CONFIDENCE BANDS FOR SURVIVAL CURVES FROM OUTCOME-DEPENDENT STRATIFIED SAMPLES

TAKUMI SAEGUSA AND PETER NANDORI

DEPARTMENT OF MATHEMATICS, UNIVERSITY OF MARYLAND AND
DEPARTMENT OF MATHEMATICS, YESHIVA UNIVERSITY

Abstract. We consider the construction of confidence bands for survival curves under the outcome-dependent stratified sampling. A main challenge of this design is that data are a biased dependent sample due to stratification and sampling without replacement. Most literature on regression approximates this design by Bernoulli sampling but variance is generally overestimated. Even with this approximation, the limiting distribution of the inverse probability weighted Kaplan-Meier estimator involves a general Gaussian process, and hence quantiles of its supremum is not analytically available. In this paper, we provide a rigorous asymptotic theory for the weighted Kaplan-Meier estimator accounting for dependence in the sample. We propose the novel hybrid method to both simulate and bootstrap parts of the limiting process to compute confidence bands with asymptotically correct coverage probability. Simulation study indicates that the proposed bands are appropriate for practical use. A Wilms tumor example is presented.

Keywords. confidence band, Gaussian process, right censoring, sampling without replacement, stratified sampling, survival curve.

1. Introduction

Outcome-dependent sampling is a widely used sampling method in epidemiologic studies to study association between exposure and survival events. This design oversamples cases and collects a random sample of controls from the study cohort. This sampling method is highly cost-effective in conducting large scale studies because covariates are measured only for much smaller subsamples from the full cohort. Examples include case-cohort design Prentice (1986) and case-control design Prentice and Pyke (1979). Instead of simple random sampling of controls, outcome-dependent stratified sampling conducts stratified sampling to further reduce the cost and effectively collect important exposures (see e.g. exposure stratified case-cohort study Borgan et al. (2000) and stratified case-control study White (1986)).

Because loss in statistical efficiency is generally small, this design has been successfully adopted to study various statistical problems with the help of the inverse probability weighting. Most applications are censored regression such as the accelerated failure time model, Nan et al. (2006, 2009), the additive hazards model Kulich and Lin (2000), the Cox proportional hazards model (Prentice, 1986; Self and Prentice, 1988), and the transformation model Lu and Tsiatis (2006); Kong et al. (2006); Zeng and Lin (2014) for different types of data such as clustered correlated data Moger et al. (2008), interval censored data Li and Nan (2011); Saegusa and Wellner (2013); Zhou et al. (2017, 2018), and data with competing risks Sø rensen and Andersen (2000); Sun et al. (2004); Kang and Cai (2009).

Despite the extensive research on censored regression, the issue of confidence bands for survival curves has been overlooked because of the challenging probabilistic structure of the outcome dependent stratified design. Randomness in this design comes from two sources: (1) sampling from the infinite population and (2) subsequent sampling from strata without replacement. The resultant sample is a dependent biased sample due to stratification and sampling without replacement. A valid confidence band must address these qualitatively different types of randomness simultaneously. In the regression setting, most research avoids the challenging issue of dependence by assuming Bernoulli sampling where data are treated as an independent sample with missing variables. Variance estimation for this i.i.d. sample is then straightforward with weighted bootstrap Ma and Kosorok (2005) even for complex semiparametric models Li and Nan (2011); Zhou et al. (2017, 2018). A price of approximation by Bernoulli sampling is overestimation of asymptotic variance Breslow and Wellner (2007) resulting in a statistically conservative conclusion in regression settings. Approximate quantification of uncertainty, however, invalidates confidence bands with correct coverage probability in our problem. Moreover, it is not clear how weighted bootstrap behaves in our setting. The weighted bootstrap is only valid for the average of zero-mean variables in the i.i.d. setting (see sections

2.9 and 3.6 of van der Vaart and Wellner (1996)). If applied to nonzero-mean variables, it estimates the second moment rather than variance. In our setting, data do not have the i.i.d. structure and the Kaplan-Meier estimator of survival curves Kaplan and Meier (1958) is a function of nonzero-mean averages.

A statistical challenge in the outcome-dependent stratified designs is not only the difficulty in asymptotic theory for a biased dependent sample. In the i.i.d. setting, the basic idea for confidence bands is to compute quantiles of the supremum of the absolute difference between the Kaplan-Meier estimator and the true function over an interval. The standard method for this purpose is to obtain analytical expressions of quantiles of the relatively simple limiting distribution. Various methods along this line have been proposed Borgan and Liestø l (1990); Gillespie and Fisher (1979); Hall and Wellner (1980); Hollander et al. (1997); Nair (1984) with different transformations of the Kaplan-Meier estimator for better coverages in the finite sample. Different transformations in fact reduce to the supremum of either well-known Brownian motion or Brownian bridge processes thanks to the celebrated martingale theory. Their quantiles can be analytically obtained with ease. In our setting, however, inverse probability weights depend on the survival event which is not predictable. The lack of a martingale structure then yields a complicated limiting Gaussian process with unknown parameters. Because this process cannot be reduced to other well-known processes, quantiles of its supremum are not analytically available. In fact, even finding a tight bound on the suprema of general Gaussian processes is an important research question in probability theory Chernozhukov et al. (2014); Harper (2013). An alternative method would be to bootstrap the supremum of absolute difference explored by Akritas (1986) and Lo and Singh (1986) in the analysis of the i.i.d. sample. For stratified samples, Bickel and Krieger (1989) applied the specialized bootstrap method for a finite population sampling to data without censoring. These bootstrap methods focus on randomness either from sampling from the infinite population or finite population sampling from strata. To the best of our knowledge, there is no valid bootstrap procedure

that simultaneously accounts for both randomness in our setting. Our statistical challenge is to compute quantiles of the supremum of the general Gaussian process without analytical computation nor bootstrap.

In this paper, we study nonparametric estimation of survival functions from the outcome-dependent stratified samples with right censored data. The estimator considered is the inverse probability weighted Kaplan-Meier estimator originally proposed by Amato (1988) for the analysis of the sample from heterogeneous populations. We provide a rigorous asymptotic theory to derive the limiting distribution of our estimator in our design. We then propose a novel procedure to construct simultaneous confidence bands for the survival curves. We show that the proposed bands have the correct coverage probability asymptotically. We illustrate our methodology in a simulation study and a data example.

The weighted Kaplan-Meier estimator has been studied in various settings with non-independent data. Existing methods focused on variance estimation at a fixed time point but have never considered confidence bands due to its theoretical difficulty. Amato (1988) and Williams (1995) who studied correlated survival data, both proposed the same variance estimator based on Greenwood's formula Greenwood (1926). The validity of Greenwood's formula is related to the martingale structure as seen in Hall and Wellner (1980) and Shorack and Wellner (1986), and hence this variance estimator is not consistent in our setting. Rebora and Valsecchi (2016) studied a general complex sampling design including our setting but their limiting distribution is different from our result. Winnett and Sasieni (2002) and Galimberti et al. (2002) studied the matched case-control design and proposed the plug-in variance estimator and a bootstrap method. In contrast to the previous work, our asymptotic theory is new and, more importantly, our proposed confidence band is the first method for survival curves under outcome-dependent stratified sampling.

Instead of nonparametric estimation of survival curves, confidence bands for baseline survival curves were proposed in the case-cohort study Huang (2014) and

nested case-control study Cai and Zheng (2013) by exploiting the regression settings with weighted bootstrap Ma and Kosorok (2005). Both research considered sampling without replacement but treated dependent data as i.i.d. samples for different reasoning. Huang (2014) used the fact that although the entire sample is dependent, the selected sample from simple random sampling is i.i.d. conditionally on the selection status. As combining non-selected cases with a simple random sample creates a biased sample in the case-cohort study, stratification in our setting breaks the i.i.d. structure in our data. Cai and Zheng (2013) used the equivalence of Bernoulli sampling (i.e., selections are all independent) and sampling without replacement because inverse probability weighting with estimated weights in the former design has the same asymptotic variance in the latter. Using independence from Bernoulli sampling, they used weighted bootstrap with the theoretical support from the uniform law of large numbers in the i.i.d. setting (see e.g. Pollard (1984); van der Vaart and Wellner (1996)). In our setting, we cannot use the same argument because our data are not independent nor identically distributed.

This paper is organized as follows. In Section 2, we provide a formal description of the outcome-dependent stratified design and introduce the weighted Kaplan-Meier estimator. Its weak convergence is presented in Section 3. We propose our method to construct a simultaneous confidence band for the survival functions in Section 4. Additional topics of variance estimation and applications to the exposure stratified case-cohort study are discussed in Section 5. The finite sample performance of the proposed method is evaluated through simulation studies in Section 6. The proposed method is applied to the national Wilms tumor study in Section 7. We conclude our paper in Section 8. Proofs are collected in the supplementary material to this paper.

2. Sampling and Estimator

Let T be a failure time and C be a censoring time. For right censored data, we observe a censored failure time $\tilde{T} = \min\{T, C\}$ and a failure indicator $\Delta = I(T \leq$

C), where $I(E)$ is the indicator of the event E. The parameter of interest is the survival functions at each level of the exposure status X given by

$$S(t|x) = P(T > t|X = x).$$

Here the exposure status X is a discrete variable. In the outcome-dependent stratified design, we observe $O_i = (\tilde{T}_i, \Delta_i, U_i)$ for all n subjects in the entire cohort as the i.i.d. sample from the infinite population. The variable $U \in \mathcal{U}$ is an auxiliary variable useful in creating strata for controls based on disjoints subsets $\mathcal{U}_j$s of $\mathcal{U}$ with $\mathcal{U} = \cup_{j=2}^{J} \mathcal{U}_j$. Each subject is classified into one of J strata consisting of all cases $\mathcal{S}_1 = \{O : \Delta = 1\}$ and stratified controls

$$\mathcal{S}_2 = \{O : \Delta = 0, U \in \mathcal{U}_2\}, \ldots, \mathcal{S}_J = \{O : \Delta = 0, U \in \mathcal{U}_J\},$$

based on the values of O_i. To obtain additional variables V, subsamples of size m_j are sampled from each stratum $\mathcal{S}_j$ of size $n_j = |\mathcal{S}_j|$ without replacement. Since all cases are selected, $m_0 = n_0$. The exposure status X is a part of V, for unweighted Kaplan-Meier estimators suffice if X is a part of U. For a concrete example, consider giving a physical exam to participants selected from stratified sampling by a certain disease, age and residence. In this case, U is a vector of age and country of residence to create strata $\mathcal{S}_2$ as controls aged 20-29 in North America, $\mathcal{S}_3$ as controls aged 30-39 in Europe, and so on. Once giving physical exams, lab results and medical history are obtained as V. Among these, a certain gene is selected as the exposure X to study the relationship between the gene and time T to the disease onset.

We denote sampling indicator for the ith subject by $\xi_i \in \{0, 1\}$ and the corresponding sampling probability by

$$\pi_i = P(\xi_i = 1|O_k, k = 1, \ldots, n) = m_j/n_j \quad \text{if } O_i \in \mathcal{S}_j.$$

We denote the stratum membership probabilities by $\nu_j = P(O \in \mathcal{S}_j)$. The observed data are $O_i^\pi, i = 1, \ldots, n$, where $O_i^\pi = (\tilde{T}_i, \Delta_i, U_i, \xi_i, V_i)$ if the ith subject is selected (i.e., $\xi_i = 1$) and $O_i^\pi = (\tilde{T}_i, \Delta_i, U_i, \xi_i)$ otherwise. Note that

$O_i = (\tilde{T}_i, \Delta_i, U_i), i = 1, \ldots, n$, are i.i.d. but the observed data $O_i^\pi, i = 1, \ldots, n$, are not because dependence is induced through the sampling indicators ξ_i.

The proposed estimator builds on the estimation of the cumulative hazard function. This estimation reduces to estimating probabilities that one experiences the event before or at time t and that one is at risk at time t, respectively. Let $N_i(t,x) = \Delta_i I(T_i \le t, X = x)$ and $Y_i(t,x) = I(\tilde{T}_i \ge t, X = x)$ be the counting process and at-risk indicator for the ith subject with exposure status $X = x$. The inverse probability weighted averages of these processes

$$\overline{N}_n^\pi(t,x) = \frac{1}{n}\sum_{i=1}^n \frac{\xi_i}{\pi_i} N_i(t,x), \quad \overline{Y}_n^\pi(t,x) = \frac{1}{n}\sum_{i=1}^n \frac{\xi_i}{\pi_i} Y_i(t,x),$$

reliably estimate $\mathcal{N}(t,x) = P(T \le t, \Delta = 1, X = x)$ and $\mathcal{Y}(t,x) = P(\tilde{T} \ge t, X = x)$. The cumulative hazard function $\Lambda(t|x) = \int_0^t \{\mathcal{Y}(u,x)\}^{-1} \mathcal{N}(du,x)$ at level x is then estimated by the plug-in estimator given by

$$\hat{\Lambda}_n(t|x) = \int_0^t \frac{1}{\overline{Y}_n^\pi(u,x)} \overline{N}_n^\pi(du,x).$$

If complete data are available and inverse probability weights are removed, this estimator reduces to the well-known Nelson-Aalen estimator. Note that multiplying and dividing by $P(X = x)$ within the integral of $\Lambda(t|x)$ does not change $\Lambda(t|x)$ and hence this can be interpreted as the conditional cumulative hazard function given $X = x$.

The survival function is the product limit integral of the cumulative hazard function. In fact, the Kaplan–Meier estimator is the product limit integral of the Nelson-Aalen estimator. Following the same idea, we obtain our estimator as the product limit integral of $\hat{\Lambda}_n$ given by

$$\hat{S}_n(t|x) = \prod_{0<u\le t} \left\{1 - \Delta_u \hat{\Lambda}_n(u|x)\right\} = \prod_{0<u\le t} \left\{1 - \frac{\Delta_u \overline{N}_n^\pi(u,x)}{\overline{Y}_n^\pi(u,x)}\right\},$$

where $\Delta_x f(x,y)$ is the jump $f(x,y)-f(x-,y)$ of f at x for a fixed y throughout the paper. This estimator is easily computed from standard software for the Kaplan–Meier estimator with weights ξ_i/π_i.

3. Limiting Distributions

To construct confidence bands, we first compute limiting processes of $\hat{\Lambda}_n$ and $\hat{S}_n$. Results in this section and Section 4 are presented in a general form in the sense that we allow strata $\mathcal{S}_j$s to be arbitrarily formed by available variables for all n subjects with different probabilities. Note that the definitions of $\overline{\mathcal{N}}_n^{\pi}$ and $\overline{\mathcal{Y}}_n^{\pi}$ also change accordingly. For example, one can consider an outome independent stratum $\mathcal{S}_j = \{O : U \in \mathcal{U}_k\}$ or stratified sampling of cases with $\mathcal{S}_j = \{O : \Delta = 1, U \in \mathcal{U}_k\}$ for some k. This generality allows us to extend our results to the exposure stratified case-cohort study discussed in Section 5.

To obtain limiting processes, we assume

Condition 1. *(a) Survival time T and censoring time C are conditionally independent given X.*

(b) Sampling probabilities m_j/n_j converges to a constant $p_j > 0$ as $n \to \infty$.

(c) There exists $\tau > 0$ such that $P(\tilde{T} \leq \tau, X = x) < 1$ for all levels $X = x$.

The first condition is standard with complete data. Sampling fractions m_j/n_j are at the disposal of the study investigator and it is natural to assume the existence of their limits as in Breslow and Wellner (2007). The third condition concerns a range of confidence bands.

Because $\hat{\Lambda}_n$ and $\hat{S}_n$ are functions of $\overline{N}_n^{\pi}$ and $\overline{Y}_n^{\pi}$, we first compute their limiting distributions.

Lemma 1. *Under Condition 1, $n^{1/2}\{\overline{N}_n^{\pi}(\cdot,x) - \mathcal{N}(\cdot,x), \overline{Y}_n^{\pi}(\cdot,x) - \mathcal{Y}(\cdot,x)\}$ weakly converges to Gaussian processes $\{\mathbb{G}_{\pi}^{\mathcal{N}}(\cdot,x), \mathbb{G}_{\pi}^{\mathcal{Y}}(\cdot,x)\}$ in $(D[0,\tau])^2$ where $D[0,\tau]$ is*

the class of càdlàg functions on $[0,\tau]$ *equipped with the uniform norm and each process is a linear combination of zero-mean Gaussian processes of the form*

$$\begin{aligned}\mathbb{G}_\pi^{\mathcal{N}} &= \mathbb{G}^{\mathcal{N}} + \sum_{j=1}^{J} \left\{ \frac{\nu_j(1-p_j)}{p_j} \right\}^{1/2} \mathbb{G}_j^{\mathcal{N}}, \\ \mathbb{G}_\pi^{\mathcal{Y}} &= \mathbb{G}^{\mathcal{Y}} + \sum_{j=1}^{J} \left\{ \frac{\nu_j(1-p_j)}{p_j} \right\}^{1/2} \mathbb{G}_j^{\mathcal{Y}}.\end{aligned}$$

Here pairs of Gaussian processes $(\mathbb{G}^{\mathcal{N}}, \mathbb{G}^{\mathcal{Y}}), (\mathbb{G}_1^{\mathcal{N}}, \mathbb{G}_1^{\mathcal{Y}}), \ldots, (\mathbb{G}_J^{\mathcal{N}}, \mathbb{G}_J^{\mathcal{Y}})$ *are all independent. The covariance function for* $(\mathbb{G}^{\mathcal{N}}, \mathbb{G}^{\mathcal{Y}})$ *is*

$$\begin{aligned}&E\mathbb{G}^{\mathcal{N}}(s,x)\mathbb{G}^{\mathcal{N}}(t,x) = \mathcal{N}(s \wedge t, x) - \mathcal{N}(s,x)\mathcal{N}(t,x), \\ &E\mathbb{G}^{\mathcal{Y}}(s,x)\mathbb{G}^{\mathcal{Y}}(t,x) = \mathcal{Y}(s \vee t, x) - \mathcal{Y}(s,x)\mathcal{Y}(t,x), \\ &E\mathbb{G}^{\mathcal{N}}(s,x)\mathbb{G}^{\mathcal{Y}}(t,x) = \{\mathcal{N}(s,x) - \mathcal{N}(t-,x)\} I(t \le s) - \mathcal{N}(s,x)\mathcal{Y}(t,x),\end{aligned}$$

and covariance functions for $(\mathbb{G}_j^{\mathcal{N}}, \mathbb{G}_j^{\mathcal{Y}}), j = 1, \ldots, J$, *can be similarly obtained by replacing the probability* $P(\cdot, X = x)$ *in* $\mathcal{N}$ *and* $\mathcal{Y}$ *above by the conditional probability* $P(\cdot, X = x | O \in \mathcal{S}_j)$ *given stratum membership in* $\mathcal{S}_j$.

In the setting of outcome dependent stratified sampling discussed in Section 2, all cases are selected (i.e., $p_1 = 1$) and $\mathcal{S}_2, \ldots, \mathcal{S}_J$ are characterized by $\Delta = 0$ (i.e., $P(T \le x, \Delta = 1, X = x | O \in \mathcal{S}_j) = 0$ for $j = 2, \ldots, J$). Thus, the limiting processes can be simplified to

$$\begin{aligned}\mathbb{G}_\pi^{\mathcal{N}} &= \mathbb{G}^{\mathcal{N}}, \\ \mathbb{G}_\pi^{\mathcal{Y}} &= \mathbb{G}^{\mathcal{Y}} + \sum_{j=2}^{J} \left\{ \frac{\nu_j(1-p_j)}{p_j} \right\}^{1/2} \mathbb{G}_j^{\mathcal{Y}}.\end{aligned}$$

A straightforward application of the functional delta method (see, for example, Section 3.9 of van der Vaart and Wellner (1996)) yields the following asymptotic result.

Theorem 1. *Under Condition 1,* $n^{1/2}\{\hat{\Lambda}_n(\cdot|x) - \Lambda(\cdot|x)\}$ *converges weakly to the zero-mean Gaussian process* $\mathbb{L}$ *in* $D[0,\tau]$ *given by*

$$\mathbb{L}(\cdot|x) = \int_0^{\cdot} \frac{\mathbb{G}_\pi^{\mathcal{N}}(du,x)}{\mathcal{Y}(u,x)} - \int_0^{\cdot} \frac{\mathbb{G}_\pi^{\mathcal{Y}}(u,x)}{\{\mathcal{Y}(u,x)\}^2}\mathcal{N}(du,x),$$

and $n^{1/2}\{\hat{S}_n(\cdot|x) - S(\cdot|x)\}$ *converges weakly to the zero-mean Gaussian process* $\mathbb{S}$ *in* $D[0,\tau]$ *given by*

$$\mathbb{S}(\cdot|x) = S(\cdot|x)\int_0^{\cdot} \frac{\mathbb{L}(du|x)}{1 - \Delta_u\Lambda(u|x)}.$$

4. Simultaneous Confidence Band

The basic idea in constructing a confidence band is to obtain $q_{1-\alpha}$ such that

$$P\left(\sup_{t\in[0,\tau]} n^{1/2}\left|\hat{S}_n(t|x) - S(t|x)\right| \le q_{1-\alpha}\right) \to 1-\alpha, \quad n\to\infty,$$

from which the large sample $100(1-\alpha)\%$ confidence band is obtained as

$$\hat{S}_n(t|x) - n^{-1/2}q_{1-\alpha} \le S(t|x) \le \hat{S}_n(t|x) + n^{-1/2}q_{1-\alpha}, \quad \text{all } t\in[0,\tau].$$

Unlike the analysis of complete data, our complicated limiting process $\mathbb{S}$ cannot be reduced to other well-known Gaussian processes, and hence the quantiles of $\sup_{t\in[0,\tau]}|\mathbb{S}(t|x)|$ are not analytically available. An alternative method is approximate $n^{1/2}(\hat{S}_n - S)$, but no bootstrap method is available for our setting.

To estimate $q_{1-\alpha}$, we propose to approximate the limiting process $\mathbb{S}$ through simulation and bootstrap. This proposal is based on the observation that $\mathbb{S}$ can be decomposed into two independent Gaussian processes. By Lemma 1 and Theorem 1, the limiting process $\mathbb{L}$ can be written as the sum of independent Gaussian processes given by $\mathbb{L}(\cdot|x) = \mathbb{L}_1(\cdot|x) + \mathbb{L}_2(\cdot|x)$ where

$$\mathbb{L}_1(\cdot|x) = \int_0^{\cdot} \frac{\mathbb{G}^{\mathcal{N}}(du,x)}{\mathcal{Y}(u,x)} - \int_0^{\cdot} \frac{\mathbb{G}^{\mathcal{Y}}(u,x)}{\{\mathcal{Y}(u,x)\}^2}\mathcal{N}(du,x),$$

$$\mathbb{L}_2(\cdot|x) = \sum_{j=1}^{J}\left\{\frac{\nu_j(1-p_j)}{p_j}\right\}^{1/2}\left\{\int_0^{\cdot} \frac{\mathbb{G}_j^{\mathcal{N}}(du,x)}{\mathcal{Y}(u,x)} - \int_0^{\cdot} \frac{\mathbb{G}_j^{\mathcal{Y}}(u,x)}{\{\mathcal{Y}(u,x)\}^2}\mathcal{N}(du,x)\right\}.$$

We can then write the limiting process $\mathbb{S}$ as the sum of two independent Gaussian processes given by $\mathbb{S}(\cdot|x) = \mathbb{S}_1(\cdot|x) + \mathbb{S}_2(\cdot|x)$ where

$$\mathbb{S}_1(\cdot|x) = S(\cdot|x) \int_0^{\cdot} \frac{d\mathbb{L}_1(u|x)}{1 - \Delta_u \Lambda(u|x)}, \quad \mathbb{S}_2(\cdot|x) = S(\cdot|x) \int_0^{\cdot} \frac{d\mathbb{L}_2(u|x)}{1 - \Delta_u \Lambda(u|x)}.$$

The key idea is to directly simulate $\mathbb{S}_1$ and to bootstrap certain quantity described below to approximate $\mathbb{S}_2$. Independence of $\mathbb{S}_1$ and $\mathbb{S}_2$ guarantees the validity of separate approximation.

This decomposition reflects two sources of randomness. The process $\mathbb{S}_1$ represents randomness due to sampling from the infinite population. In fact, this process is the exactly the same as the limiting process of the Kaplan–Meier estimator when complete data woud be obtained Breslow and Crowley (1974). The process $\mathbb{S}_2$ thus represents randomness due to additional sampling from strata. This interpretation leads to the decomposition in the estimator with the hypothetically computed standard Kaplan–Meier estimator S_n from complete data of size n. Since $n^{1/2}(\hat{S}_n - S) = n^{1/2}(S_n - S) + n^{1/2}(\hat{S}_n - S_n)$, the process $n^{1/2}(S_n - S)$ and $n^{1/2}(\hat{S}_n - S_n)$ weakly converge to $\mathbb{S}_1$ and $\mathbb{S}_2$ respectively.

The well-studied process $\mathbb{S}_1$ is the zero-mean Gaussian process with a covariance function $\rho(s, t|x)$ at times s and t given by

$$\rho(s, t|x) = S(s|x)S(t|x) \int_0^{s \wedge t} \frac{\mathcal{N}(du, x)}{\mathcal{Y}(u, x)\mathcal{Y}(u-, x)}.$$

This covariance function can be estimated by the consistent plug-in estimator

$$\hat{\rho}_n(s, t|x) = \hat{S}_n(s|x)\hat{S}_n(t|x) \int_0^{s \wedge t} \frac{\overline{N}_n^{\pi}(du, x)}{\overline{Y}_n^{\pi}(u, x)\overline{Y}_n^{\pi}(u-, x)}.$$

If we remove the inverse probability weights, this estimator reduces to the well-known Greenwood's formula. We generate a zero-mean Gaussian process $\hat{\mathbb{S}}_{n,1}$ with the estimated covariance function $\hat{\rho}_n(s, t|x)$.

Unlike $\mathbb{S}_1$, the process $\mathbb{S}_2$ is not a Gaussian martingale in general, and its complicated covariance function cannot be estimated by Greenwood's formula. Instead, we approximate $\mathbb{S}_2$ by bootstrapping $n^{1/2}(\hat{S}_n - S_n)$ which weakly converges to $\mathbb{S}_2$.

The bootstrap procedure we use is the extension of Gross' bootstrap Gross (1980) by Bickel and Freedman (1984). There are other bootstrap methods (see e.g. Booth et al. (1994); Chao and Lo (1985); Mashreghi et al. (2016); Sitter (1992)) for the finite population but the weak convergence of bootstrap empirical process is only proved for Gross's bootstrap Saegusa (2015) which requires the theoretical properties of confidence bands. This bootstrap method concerns a finite population sampling and is suitable for reproducing randomness only from stratified sampling in our context. The basic idea of this bootstrap method is the following. If we sample 50 subjects from 100 people without replacement, we double 50 selected subjects to create a bootstrap population of size 100 from which 50 items are selected without replacement into a bootstrap sample. The formal description is given below.

Let W_i be the count of how many times the ith subject is selected in a bootstrap sample. For a subject with $\xi_i = 0$, $W_i = 0$. For m_j selected subjects in stratum $\mathcal{S}_j$, suppose n_j/m_j is an integer k_j. Then create a bootstrap population of size $n_j = k_j m_j$ consisting of k_j copies of each selected subject in stratum $\mathcal{S}_j$. Select m_j subjects from the bootstrap population without replacement into a bootstrap sample. This determines W_i for the m_j selected subjects. Suppose $n_j = m_j k_j + r_j$ where m_j is a divisor and r_j is a remainder. With probability $s_j = (1 - r_j/m_j)\{1 - r_j/(n_j - 1)\}$, create a bootstrap population of size $k_j m_j$ as before and sample m_j subjects without replacement. Otherwise, create a bootstrap population of size $(k_j + 1)m_j$ consisting of $k_j + 1$ copies of each selected subjects and select m_j subjects without replacement. For the stratum $\mathcal{S}_1$ in the outcome dependent stratified design, all subjects are selected and hence $W_i = 1$.

The bootstrap estimator of Λ and S are computed as

$$\hat{\Lambda}_n^b(\cdot|x) = \int_0^{\cdot} \frac{1}{\overline{Y}_n^{b,\pi}(u,x)} \overline{N}_n^{b,\pi}(du,x),$$

$$\hat{S}_n^b(\cdot|x) = \prod_{0<u\leq\cdot} \left\{1 - \Delta_u \hat{\Lambda}_n^b(u|x)\right\} = \prod_{0<u\leq\cdot} \left\{1 - \frac{\Delta_u \overline{N}_n^{b,\pi}(u,x)}{\overline{Y}_n^{b,\pi}(u,x)}\right\},$$

where

$$\overline{N}_n^{b,\pi}(t,x) = \frac{1}{n}\sum_{i=1}^{n} W_i \frac{\xi_i}{\pi_i} N_i(t|x), \quad \overline{Y}_n^{b,\pi}(t,x) = \frac{1}{n}\sum_{i=1}^{n} W_i \frac{\xi_i}{\pi_i} Y_i(t,x).$$

For bootstrapping $n^{1/2}(\hat{S}_n - S_n)$, we compute $\hat{\mathbb{S}}_{n,2} = n^{1/2}(\hat{S}_n^b - \hat{S}_n)$.

The approximation of $\mathbb{S}$ by $\hat{\mathbb{S}}_{n,1}+\hat{\mathbb{S}}_{n,2}$ is asymptotically valid. However, $\sup_{t\in[0,\tau]} |\mathbb{S}(t|x)|$ may have a jump at the lower end of the support of T though $\mathbb{S}$ is sample continuous Tsirelson (1975). To exclude the possibility that a jump occurs at $q_{1-\alpha}$, we assume the following condition. A similar condition is imposed by Bickel and Krieger (1989) for non-censored data.

Condition 2. *The distribution of* $\sup_{t\in[0,\tau]} |\mathbb{S}(t|x)|$ *is continuous.*

Theorem 2. *Let* $q > 0$. *Under Conditions 1 and 2,*

$$P\left(\sup_{t\in[0,\tau]} \left|\hat{\mathbb{S}}_{n,1}(t|x) + \hat{\mathbb{S}}_{n,2}(t|x)\right| \le q\right) \to P\left(\sup_{t\in[0,\tau]} |\mathbb{S}(t|x)| \le q\right), \quad \text{as } n \to \infty.$$

We propose the following simulation-based procedure to construct confidence band.

Step 1. Generate $\hat{\mathbb{S}}_{n,1}$ and $\hat{\mathbb{S}}_{n,2}$ described above to obtain $c_{1,x} = \sup_{t\in[0,\tau]} |\hat{\mathbb{S}}_{n,1}(t|x)+ \hat{\mathbb{S}}_{n,2}(t|x)|$.

Step 2. Repeat Step 1 B times and compute $100(1-\alpha)\%$tile $\hat{q}_{1-\alpha,x}$ of $c_{1,x}, \ldots, c_{B,x}$.

Step 3. Compute our $100(1-\alpha)\%$ confidence band given by

$$\hat{S}_n(t|x) \pm n^{-1/2}\hat{q}_{1-\alpha,x}, \quad t \in [0,\tau]. \tag{1}$$

The proposed confidence band achieves correct coverage probability asymptotically.

Theorem 3. *Let* $\alpha \in (0,1)$. *Under Conditions 1 and 2, as* $n \to \infty$ *and* $B \to \infty$,

$$P\left(\hat{S}_n(t|x) - n^{-1/2}\hat{q}_{1-\alpha,x} \le S(t|x) \le \hat{S}_n(t|x) + n^{-1/2}\hat{q}_{1-\alpha,x},\ t \in [0,\tau]\right) \to 1-\alpha.$$

The proposed confidence band has the same width at each time $t \in [0, \tau]$ but estimation of S is less reliable on the right tail. More desirable confidence bands have smaller width for earlier time and larger width for later time. To achieve this property, we can modify the proposed procedure by computing $c^*_{k,x} = \sup_{t\in[0,\tau]} |\{\hat{\mathbb{S}}_{n,1}(t|x) + \hat{\mathbb{S}}_{n,2}(t|x)\}/f(t)|$ for a fixed positive and increasing function f on $[0, \tau]$ and obtain the corresponding $100(1-\alpha)\%$tile $\hat{q}^*_{1-\alpha,x}$. The resultant $100(1-\alpha)\%$ confidence band of variable width is

$$\hat{S}_n(t|x) \pm n^{-1/2} f(t) \hat{q}^*_{1-\alpha,x}, \quad t \in [0, \tau]. \tag{2}$$

In this paper, we use exponential functions for f in simulation and data analysis. Choice of f possibly in an adaptive way deserves further investigation.

Note that n in the confidence band is the size of the entire cohort. In complete data analysis, data are split by the exposure status $X = x$ and n appearing in the confidence bands is the size of split data. In our design, the inverse probability weighting corrects for biased sampling and the size of the entire cohort appears in the computation of the limiting distribution.

5. Additional Results

5.1. **Variance Estimation.** In the previous section, we develop the method of constructing confidence bands for survival functions. The same idea naturally applies to a simpler question of variance estimation. Consider estimating a variance of $\hat{S}_n(t|x)$ at a fixed $t \in (0, \tau]$. It follows by Theorem 1 and discussion in Section 4, $n^{1/2}(\hat{S}(t|x) - S(t|x))$ converges in distribution to the sum of independent variables $\mathbb{S}_1(t|x)$ and $\mathbb{S}_2(t|x)$. The variance of $\mathbb{S}_1(t|x)$ is $\rho(t, t|x)$. This is estimated by the weighted Greenwood's formula $\hat{\rho}_n(t, t|x)$. For variance of $\mathbb{S}_2(t|x)$, we generate bootstrap estimator $n^{1/2}(\hat{S}^b_n(t|x) - \hat{S}_n(t|x))$ of $n^{1/2}(\hat{S}_n(t|x) - S_n(t|x))$ and compute its sample variance. Adding these quantities and dividing them by n, we obtain our variance estimator of $\hat{S}_n(t|x)$.

For two-phase cohort studies, Rebora and Valsecchi (2016) studied the variance estimation of the weighted Kaplan-Meier estimator assuming the continuity of the survival time. They computed the limiting process as $\mathbb{S}_1$ for general complex sampling, not $\mathbb{S}_1 + \mathbb{S}_2$, from which they considered two terms for variance estimator. The first term is essentially the same as our weighted Greenwood's formula derived from $\mathbb{S}_1$. They instead used $\mathbb{S}_1$ to obtain the second term as conditional variance of the sampling indicators in $\hat{\Lambda}_n$ given $(\tilde{T}_i, \Delta_i), i = 1 \ldots, n$. Given the two-phase framework in survey sampling Rubin-Bleuer and Schiopu Kratina (2005), this estimator that captures two sources of randomness seems reasonable. They also proposed an alternative to the second term by the linearization Demnati and Rao (2004, 2010). Both estimators for the second term is based on standard techniques in survey sampling for the analysis of sums and totals. It is of a theoretical interest to study asymptotic properties of their estimator because it is not straightforward to see how those techniques behave when applying to random functions.

5.2. **Exposure Stratified Case-Cohort Study.** Our methodology is easily extended to the exposure stratified case-cohort study Borgan et al. (2000) with the modification of strata. Data from the exposure stratified case-cohort study are a stratified sample from the entire cohort and all other cases not selected from stratified sampling. Borgan et al. (2000) considered three estimators for the Cox model but one of those estimator (Estimator I) only uses a stratified sample. In this case, one can view the exposure stratified case-cohort design as stratified sampling without dependence on outcomes. Consider stratification $\mathcal{S}_1 = \{O : U \in \mathcal{U}_2\}, \ldots, \mathcal{S}_{J-1} = \{O : U \in \mathcal{U}_J \text{ with } \pi_i = m_j/n_j \text{ if } O_i \in \mathcal{S}_j, j = 1, \ldots, J$. Then we can compute the inverse probability weighted Kaplan-Meier estimator in the same way as above. The limiting process takes the same form with a modification of $\{\mathbb{G}_\pi^{\mathcal{N}}(\cdot, x), \mathbb{G}_\pi^{\mathcal{Y}}(\cdot, x)\}$ that only reflects different definitions of strata in Lemma 1 and Theorem 1. The methodology for confidence bands and variance estimation can be carried out in the same way as above. A limitation of this approach which also holds for Estimator I of Borgan et al. (2000) is that we

do not fully use all available data of cases. Especially when the event is rare, a stratified sample may not contain enough cases so that resultant confidence bands would be unstable. An extension that accommodates all cases is desired in the future research.

6. Simulation Study

We performed a simulation study to evaluate the finite-sample performance of the proposed methodology. The failure time T is generated from the Cox proportional hazards model with two independent binary covariates (V, X) with same prevalence 30%. The auxiliary binary variable U is related to X with sensitivity and specificity 0.9. The baseline hazard functions corresponds to the Weibull distribution with parameter $\alpha = 0.2$ and $\beta \in \{0.5, 1, 3\}$. The regression coefficients are both $\log 2$. The censoring variable C follows the uniform distribution on $[0, c]$ where c is selected to achieve heavy censoring proportion of 80%. Three strata are considered with $\mathcal{S}_1 = \{O : \Delta = 1\}$, $\mathcal{S}_2 = \{O : \Delta = 0, U = 0\}$, and $\mathcal{S}_3 = \{O : \Delta = 0, U = 1\}$. All cases are selected from $\mathcal{S}_1$ and 30 percent of subjects are selected without replacement from $\mathcal{S}_2$ and $\mathcal{S}_3$ respectively. Confidence bands were constructed over the interval $[0, c-.2]$ 500 times where each was based on 3000 iterations to generate $\hat{\mathbb{S}}_{n,1}$ and $\hat{\mathbb{S}}_{n,2}$. In addition to the equal-width band (1), we created an variable-width band (2) with $f(t) = \exp(t)$.

Table 1 shows simulated coverage probabilities of the proposed confidence bands at the nominal level 95%. The variable-width bands achieve accurate coverages even for smaller sample sizes in general, showing superior performance over the equal-width bands. The equal-width bands show reasonable performance in most settings where coverage is improved with a larger sample size. The coverage probabilities for this method are generally closer to the nominal level at $X = 0$ than $X = 1$ because there are fewer cases for the group with $X = 1$ in the final sample. On average, censoring proportions are about 60 percent and 45 percent for groups with $X = 0$ and $X = 1$ in the final sample, which yields about 88 cases and 35 cases

respectively when $n = 500$ and about 175 cases and 69 cases respectively when $n = 1000$. When $\beta = 3$, a larger sample size is required for more accurate results but relatively poor performance of equal precision bands compared to settings with $\beta = 0.5$ and $\beta = 1$ indicates that this setting is most difficult among three.

Figure 1 presents four different confidence bands based on the data sets generated with the Weibull distribution with $\beta = 0.5$ and $\beta = 3$. In addition to our method, we consider Nair's equal-precision bands Nair (1984) and Hall-Wellner bands Hall and Wellner (1980) for the i.i.d. setting. For these methods, we additionally generated the i.i.d. sample of size comparable to our final sample sizes. Note that we cannot apply these methods to the same stratified data because unweighted Kaplan-Meier curves are biased and do not lead to the fair and meaningful reference. All methods contain true survival curves. The width of the equal-width bands are narrower than all other methods on the right tail. The failure to account for increasing uncertainty in estimation of the survival curves on the right leaves room for improvement to the variable-width band. The variable-width bands are narrow on earlier time and increase their width as time progresses. Both methods have reasonable shapes compared to other two methods with complete data. For graphical representation, upper limits of the bands, when exceeding 1, are reduced to 1. For the variable-width bands, upper or lower limits may be increasing on some time intervals. Since survival curves are decreasing, we force limits of the bands to be decreasing by choosing smallest values on the left.

7. Application

We apply our method to analyze the the national Wilms tumor study Green et al. (1998) where 3915 patients with Wilms tumor were followed until the disease progression during 1980-1994. The outcome of interest is time to relapse of cancer in years. Nine strata were formed based on case status defined by tumor relapse, age at diagnosis (less than a year old or older), stage of cancer (I and II, or III and IV), and histology measured at the hospital (favorable or unfavorable). All patients were

selected except three strata with favorable histology. From the stratum of children less than a year old with early stage of cancer, 120 patients were selected from 452 patients. From the stratum of children older than a year old with early stage of cancer, 160 patients were selected from 1620 patients. From the stratum of children older than a year old with advanced stage of cancer, 120 patients were selected from 914 patients. In total, 1329 patients were selected consisting of 669 cases and 660 controls, and their histology was measured again at the central reference laboratory, which is considered as the gold standard. Our goal is to construct confidence bands of two relapse-free survival curves with favorable and unfavorable histology measured at the second time. To determine confidence bands, we repeated our simulation and bootstrap 1000 times. For the variable-width confidence bands, we used $f(t) = \exp(t/22)$ based on the maximum censored time is 22 years.

Figure 2 shows inverse probability weighted Kaplan-Meier estimators and corresponding confidence bands. Both estimates show more relapse of cancer short time after the first onset, and then reach a plateau. The drop in relapse-free survival probability is clearly larger for the group with unfavorable histology, which indicates histology is an important predictor of cancer relapse. This finding was confirmed by Breslow et al. (2009) in their proportional hazards regression analysis. Our proposed confidence bands provide uncertainty of estimates in this graphical comparison. Both bands are comparable due to a large sample size.

8. Discussion

In this paper, we propose the method of constructing the confidence bands for the survival functions over the interval based on the inverse probability weighted Kaplan-Meier curves. The proposed method is the first valid confidence bands in survival analysis when data are collected from outcome-dependent sampling design with sampling without replacement. Despite the simplicity of weighted Kaplan-Meier curves, great challenges for constructing valid confidence bands are (1) the difficulty in deriving the limiting distribution in the presence of dependence, and (2)

the lack of analytical formulae for quantiles of the supremum of general Gaussian processes. For the first issue, we apply the uniform central limit theorem and the functional delta method from the special empirical process theory for two-phase stratified samples Breslow and Wellner (2007); Saegusa and Wellner (2013) (see the supplementary materials for the distinction in the meaning of two-phase sampling in biostatistics and survey sampling). For the second issue, we separate the limiting process according to different sampling phases. We then approximate the limiting process corresponding to sampling from the infinite population by simulating the estimated limiting process itself, while we adopted the specialized bootstrap procedure in the finite population sampling for the rest of the limiting process. The proposed procedure showed good coverage in simulation studies.

A limitation of our approach is computational costs arising from both simulating Gaussian processes and bootstrapping stratified samples. One potential way to reduce computational costs is to omit one of the two extensive computations. For example, if one constructs a new bootstrap method that reproduce randomness from sampling from the infinite population and subsequent sampling from strata at the same time, simulation on Gaussian processes can be omitted. Or if one analytically simplifies the linear combination of Gaussian processes to the single Gaussian process, one have only to generate the resultant Gaussian processes. In this paper, we provide a simple method to take care of distinct randomness separately, but these directions are worthwhile to explore in the future research.

There are several other directions for further work. In the i.i.d. setting, the confidence bands for the conditional hazard and survival functions have been extensively studied when covariate are available (see Dobler et al. (2019) and reference therein). Because regression modeling sheds further insight on survival data, constructing similar confidence bands in our setting facilitates the use of cost-effective outcome-dependent stratified designs in public health research. The limiting distribution of the conditional hazard function was shown to be the linear combination of Gaussian processes Breslow et al. (2015) as in our case. Our approach demonstrated in this

paper with certain modifications is expected to address this challenging question. Another direction of research is to consider different sampling designs. Because our method counts on empirical process theory for outcome-dependent stratified sampling, different designs require different empirical process theory. Furthermore, our bootstrap method for sampling without replacement must be replaced by new bootstrap methods specializing in different sampling designs.

Acknowledgements. The first author is supported by the National Science Foundation, USA (DMS 2014971). We would like to thank the editor, the associate editor, and the referees for their constructive comments and suggestions.

Supporting Information. Additional information for this article is available online, containing proofs for Lemma 1, Theorem 1, Theorem 2, and Theorem 3, and R code.

References

Akritas, M. G. (1986). Bootstrapping the Kaplan-Meier estimator. *J. Amer. Statist. Assoc.*, 81(396):1032–1038.

Amato, D. A. (1988). A generalized Kaplan-Meier estimator for heterogeneous populations. *Comm. Statist. Theory Methods*, 17(1):263–286.

Bickel, P. J. and Freedman, D. A. (1984). Asymptotic normality and the bootstrap in stratified sampling. *Ann. Statist.*, 12(2):470–482.

Bickel, P. J. and Krieger, A. M. (1989). Confidence bands for a distribution function using the bootstrap. *J. Amer. Statist. Assoc.*, 84(405):95–100.

Booth, J. G., Butler, R. W., and Hall, P. (1994). Bootstrap methods for finite populations. *J. Amer. Statist. Assoc.*, 89(428):1282–1289.

Borgan, Ø., Langholz, B., Samuelsen, S. O., Goldstein, L., and Pogoda, J. (2000). Exposure stratified case-cohort designs. *Lifetime Data Anal.*, 6(1):39–58.

Borgan, O. r. and Liestø l, K. (1990). A note on confidence intervals and bands for the survival function based on transformations. *Scand. J. Statist.*, 17(1):35–41.

Breslow, N. and Crowley, J. (1974). A large sample study of the life table and product limit estimates under random censorship. *Ann. Statist.*, 2:437–453.

Breslow, N. E., Hu, J., and Wellner, J. A. (2015). Z-estimation and stratified samples: application to survival models. *Lifetime Data Anal.*, 21(4):493–516.

Breslow, N. E., Lumley, T., Ballantyne, C., Chambless, L., and Kulich, M. (2009). Using the whole cohort in the analysis of case-cohort data. *American J. Epidemiol.*, 169:1398–1405.

Breslow, N. E. and Wellner, J. A. (2007). Weighted likelihood for semiparametric models and two-phase stratified samples, with application to Cox regression. *Scand. J. Statist.*, 34(1):86–102.

Cai, T. and Zheng, Y. (2013). Resampling procedures for making inference under nested case-control studies. *J. Amer. Statist. Assoc.*, 108(504):1532–1544.

Chao, M. T. and Lo, S.-H. (1985). A bootstrap method for finite population. *Sankhyā Ser. A*, 47(3):399–405.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2014). Gaussian approximation of suprema of empirical processes. *Ann. Statist.*, 42(4):1564–1597.

Demnati, A. and Rao, J. (2004). Linearization variance estimators for survey data. *Survey Methodology*, 30(1):pp. 17–26.

Demnati, A. and Rao, J. (2010). Linearization variance estimators for model parameters from complex survey data. *Survey Methodology*, 36(2):pp. 193–201.

Dobler, D., Pauly, M., and Scheike, T. H. (2019). Confidence bands for multiplicative hazards models: flexible resampling approaches. *Biometrics*, 75(3):906–916.

Galimberti, S., Sasieni, P., and Valsecchi, M. G. (2002). A weighted kaplan–meier estimator for matched data with application to the comparison of chemotherapy and bone-marrow transplant in leukaemia. *Statistics in Medicine*, 21(24):3847–3864.

Gillespie, M. J. and Fisher, L. (1979). Confidence bands for the Kaplan-Meier survival curve estimate. *Ann. Statist.*, 7(4):920–924.

Green, D. M., Breslow, N. E., Beckwith, J. B., Finklestein, J. Z., Grundy, P. E., Thomas, P. R., Kim, T., Shochat, S. J., Haase, G. M., Ritchey, M. L., Kelalis, P. P., and D'Angio, G. J. (1998). Comparison between single-dose and divided-dose administration of dactinomycin and doxorubicin for patients with Wilms' tumor: a report from the National Wilms' Tumor Study Group. *J. Clin. Oncol.*, 16(1):237–245.

Greenwood, M. (1926). *A Report on the Natural Duration of Cancer*. Reports on public health and medical subjects. H.M. Stationery Office.

Gross, S. (1980). Median estimation in sample surveys. In *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pages 181–184.

Hall, W. J. and Wellner, J. A. (1980). Confidence bands for a survival curve from censored data. *Biometrika*, 67(1):133–143.

Harper, A. J. (2013). Bounds on the suprema of Gaussian processes, and omega results for the sum of a random multiplicative function. *Ann. Appl. Probab.*, 23(2):584–616.

Hollander, M., McKeague, I. W., and Yang, J. (1997). Likelihood ratio-based confidence bands for survival functions. *J. Amer. Statist. Assoc.*, 92(437):215–226.

Huang, Y. (2014). Bootstrap for the case-cohort design. *Biometrika*, 101(2):465–476.

Kang, S. and Cai, J. (2009). Marginal hazards model for case-cohort studies with multiple disease outcomes. *Biometrika*, 96(4):887–901.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *J. Amer. Statist. Assoc.*, 53:457–481.

Kong, L., Cai, J., and Sen, P. K. (2006). Asymptotic results for fitting semiparametric transformation models to failure time data from case-cohort studies. *Statist. Sinica*, 16(1):135–151.

Kulich, M. and Lin, D. Y. (2000). Additive hazards regression for case-cohort studies. *Biometrika*, 87(1):73–87.

Li, Z. and Nan, B. (2011). Relative risk regression for current status data in case-cohort studies. *Canad. J. Statist.*, 39(4):557–577.

Lo, S.-H. and Singh, K. (1986). The product-limit estimator and the bootstrap: some asymptotic representations. *Probab. Theory Relat. Fields*, 71(3):455–465.

Lu, W. and Tsiatis, A. A. (2006). Semiparametric transformation models for the case-cohort study. *Biometrika*, 93(1):207–214.

Ma, S. and Kosorok, M. R. (2005). Robust semiparametric M-estimation and the weighted bootstrap. *J. Multivariate Anal.*, 96(1):190–217.

Mashreghi, Z., Haziza, D., and Léger, C. (2016). A survey of bootstrap methods in finite population sampling. *Stat. Surv.*, 10:1–52.

Moger, T. A., Pawitan, Y., and Borgan, Ø. (2008). Case–cohort methods for survival data on families from routine registers. *Statistics in Medicine*, 27(7):1062–1074.

Nair, V. N. (1984). Confidence bands for survival functions with censored data: A comparative study. *Technometrics*, 26(3):265–275.

Nan, B., Kalbfleisch, J. D., and Yu, M. (2009). Asymptotic theory for the semiparametric accelerated failure time model with missing data. *Ann. Statist.*, 37(5A):2351–2376.

Nan, B., Yu, M., and Kalbfleisch, J. D. (2006). Censored linear regression for case-cohort studies. *Biometrika*, 93(4):747–762.

Pollard, D. (1984). *Convergence of Stochastic Processes*. Springer-Verlag.

Prentice, R. L. (1986). A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika*, 73:pp. 1–11.

Prentice, R. L. and Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika*, 66(3):403–411.

Rebora, P. and Valsecchi, M. G. (2016). Survival estimation in two-phase cohort studies with application to biomarkers evaluation. *Stat. Methods Med. Res.*, 25(6):2895–2908.

Rubin-Bleuer, S. and Schiopu Kratina, I. (2005). On the two-phase framework for joint model and design-based inference. *Ann. Statist.*, 33(6):2789–2810.

Saegusa, T. (2015). Variance estimation under two-phase sampling. *Scandinavian Journal of Statistics*, 42(4):1078–1091.

Saegusa, T. and Wellner, J. A. (2013). Weighted likelihood estimation under two-phase sampling. *Ann. Statist.*, 41(1):269–295.

Self, S. G. and Prentice, R. L. (1988). Asymptotic distribution theory and efficiency results for case-cohort studies. *Ann. Statist.*, 16(1):64–81.

Shorack, G. R. and Wellner, J. A. (1986). *Empirical processes with applications to statistics*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons, Inc., New York.

Sitter, R. R. (1992). Comparing three bootstrap methods for survey data. *Canad. J. Statist.*, 20(2):135–154.

Sø rensen, P. and Andersen, P. K. (2000). Competing risks analysis of the case-cohort design. *Biometrika*, 87(1):49–59.

Sun, J., Sun, L., and Flournoy, N. (2004). Additive hazards model for competing risks analysis of the case-cohort design. *Comm. Statist. Theory Methods*, 33(2):351–366.

Tsirelson, V. S. (1975). The density of the distribution of the maximum of a Gaussian process. *Theory of Probability and its Applications*, 20:847–865.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer Series in Statistics. Springer-Verlag, New York.

White, J. E. (1986). A two stage design for the study of the relationship between a rare exposure and and a rare disease. *Am. J. Epidemiol.*, 115(1):119–128.

Williams, R. (1995). Product-limit survival functions with correlated survival times. *Lifetime Data Analysis*, 1(2):171–186.

Winnett, A. and Sasieni, P. (2002). Adjusted Nelson-Aalen estimates with retrospective matching. *J. Amer. Statist. Assoc.*, 97(457):245–256.

Zeng, D. and Lin, D. Y. (2014). Efficient estimation of semiparametric transformation models for two-phase cohort studies. *J. Amer. Statist. Assoc.*, 109(505):371–383.

Zhou, Q., Cai, J., and Zhou, H. (2018). Outcome-dependent sampling with interval-censored failure time data. *Biometrics*, 74(1):58–67.

Zhou, Q., Zhou, H., and Cai, J. (2017). Case-cohort studies with interval-censored failure time data. *Biometrika*, 104(1):17–29.

Takumi Saegusa

Department of Mathematics, University of Maryland

College Park, MD, United States of America

E-mail address: `tsaegusa@umd.edu`

Peter Nandori

Department of Mathematics, Yeshiva University

New York, NY, United States of America

E-mail address: `peter.nandori@yu.edu`

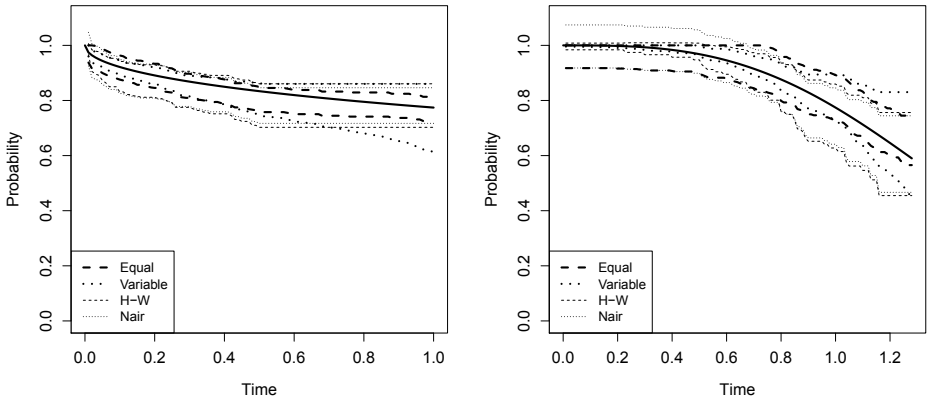

FIGURE 1. Confidence bands for survival curves at $X = 0$ with $n = 1000$ and $\beta = 0.5$ (left panel) and $\beta = 3$ (right panel). Solid lines for true survival curves, wide dashed lines for the equal-width bands, wide dotted lines for the variable-width bands, narrow dashed lines for the Hall-Wellner band, and narrow dotted lines for Nair's equal precision bands

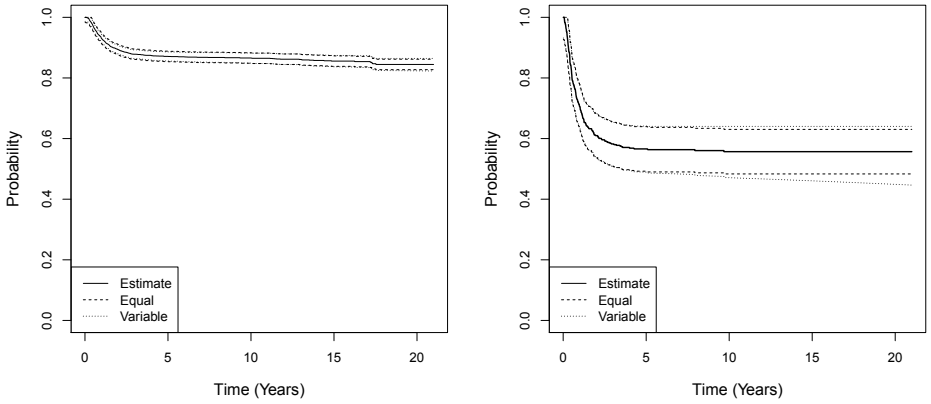

FIGURE 2. Weighted Kaplan-Meier estimators and confidence bands for survival curves with favorable histology (left panel) and unfavorable histology (right panel) based on data from the national Wilms tumor study. Dashed lines for equal-width bands, dotted lines for variable-width bands

TABLE 1. Empirical coverages of 95% confidence bands in different data generating mechanism.

X	β	n	m	equal	variable
$X=0$	0.5	500	146	91.8%	93.6%
		1000	292	94.8%	95.0%
	1	500	146	95.5%	96.6%
		1000	293	95.8%	94.4%
	3	500	146	93.9%	95.1%
		1000	291	93.8%	95.4%
$X=1$	0.5	500	77	93.6%	94.4%
		1000	154	94.4%	94.8%
	1	500	76	93.8%	96.6%
		1000	153	94.7%	94.6%
	3	500	73	91.0%	97.2%
		1000	147	91.9%	95.4%