $\mathcal{F}(t) = \mathcal{F}_1(t) \vee \mathcal{F}_2(t)$ generated by all the failure and censoring histories up to time t . Hence, the standard martingale central limit theorem cannot be applied directly to obtain the joint asymptotic normality result in the above theorem. A different proof is indeed needed and provided in the Appendix.

It follows from theorem 1 that for a given time t , the joint distribution of $\widehat{\Lambda}_1^V(t)$ and $\widehat{\Lambda}_2^V(t)$ is approximately bivariate normal:

$$\mathcal{N}\left(\begin{pmatrix} \Lambda_1(t) \\ \Lambda_2(t) \end{pmatrix}, \frac{1}{n} \begin{pmatrix} \sigma_{11}(t) & \sigma_{12}(t) \\ \sigma_{12}(t) & \sigma_{22}(t) \end{pmatrix}\right),$$

with $\sigma_{12}(t) = \sigma_{12}(t, t)$. Therefore,

$$E\left(\widehat{\Lambda}_2^V(t) - \Lambda_2(t) | \widehat{\Lambda}_1^V(t) - \Lambda_1(t)\right) \approx \sigma_{12}(t) \sigma_{11}^{-1}(t) (\widehat{\Lambda}_1^V(t) - \Lambda_1(t)).$$

This suggests that $\Lambda_2(t)$ be estimated by

$$\begin{aligned} \widehat{\Lambda}_2(t) = & \widehat{\Lambda}_2^V(t) \\ & - \widehat{\sigma}_{12}(t) \widehat{\sigma}_{11}^{-1}(t) (\widehat{\Lambda}_1^V(t) \\ & - \widehat{\Lambda}_1(t)), \end{aligned} \quad (6)$$

where $\widehat{\Lambda}_1(t)$, $\widehat{\Lambda}_1^V(t)$, $\widehat{\Lambda}_2^V(t)$, $\widehat{\sigma}_{11}(t)$ and $\widehat{\sigma}_{12}(t) = \widehat{\sigma}_{12}(t, t)$ are defined by (1), (2), (3) and (5) respectively.

Consequently, we propose to estimate the survival function $S_2(t)$ by the corresponding product-limit estimator,

$$\widehat{S}_2(t) = \prod_{s \leq t} (1 - d\widehat{\Lambda}_2(s)). \quad (7)$$

Remark 2: If $\widehat{\sigma}_{12}(t)$ is close to zero, then the proposed estimator $\widehat{\Lambda}_2(t)$ is close to $\widehat{\Lambda}_2^V(t)$. This is a nice property because the first stage data are not expected to contribute much useful information on the estimation of $\Lambda_2(t)$ if the first stage data are barely correlated with the second stage validation sample.

Remark 3: If the first stage observations are identical to the second stage observations for individuals in the validation sample $V(n)$, then it is easy to see from the definitions that $\widehat{\Lambda}_2^V(t) = \widehat{\Lambda}_1^V(t)$ and $\widehat{\sigma}_{12}(t) = \widehat{\sigma}_{11}(t)$. Therefore, $\widehat{\Lambda}_2(t) = \widehat{\Lambda}_1(t)$. This is again an appealing property because if the first stage data are indeed without any bias, we should just estimate the underlying distribution using the larger first stage data which include the second stage sample as a subset.

Remark 4: A shortcoming of the proposed estimator $\widehat{S}_2(t)$ is that it is not necessarily monotonically non-increasing in t . This problem is, however, local and minor especially for large samples. In practice, some simple modifications such as $\widehat{S}_2^*(t) = \max_{s \leq t} \widehat{S}_2(s)$ (Lin & Ying, 1994), $\widehat{S}_2^{**}(t) = \min_{s \geq t} \widehat{S}_2(s)$, or the ‘Pool-Adjacent-Violator’ algorithm (Barlow *et al.*, 1972) can be applied to correct this problem. The modified estimators have been observed to possess almost identical large sample properties to $\widehat{S}_2(t)$ in our limited simulation studies (Tseng, 2004).

2.3. Weak convergence of the estimators

Theorem 2 states the large sample properties of $\hat{\Lambda}_2(t)$ and $\hat{S}_2(t)$. The proofs are given in the Appendix. Let $D[0, \tau]$ be the standard Skorohod space in $[0, \tau]$.

Theorem 2—Let $\hat{\Lambda}_2(t)$ and $\hat{S}_2(t)$ be defined by (6) and (7) respectively. Assume that $n/N \rightarrow \rho$ for some constant $0 < \rho < 1$ as $n, N \rightarrow \infty$. Then, under conditions 1, 2 and 3,

$$\begin{aligned} \sqrt{n}(\hat{\Lambda}_2(\cdot) \\ - \Lambda_2(\cdot)) \xrightarrow{d} \mathcal{W}_2(\cdot) \text{ in } D[0, \tau] \end{aligned}$$

and

$$\begin{aligned} \sqrt{n}(\hat{S}_2(\cdot) \\ - S_2(\cdot)) \xrightarrow{d} S_2(\cdot)\mathcal{W}_2(\cdot) \text{ in } D[0, \tau] \end{aligned}$$

as $n \rightarrow \infty$, where $\mathcal{W}_2(t)$ is a zero-mean Gaussian process with covariance function given by

$$\begin{aligned} \xi(s, t) &= \text{cov}(\mathcal{W}_2(s) \\ &\quad, \mathcal{W}_2(t)) \\ &= \sigma_{22}(s \wedge t) \\ &\quad - (1 - \rho)\sigma_{12}(t)\sigma_{11}^{-1}(s \vee t)\sigma_{12}(s) \text{ for } 0 \leq s, t \leq \tau. \end{aligned}$$

Furthermore, the covariance function $\xi(s, t)$ can be consistently estimated by

$$\begin{aligned} \hat{\xi}_{22}(s, t) &= \hat{\sigma}_{22}(s \wedge t) \\ &\quad - \left(1 - \frac{n}{N}\right) \hat{\sigma}_{12}(t) \hat{\sigma}_{11}^{-1}(s \vee t) \hat{\sigma}_{12}(s) \text{ for } 0 \leq s, t \leq \tau. \end{aligned}$$

Consequently, in the special case of $t = s$, the variance function $\sigma_{22}^*(t) = \xi(t, t)$ can be consistently estimated by

$$\begin{aligned} \hat{\sigma}_{22}^*(t) &= \hat{\sigma}_{22}(t) \\ &\quad - \left(1 - \frac{n}{N}\right) \hat{\sigma}_{12}(t) \hat{\sigma}_{11}^{-1}(t) \hat{\sigma}_{12}(t) \text{ for any } t \in [0, \tau]. \end{aligned}$$

2.4. Pointwise and simultaneous confidence intervals for survival probabilities

It follows immediately from theorem 2 that for any given t , a $100(1 - \alpha)\%$ confidence interval for $S_2(t)$ is given by

$$\widehat{S}_2(t) \pm z_{1-\alpha/2} \widehat{S}_2(t) \sqrt{\sigma_{22}^*(t)}, \quad (8)$$

where $z_{1-\alpha/2}$ is the $(1 - \alpha/2)$ th percentile of the standard normal distribution.

However, the results of theorem 2 cannot be readily applied to construct simultaneous confidence bands for the survival function S_2 over a given interval $[t_1, t_2]$ as the asymptotic distribution of

$$\sup_{t \in [t_1, t_2]} |\widehat{S}_2(t) - S_2(t)|$$

is intractable. Below we develop a Monte Carlo method to construct confidence bands for Λ_2 and S_2 , using an idea similar to Lin *et al.* (1994).

Note that

$$\begin{aligned} U_n(t) &= \sqrt{n}(\widehat{\Lambda}_2(t) - \Lambda_2(t)) \\ &= \sqrt{n}(\widehat{\Lambda}_2^V(t) - \Lambda_2(t)) \\ &= -\widehat{\sigma}_{21}(t)\widehat{\sigma}_{11}^{-1}(t) \left\{ \sqrt{n}(\widehat{\Lambda}_1^V(t) - \Lambda_1(t)) - \sqrt{\frac{n}{N}} \sqrt{N}(\widehat{\Lambda}_1(t) - \Lambda_1(t)) \right\} \\ &= \sqrt{n} \sum_{i \in V} \int_0^t \frac{dM_{2i}(u)}{\widehat{Y}_2^V(u)} \\ &= -\widehat{\sigma}_{21}(t)\widehat{\sigma}_{11}^{-1}(t) \left\{ \sqrt{n} \sum_{i \in V} \int_0^t \frac{dM_{1i}(u)}{\widehat{Y}_1^V(u)} - \sqrt{\frac{n}{N}} \sqrt{N} \sum_{i=1}^N \int_0^t \frac{dM_{1i}(u)}{\widehat{Y}_1(u)} \right\}. \end{aligned}$$

To approximate the distribution of the process $U_n(t)$, we define the following Gaussian process

$$\begin{aligned} \widehat{U}_n(t) &= \sqrt{n} \sum_{i \in V} \int_0^t \frac{d\widehat{M}_{2i}(u)G_i}{\widehat{Y}_2^V(u)} \\ &= -\widehat{\sigma}_{21}(t)\widehat{\sigma}_{11}^{-1}(t) \left\{ \sqrt{n} \sum_{i \in V} \int_0^t \frac{d\widehat{M}_{1i}(u)G_i}{\widehat{Y}_1^V(u)} - \sqrt{\frac{n}{N}} \sqrt{N} \sum_{i=1}^N \int_0^t \frac{d\widehat{M}_{1i}(u)G_i}{\widehat{Y}_1(u)} \right\}, \end{aligned}$$

where $G_1, \dots, G_N$ are independent and identically distributed (i.i.d.) standard normal random variables. Theorem 3 establishes that $U_n(t)$ and $\widehat{U}_n(t)$ have the same limiting distribution.

Theorem 3—Conditionally on the data

$$\begin{aligned} &\{N_{1i}(t), Y_{1i}(t) \\ &\quad, N_{2j}(t), Y_{2j}(t); t \in [0 \\ &\quad, \tau]\} \\ &\quad, i=1, \dots, N; j \in V(n), \end{aligned}$$

the random process $\widehat{U}_n(t)$ converges weakly to $\mathcal{W}_2(t)$ on $[0, \tau]$.

It follows from theorem 3 that the distribution of the process $U_n(t)$ can be approximated by the conditional distribution of the process $\hat{U}_n(t)$ given the observed data. The conditional distribution of the process $\hat{U}_n(t)$ given the observed data can be easily simulated by generating a large number of independent Monte Carlo random samples $G_1, \dots, G_N$ from the standard normal distribution.

When constructing confidence bands for Λ_2 and S_2 , it is useful to consider a transformation process (Lin *et al.*, 1994),

$$B(t) = \sqrt{n}g(t)[\varphi(\hat{\Lambda}_2(t)) - \varphi(\Lambda_2(t))],$$

where g is the weight function and φ is a known transformation function, whose derivation φ' is continuous and non-zero in the interval of interest $[\tau_1, \tau_2]$. $B(t)$ is asymptotically equivalent to $\sqrt{n}g(t)\varphi'(\hat{\Lambda}_2(t))U_n(t)$, and, therefore, can be simulated by

$$\widehat{B}(t) = \sqrt{n}g(t)\varphi'(\hat{\Lambda}_2(t))\widehat{U}_n(t).$$

By simulation, let q_α be the boundary value of

$$\Pr\left(\sup_{t \in [t_1, t_2]} |\widehat{B}(t)| > q_\alpha\right) = \alpha.$$

An approximate $(1 - \alpha)$ confidence band for $\varphi(\Lambda_2(t))$ on $[t_1, t_2]$ is given by,

$$\varphi(\hat{\Lambda}_2(t)) \mp n^{-1/2}q_\alpha/g(t).$$

The corresponding confidence bands for $\Lambda_2(t)$ and $S_2(t)$ are then given by,

$$\varphi^{-1}[\varphi(\hat{\Lambda}_2(t)) \mp n^{-1/2}q_\alpha/g(t)]$$

and

$$\exp\{-\varphi^{-1}[\varphi(\hat{\Lambda}_2(t)) \mp n^{-1/2}q_\alpha/g(t)]\}.$$

Typically, we consider $\varphi(t) = \log(t)$ to obtain bands in meaningful ranges and to attain better coverage probabilities. We use $g(t) = \hat{\Lambda}_2(t)/\sigma_{22}^*(t)$ to obtain the equal-precision band.

3. An example

Now we illustrate our method on a two-stage survival data from a medical device postmarketing surveillance study. As discussed earlier in section 1, the postmarked surveillance study is designed to collect information about a marketed device, which can reveal unforeseen adverse events, the actual rate of anticipated adverse events or other information necessary to protect the public health. Physicians can also use this information to evaluate the performance of a marketed device for better patient care.

In August 2002, St Jude Medical conducted an active study to gather information on devices failures and adverse events of five pacing electrode groups. The grouping of the electrodes is based on their insulation, material and fixation technique. A random sample of devices was selected within each electrode group to be in the active study based on the company administrative data. One unique patient can be identified with each device. The corresponding patients and their follow-up physicians were contacted to verify the device status, device failure, associated adverse events and patient death. For pacing electrode, when it is not working properly, no repair is possible; the old electrode has to be replaced by a new one. Because of the potential under-reporting problem, it is important to calibrate the adverse event rate based on the company administrative data by the active validation sample.

For the purpose of illustration, we only present an analysis of group 2 here. Group 2 electrodes are passive fixation, platinum/iridium electrodes with silicone insulation. Table 1 gives a brief summary statistics of the group 2 devices. Two hundred and thirty-six electrodes were included in the active study, and five devices (2.12%) were found to have adverse events, whereas in the company data, 261 devices (0.83%) are recorded with associated adverse events, of a total of 31,438 devices.

Table 2 gives three different estimates of $1000(1 - S(t))$, the expected number of adverse events per 1000 electrodes in t years, where method 1 is based on $\hat{S}_1$, the Kaplan–Meier estimator using the company data only; method 2 is based on $\hat{S}_2$, our proposed estimator combining the two-stage data; and method 3 is based on $\hat{S}_2^V$, the Kaplan–Meier estimator based on the active sample only. It is clear from Table 2 that, except for the first and second year, using the company data only would underestimate the adverse event rate by more than 50%. For instance, the estimated number of adverse events within 5 years is about nine per 1000 based on the company data only, while the corrected estimate is 21 per 1000 using our proposed method based on the two-stage data. Finally, the last column reports the estimate based on the active sample only. It differs from our proposed method except for $t = 1$ and 2 years, which could be explained by the well-known fact that the Kaplan–Meier estimator does not behave well in its far right tail for relatively small samples. As shown by our simulations in section 4, our proposed estimate $\hat{S}_2$ is less biased and more stable in the right tail than the active sample Kaplan–Meier estimator $\hat{S}_2^V$.

4. Simulation study

We conducted some simulations to investigate the finite-sample performance of our proposed estimator $\hat{S}_2$. We also compared it with the Kaplan–Meier estimator $\hat{S}_2^V$ based on the second stage validation sample only.

The following model is considered. For the second stage validation sample, the survival time T_2 is assumed to have an exponential(1) distribution. For the first stage sample, the biased survival time is generated by $T_1 = T_2 + \varepsilon$, where $\varepsilon \sim \text{Uniform}(0, \theta)$. Clearly, a larger θ gives a larger bias. The censoring distribution is assumed to be Exponential(v) with v adjusted to give

a desired censoring rate. The sample sizes for the first and second stage samples are $N = 1000$ and $n = 100$ respectively. All simulations were done using 1000 Monte Carlo samples.

Figure 1 displays the true survival function $S_2(t)$, the simulated mean functions of the Kaplan–Meier estimator $\hat{S}_1(t)$ from the first stage sample, the Kaplan–Meier estimator $\widehat{S}_2^V(t)$ from the second stage sample, and our proposed estimator $\hat{S}_2(t)$ using the two-stage data respectively. As expected, $\hat{S}_1(t)$ is biased, while our proposed estimator $\hat{S}_2(t)$ has effectively corrected the bias. The Kaplan–Meier estimator $\widehat{S}_2^V(t)$ from the second stage sample behaves well except that it shows more bias in its right tail than our proposed estimator $\hat{S}_2(t)$ because it does not make use of any information from the first stage data.

Figure 2 depicts the ratio of the mean-squared errors between $\widehat{S}_2^V(t)$ and $\hat{S}_2(t)$ over time t . It is observed that the mean-squared error of $\widehat{S}_2^V(t)$ is always larger than our proposed estimator $\hat{S}_2(t)$. The improvements are more substantial in the right tail. This is encouraging as the Kaplan–Meier estimator is notorious for its unstable behaviour in the right tail especially under heavy censoring.

We further investigated the differences between $\hat{S}_2(t)$ and $\widehat{S}_2^V(t)$ by examining their bias and variance. Figures 3, 4 and 5 depict the average and variance of $\hat{S}_2(t)$ and $\widehat{S}_2^V(t)$ over 1000 Monte Carlo samples for $\theta = 0.6$ and 0.3 with 70%, 50% and 30% censoring in the second stage sample respectively. The corresponding censoring probabilities for the first stage are 78%, 57% and 34% when $\theta = 0.3$, and 84%, 61% and 38% when $\theta = 0.6$. It is seen that under heavy censoring, say 70% (Figure 3), our estimator $\hat{S}_2(t)$ not only has a smaller variance, but also has a smaller bias in the right tail than $\widehat{S}_2^V(t)$. As the censoring rate decreases, both estimators become essentially unbiased (upper panels of Figures 4 and 5), but our estimator $\hat{S}_2(t)$ still has a smaller variance than $\widehat{S}_2^V(t)$. This is consistent with the asymptotic variance formula obtained in theorem 2.

We also studied the performance of the confidence interval (8) for $S_2(t)$ based on our proposed estimator. Table 3 reports the achieved coverage probability of the confidence interval (8) for different combinations of $S_2(t)$, θ and censoring rate. Each entry in the table is based on 1000 Monte Carlo samples. The sample sizes are $N = 1000$ and $n = 100$. The achieved coverage probability are pretty close to the 95% nominal level in almost all cases except that a larger sample size n may be needed under heavy censoring (70%).

Finally, we conducted a simulation to evaluate the actual coverage probability of the equal-precision bands described in section 2.4. One thousand sets of G_i are generated to construct each confidence band. Table 4 gives the actual coverage probability of the equal-precision band for $S_2(t)$ over the time interval between the 10th percentile and the 70th percentile of the failure time distribution for different combinations of sample sizes and censoring rates. The nominal level is 95% and $\theta = 0.3$. Each entry in the table is based on 1000 Monte Carlo samples. It is seen that for the sample sizes considered, the achieved coverage probability of the confidence band is reasonably close to the nominal level when the censoring rate is moderate (30% or 50%) or the second stage sample size n is large enough. When the censoring rate is high (70%) and the second stage sample size is small, the coverage probability is a little lower than the nominal level.

5. Discussion

We developed a non-parametric method for the estimation of a survival function with right-censored survival data from a two-stage design. It has applications to medical device

postmarked surveillance study as well as epidemiology and clinical trials studies. Our approach does not make any assumption on the type of bias associated with the first stage data. By combining the two-stage data, our proposed estimator of the survival function corrects the bias in the first stage data. In the mean time, it has a smaller asymptotic variance than the Kaplan–Meier estimator based on the second stage data alone. Furthermore, our simulation shows that the proposed estimator has a better small-sample behaviour than the Kaplan–Meier estimator based on the second stage data alone, with a smaller variance and a smaller bias in the right tail of the distribution. Our confidence intervals and confidence bands for the survival function perform well for a moderate validation sample size (say $n = 100$) and a moderate censoring rate (say 50% or less). Our results are derived using the counting process techniques and therefore apply to other types of survival data including right-censored and left-truncated data (cf. Andersen *et al.*, 1993). Finally, the ideas of this paper can be used to study other survival models for two-stage designs. In a sequel, we study the regression analysis of survival data for two-stage designs using proportional hazards and additive risks models.

Acknowledgments

Gang Li's research was supported in part by NIH grant CA78314-03. We thank St Jude Medical Inc. CRMD for providing the electrode postmarket surveillance data, and two referees for their helpful comments.

References

- Andersen, PK.; Borgan, Ø.; Gill, RD.; Keiding, N. Statistical models based on counting processes. Springer Verlag; New York: 1993.
- Barlow, RE.; Bartholomew, DJ.; Bremner, JM.; Brunk, H. Statistical inference under order restrictions. Wiley; New York: 1972.
- Breslow NE, Cain KC. Logistic regression for two-stage case–control data. *Biometrika* 1988;75:11–20.
- Breslow NE, Holubkov R. Weighted likelihood, pseudo-likelihood and maximum likelihood methods for logistic regression analysis of two-stage data. *Statist Med* 1997;16:103–116.
- Brookmeyer R, Yasui Y. Statistical analysis of passive surveillance disease registry data. *Biometrics* 1995;51:831–842. [PubMed: 7548702]
- Carroll RJ, Wand MP. Semiparametric estimation in logistic measurement error models. *J Roy Statist Soc Ser B* 1991;53:573–585.
- Chatterjee N, Chen YH, Breslow NE. A pseudoscore estimator for regression problems with two phase sampling. *J Amer Statist Assoc* 2003;98:158–168.
- Chen YH. Cox regression in cohort studies with validation sampling. *J Roy Statist Soc Ser B* 2002;64:51–62.
- Chen YH, Chen H. A unified approach to regression analysis under double sampling design. *J Roy Statist Soc Ser B* 2000;64:449–460.
- Chernick MR, Poulsen E, Wang Y. Effects of bias adjustment on actuarial survival curves. *Drug Inf J* 2002;36:595–609.
- Cox DR. Regression models and life tables (with discussion). *J Roy Statist Soc Ser B* 1972;34:187–220.
- Flanders TR, Greenland S. Analytic methods for two-stage case–control study and other stratified designs. *Statist Med* 1991;10:739–747.
- Lin DY, Ying Z. Semiparametric analysis of additive risk model. *Biometrika* 1994;81:61–71.
- Lin DY, Ying Z. Semiparametric and nonparametric regression analysis of longitudinal data. *J Amer Statist Assoc* 2001;96:103–113.
- Lin DY, Fleming TR, Wei LJ. Confidence bands for survival curves under the proportional hazards model. *Biometrika* 1994;81:73–81.
- Muller P, Roeder K. A semiparametric model for case–control studies with errors in variables. *Biometrika* 1997;84:523–537.
- Pepe MS, Fleming TR. A nonparametric method for dealing with mismeasured covariate data. *J Amer Statist Assoc* 1991;96:106–113.

- Prentice RL. Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika* 1982;69:331–342.
- Rao RR. The law of large numbers for $D[0, 1]$ -valued random variables. *Theory Probab Appl* 1963;8:70–74.
- Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. *J Amer Statist Assoc* 1994;89:846–866.
- Robins JM, Hsieh F, Newey W. Semiparametric efficient estimation of a conditional density with missing or mismeasured covariates. *J Roy Statist Soc Ser B* 1995;57:409–424.
- Scott AJ, Wild CJ. Fitting regression models to case–control data by maximum likelihood. *Biometrika* 1997;84:57–71.
- Shorack, GR.; Wellner, JA. Empirical processes with applications to statistics. John Wiley and Sons; New York: 1986.
- Tseng, CH. PhD Dissertation. University of California; Los Angeles: 2004. Survival analysis with two stage design studies.
- van der Vaart, AW.; Wellner, JA. Weak convergence and empirical processes. Springer Verlag; New York: 1996.
- Wang CY, Hsu L, Feng ZD, Prentice RL. Regression calibration in failure time regression. *Biometrics* 1997;53:131–145. [PubMed: 9147589]
- White JE. A two stage design for the study of the relationship between a rare exposure and a rare disease. *Amer J Epidemiol* 1982;115:119–128. [PubMed: 7055123]
- Zhou H, Pepe MS. Auxiliary covariate data in failure time regression. *Biometrika* 1995;82:139–149.
- Zhou H, Wang CY. Failure time regression with continuous covariates measured with error. *J Roy Statist Soc Ser B* 2000;62:654–665.

Appendix

Because of the intraclass dependence among the processes $\widehat{\Lambda}_1^V(t)$, $\widehat{\Lambda}_2^V(t)$ and $\widehat{\Lambda}_1(t)$, the standard martingale theory for survival analysis cannot be applied directly to establish the joint weak convergence. We will use the following lemma 4 from empirical process theory to facilitate our development of the large sample properties of the estimators in this article.

Lemma 4

(Lin & Ying, 2001) *Let $H_n(t)$ and $M_n(t)$ be two sequences of bounded process. Suppose that $H_n(t)$ is monotone and converges uniformly to $H(t)$ in probability and that $M_n(t)$ converges to a zero-mean process with continuous sample path. Then*

$$\int_0^t [H_n(s) - H(s)] dM_n(s)$$

converges in probability to zero uniformly in t .

We are now ready to prove the main results of the paper.

Proof of theorem 1

Note that

$$\begin{aligned}
\sqrt{n}(\widehat{\Lambda}_1^V(t) - \Lambda_1(t)) &= \sqrt{n} \int_0^t \left(\frac{d\widehat{N}_1^V(u)}{\widehat{Y}_1^V(u)} - \int_0^t d\Lambda_1(u) \right) \\
&= \frac{1}{\sqrt{n}} \int_0^t \frac{d\widehat{N}_1^V(u) - \widehat{Y}_1^V(u) d\Lambda_1(u)}{\widehat{Y}_1^V(u)/n} \\
&= \frac{1}{\sqrt{n}} \int_0^t \frac{d\widehat{M}_1^V(u)}{\widehat{Y}_1^V(u)/n} \\
&= \frac{1}{\sqrt{n}} \int_0^t \frac{d\widehat{M}_1^V(u)}{y_1(u)} + \frac{1}{\sqrt{n}} \int_0^t \left(\frac{1}{\widehat{Y}_1^V(u)/n} - \frac{1}{y_1(u)} \right) d\widehat{M}_1^V(u).
\end{aligned}$$

Under condition 1, $\widehat{M}_1^V / \sqrt{n}$ converges weakly to a zero-mean Gaussian process by van der Vaart & Wellner (1996, Example 2.11.16). Under conditions 2 and 3, $1/(\widehat{Y}_1^V(u)/n)$ is monotone, bounded and converges uniformly to $1/y_1(u)$. It follows from lemma 4 that

$$\frac{1}{\sqrt{n}} \int_0^t \left(\frac{1}{\widehat{Y}_1^V(u)/n} - \frac{1}{y_1(u)} \right) d\widehat{M}_1^V(u) \xrightarrow{P} 0$$

uniformly in t , as $n \rightarrow \infty$. Thus, as $n, N \rightarrow \infty$,

$$\begin{aligned}
&\sqrt{N}(\widehat{\Lambda}_1^V(t) - \Lambda_1(t)) \\
&= \frac{\sqrt{N}}{\sqrt{n}} \frac{1}{\sqrt{n}} \sum_{i=1}^N \int_0^t \frac{dM_{1i}(u) \delta_i}{y_1(u)} \\
&\quad + o_p(1) \text{ uniformly in } t,
\end{aligned} \tag{9}$$

with $\delta_i = I(i \in V(n))$. Similarly, we have

$$\begin{aligned}
&\sqrt{N}(\widehat{\Lambda}_2^V(t) - \Lambda_2(t)) \\
&= \frac{\sqrt{N}}{\sqrt{n}} \frac{1}{\sqrt{n}} \sum_{i=1}^N \int_0^t \frac{dM_{2i}(u) \delta_i}{y_2(u)} \\
&\quad + o_p(1) \text{ uniformly in } t,
\end{aligned} \tag{10}$$

as $n, N \rightarrow \infty$, and

$$\begin{aligned}
&\sqrt{N}(\widehat{\Lambda}_1(t) - \Lambda_1(t)) \\
&= \frac{1}{\sqrt{N}} \sum_{i=1}^N \int_0^t \frac{dM_{1i}(u)}{y_1(u)} \\
&\quad + o_p(1) \text{ uniformly in } t,
\end{aligned} \tag{11}$$

as $N \rightarrow \infty$.

We define the joint stochastic processes as

$$W_N(t) = N^{1/2} [\widehat{\Lambda}_1^V(t) - \Lambda_1(t), \widehat{\Lambda}_2^V(t) - \Lambda_2(t), \widehat{\Lambda}_1(t) - \Lambda_1(t)].$$

Because each element of W_N is essentially a sum of N i.i.d. random variables, by the multi-variate central limit theorem, the finite-dimensional distributions of $W_N(t)$ converge to those of $\mathcal{W}(t)$ asymptotically. Because each element of W_N is a square-integrable martingale with respect to its own marginal filtration, the tightness follows directly from standard martingale proof using Lenglart's Inequality (Andersen *et al.*, 1993). Hence, $W_N(t)$ converges to a zero-mean Gaussian random field in $\mathcal{D}[0, \tau]^3$.

To prove the consistency of estimated variances and covariance, we have

$$\begin{aligned} \widehat{\sigma}_{12}(t_1, t_2) &= \frac{1}{n} \sum_{i \in V(n)} \int_0^{t_1} \frac{d\widehat{M}_{1i}(u)}{\widehat{Y}_1^V(u)/n} \int_0^{t_2} \frac{d\widehat{M}_{2i}(u)}{\widehat{Y}_2^V(u)/n} \\ &= \frac{1}{n} \sum_{i \in V(n)} \left[\int_0^{t_1} \frac{d\widehat{M}_{1i}(u)}{\widehat{Y}_1^V(u)/n} \int_0^{t_2} \frac{d\widehat{M}_{2i}(u)}{\widehat{Y}_2^V(u)/n} - \int_0^{t_1} \frac{dM_{1i}(u)}{y_1(u)} \int_0^{t_2} \frac{dM_{2i}(u)}{y_2(u)} \right] \\ &\quad + \frac{1}{n} \sum_{i \in V(n)} \int_0^{t_1} \frac{dM_{1i}(u)}{y_1(u)} \int_0^{t_2} \frac{dM_{2i}(u)}{y_2(u)}. \end{aligned}$$

By the uniform convergence of $\widehat{\Lambda}_1^V(t)$, $\widehat{\Lambda}_2^V(t)$, $\widehat{Y}_1^V(t)$ and $\widehat{Y}_2^V(t)$, we have

$$\frac{1}{n} \sum_{i \in V(n)} \left| \int_0^{t_1} \frac{d\widehat{M}_{1i}(u)}{\widehat{Y}_1^V(u)/n} \int_0^{t_2} \frac{d\widehat{M}_{2i}(u)}{\widehat{Y}_2^V(u)/n} - \int_0^{t_1} \frac{dM_{1i}(u)}{y_1(u)} \int_0^{t_2} \frac{dM_{2i}(u)}{y_2(u)} \right| \xrightarrow{P} 0$$

uniformly. Furthermore, with the assumption that

$$E \left[\sup_{t_1, t_2 \in [0, \tau]} \int_0^{t_1} \frac{dM_{1i}(u)}{y_1(u)} \int_0^{t_2} \frac{dM_{2i}(u)}{y_2(u)} \right] < \infty$$

we can conclude the uniform convergence of the covariance estimator (Rao, 1963):

$$\frac{1}{n} \sum_{i \in V(n)} \int_0^{t_1} \frac{dM_{1i}(u)}{y_1(u)} \int_0^{t_2} \frac{dM_{2i}(u)}{y_2(u)} \xrightarrow{a.s.} E \left[\int_0^{t_1} \frac{dM_{1i}(u)}{y_1(u)} \int_0^{t_2} \frac{dM_{2i}(u)}{y_2(u)} \right] \equiv \sigma_{12}(t)$$

uniformly. Therefore,

$$\widehat{\sigma}_{12}(t_1, t_2) \xrightarrow{P} \sigma_{12}(t_1, t_2)$$

uniformly. The uniform consistency of $\widehat{\sigma}_{11}(t)$ and $\widehat{\sigma}_{22}(t)$ can be established along the same lines.

Proof of theorem 2

Because $\widehat{\sigma}_{11}$ and $\widehat{\sigma}_{12}$ converge uniformly to σ_{11} and σ_{12} , and

$$\begin{aligned} W_N(t) = & N^{1/2} [\{ \widehat{\Lambda}_1^V(t) \\ & - \Lambda_1(t) \} \\ & , \{ \widehat{\Lambda}_2^V(t) \\ & - \Lambda_2(t) \} \\ & , \{ \widehat{\Lambda}_1(t) - \Lambda_1(t) \}] \end{aligned}$$

converges weakly to a zero-mean Gaussian process $\mathcal{W}$ (theorem 1), the weak convergence of $U_n(t)$ follows directly by the Slutsky's theorem and the Continuous Mapping theorem (van der Vaart & Wellner, 1996):

$$\begin{aligned} U_n(t) &\equiv \sqrt{n}(\widehat{\Lambda}_2(t) - \Lambda_2(t)) \\ &= \sqrt{n}(\widehat{\Lambda}_2^V(t) - \Lambda_2(t)) \\ -\widehat{\sigma}_{12}(t)\widehat{\sigma}_{11}^{-1}(t) &\left[\sqrt{n}(\widehat{\Lambda}_1^V(t) - \Lambda_1(t)) - \sqrt{\frac{n}{N}}\sqrt{N}(\widehat{\Lambda}_1(t) - \Lambda_1(t)) \right] \\ &\xrightarrow{d} \mathcal{W}_2^V(t) - \sigma_{12}(t)\sigma_{11}^{-1}(t)[\mathcal{W}_1^V(t) - \rho^{1/2}\mathcal{W}_1(t)] \\ &\equiv \mathcal{W}_2(t). \end{aligned}$$

It can be verified that the covariance function of $\mathcal{W}_2(t)$ is given by

$$\begin{aligned} \xi(s, t) &= \text{cov}(\mathcal{W}_2(s) \\ &\quad , \mathcal{W}_2(t)) \\ &= \sigma_{22}(s \wedge t) \\ &\quad - (1 - \rho)\sigma_{12}(t)\sigma_{11}^{-1}(s \vee t)\sigma_{12}(s) \end{aligned}$$

for $0 \leq s, t \leq \tau$. By theorem 1, the covariance of function $\xi(s, t)$ can be consistently estimated by

$$\begin{aligned} \xi(s, t) &= \widehat{\sigma}_{22}(s \wedge t) \\ &\quad - (1 - \frac{n}{N})\widehat{\sigma}_{12}(t)\widehat{\sigma}_{11}^{-1}(s \vee t)\widehat{\sigma}_{12}(s) \text{ for } 0 \leq s, t \leq \tau. \end{aligned}$$

As a special case, the variance of function $\sigma_{22}^*(t) = \xi(t, t)$ can be consistently estimated by

$$\sigma_{22}^*(t) = \widehat{\sigma}_{22}(t) - \left(1 - \frac{n}{N}\right) \widehat{\sigma}_{12}(t) \sigma_{11}^{-1}(t) \widehat{\sigma}_{12}(t) \text{ for any } t \in [0, \tau].$$

Proof of theorem 3

Let

$$U_n^*(t) = n^{-1/2} \sum_{i \in V} \int_0^t \frac{dM_{2i}(u)}{y_2(u)} - \sigma_{21}(t) \sigma_{11}^{-1}(t) \left(n^{-1/2} \sum_{i \in V} \int_0^t \frac{dM_{1i}(u)}{y_1(u)} - \sqrt{\frac{n}{N}} N^{-1/2} \sum_{i=1}^N \int_0^t \frac{dM_{1i}(u)}{y_1(u)} \right),$$

and

$$\widehat{U}_n^*(t) = n^{-1/2} \sum_{i \in V} \int_0^t \frac{dM_{2i}(u)G_i}{y_2(u)} - \sigma_{21}(t) \sigma_{11}^{-1}(t) \left(n^{-1/2} \sum_{i \in V} \int_0^t \frac{dM_{1i}(u)G_i}{y_1(u)} - \sqrt{\frac{n}{N}} N^{-1/2} \sum_{i=1}^N \int_0^t \frac{dM_{1i}(u)G_i}{y_1(u)} \right).$$

Using lemma 4 and the uniform consistency of $\bar{Y}_1$, $\bar{Y}_2$, $\hat{\sigma}_{21}$ and $\hat{\sigma}_{11}$, we can verify that $U_n^*(t) = U_n(t) + o_p(1)$ with similar arguments in the proof of theorem 1. Therefore, $U_n^*(t)$ also converges unconditionally to the zero-mean Gaussian process $\mathcal{W}_2$ (see theorem 2). It follows that $\widehat{U}_n^*(t)$ converges weakly in probability to $\mathcal{W}_2(t)$ conditional on the data, according to the conditional multiplier central limit theorem (van der Vaart & Wellner, 1996, theorem 2.9.6). To complete the proof, we would show in the following that

$$\| \widehat{U}_n^*(t) - \widehat{U}_n(t) \| \xrightarrow{d} 0$$

uniformly in t .

It is not difficult to show that

$$\begin{aligned} \| \widehat{U}_n^*(t) - \widehat{U}_n(t) \| &\leq \| U_{n1}^d(t) \| \\ &+ \| U_{n2}^d(t) \| \\ &+ \| U_{n3}^d(t) \| \end{aligned}$$

where

$$\begin{aligned}
\|U_{n1}^d(t)\| &= \|n^{-1/2} \sum_{i \in V} \int_0^t \frac{dM_{2i}(u)G_i}{y_2(u)} - n^{-1/2} \sum_{i \in V} \int_0^t \frac{d\widehat{M}_{2i}(u)G_i}{\widehat{Y}_2^V(u)/n}\| \\
\|U_{n2}^d(t)\| &= \|n^{-1/2} \sigma_{12} \sigma_{11}^{-1} \sum_{i \in V} \int_0^t \frac{dM_{1i}(u)G_i}{y_1(u)} - n^{-1/2} \widehat{\sigma}_{12} \widehat{\sigma}_{11}^{-1} \sum_{i \in V} \int_0^t \frac{d\widehat{M}_{1i}(u)G_i}{\widehat{Y}_1^V(u)/n}\| \\
\|U_{n3}^d(t)\| &= \left\| \sqrt{\frac{n}{N}} N^{-1/2} \sigma_{12} \sigma_{11}^{-1} \sum_{i=1}^N \int_0^t \frac{dM_{1i}(u)G_i}{y_1(u)} \right. \\
&\quad \left. - \sqrt{\frac{n}{N}} N^{-1/2} \widehat{\sigma}_{12} \widehat{\sigma}_{11}^{-1} \sum_{i=1}^N \int_0^t \frac{d\widehat{M}_{1i}(u)G_i}{\widehat{Y}_1(u)/N} \right\|.
\end{aligned}$$

We first show that $\|U_{n1}^d(t)\|$ converges to zero in probability

$$\begin{aligned}
\|U_{n1}^d(t)\| &\leq \|n^{-1/2} \sum_{i \in V} \int_0^t \left(\frac{1}{y_2(u)} - \frac{1}{\widehat{Y}_2(u)/n} \right) G_i dM_{2i}(u)\| \\
&\quad + \|n^{-1/2} \sum_{i \in V} \int_0^t \frac{Y_{2i}(u)G_i}{\widehat{Y}_2^V(u)/n} d(\widehat{\Lambda}_2^V(u) - \Lambda_2(u))\|.
\end{aligned}$$

Applying the conditional multiplier central limit theorem (van der Vaart & Wellner, 1996, theorem 2.9.6), we have $n^{-1/2} \sum_{i \in V} M_{2i}(u)G_i$ converges to a zero-mean Gaussian process. It follows from lemma 4 and conditions 2 and 3 that the first term converges to zero in probability uniformly in t . Under conditions 2 and 3, there exists a real number K and some n_0 such that $\widehat{Y}_2^V(u)/n > 1/K$ for all $n > n_0$. Then for the second term, we have

$$\begin{aligned}
&\|n^{-1/2} \sum_{i \in V} \int_0^t \frac{Y_{2i}(u)G_i}{\widehat{Y}_2^V(u)/n} d(\widehat{\Lambda}_2^V(u) - \Lambda_2(u))\| \\
&\leq K \left\| \int_0^t \sum_{i \in V} \frac{Y_{2i}(u)G_i}{n} d(\sqrt{n}(\widehat{\Lambda}_2^V(u) - \Lambda_2(u))) \right\|,
\end{aligned}$$

for $n > n_0$. Let us denote $G_i^+ = \max(G_i, 0)$ and $G_i^- = \max(-G_i, 0)$. It is easy to verify that

$$\sum_{i \in V} (Y_{2i}(u)G_i)/n$$

is the difference of two monotone functions,

$$\begin{aligned}
&\sum_{i \in V} (Y_{2i}(u)G_i^+)/n \text{ and } \sum_{i \in V} (Y_{2i}(u)G_i^-)/n,
\end{aligned}$$

which converge to the same limit. Because $\sqrt{n}(\widehat{\Lambda}_2^V(u) - \Lambda_2(u))$ converges to a zero-mean process with continuous path (theorem 1), we can again apply lemma 4 and conclude that the second term converges to zero in probability uniformly in t . It can be shown in the same manner that $\|U_{n2}^d(t)\|$ and $\|U_{n3}^d(t)\|$ also converge to zero in probability uniformly in t .

Fig. 1.

Estimated survival functions $\hat{S}_1(t)$ (thick solid line), $\hat{S}_2^V(t)$ (dashed line), $\hat{S}_2(t)$ (dotted line) and the true underlying survival function $S_2(t)$ (thin solid line).

Fig. 2.

Ratio of the mean-squared errors $[\text{MSE}(\hat{S}_2^V(t))/\text{MSE}(\hat{S}_2(t))]$ over time between our proposed estimator $\hat{S}_2(t)$ and the Kaplan–Meier estimator $\hat{S}_2^V(t)$ based on the second stage validation sample alone.

Fig. 3. Estimated survival functions and their corresponding variance functions with 70% censoring: the upper two panels are the estimated survival functions $\hat{S}_2(t)$ (dashed line $--$) and $\hat{S}_2^V(t)$ (dotted line $\cdots$) together with the true survival function $S_2(t)$ (solid line $—$) and the lower two panels are the corresponding variance functions.

Fig. 4. Estimated survival functions and their corresponding variance functions with 50% censoring: the upper two panels are the estimated survival functions $\hat{S}_2(t)$ (dashed line $--$) and $\hat{S}_2^V(t)$ (dotted line $\cdots$) together with the true survival function $S_2(t)$ (solid line $—$) and the lower two panels are the corresponding variance functions.

Fig. 5. Estimated survival functions and their associated variance functions with 30% censoring: the upper two panels are the estimated survival functions $\hat{S}_2(t)$ (dashed line ---) and $\widehat{S}_2^V(t)$ (dotted line $\cdots$) together with the true survival function $S_2(t)$ (solid line —) and the lower two panels are the corresponding variance functions.

Table 1

Summary of group 2 devices in active study and company database

Source	<i>N</i>	Number of adverse events	Per cent	Follow-up duration [mean (SD) in years]
Company data	31,483	261	0.83	4.76 (2.55)
Active study	236	5	2.12	4.09 (2.21)

Table 2Estimated number of adverse events per 1000 electrodes in t years

Time t (in years)	Method 1 (company data only)	Method 2 (two-stage data)	Method 3 (active sample only)
1.0	4 ± 0.00029	4 ± 0.0380	4 ± 0.0380
2.0	5 ± 0.00036	10 ± 0.0386	10 ± 0.0909
3.0	7 ± 0.00046	11 ± 0.0386	10 ± 0.0909
4.0	8 ± 0.00055	19 ± 0.1224	16 ± 0.1740
5.0	9 ± 0.00066	20 ± 0.1247	24 ± 0.2889
6.0	9 ± 0.00072	21 ± 0.1247	24 ± 0.2889

Table 3

Achieved coverage probability of the 95% confidence interval for $S_2(t)$

$\theta = 0.3$		$\theta = 0.6$				
Censoring rate		Censoring rate				
$S_2(t)$	70%	50%	30%	70%	50%	30%
0.80	93.2	93.6	94.0	92.5	93.0	93.8
0.50	91.4	93.0	93.6	92.6	92.5	93.3
0.20	85.7	92.8	95.0	87.4	91.4	94.9

The sample sizes are $N = 1000$ and $n = 100$. All values are given as percentage.

Table 4

Achieved coverage probability of the 95% confidence bands

<i>N</i>	<i>n</i>	Censoring rate		
		30%	50%	70%
1000	100	93.3	91.0	88.2
1000	200	93.7	92.4	88.7
1000	300	94.0	94.3	91.6
1000	400	95.1	94.5	92.6
1000	500	95.2	94.4	94.3

All values are given as percentage.