Published in final edited form as:

Biometrics. 2012 June ; 68(2): 521–531. doi:10.1111/j.1541-0420.2011.01708.x.

Estimating Incident Population Distribution from Prevalent Data

Kwun Chuen Gary Chan^{1,*} and Mei-Cheng Wang^{2,**}

¹Department of Biostatistics and Department of Health Services, University of Washington, Seattle, Washington 98195, U.S.A

²Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, Maryland 21205, U.S.A

Summary

A prevalent sample consists of individuals who have experienced disease incidence but not failure event at the sampling time. We discuss methods for estimating the distribution function of a random vector defined at baseline for an incident disease population when data are collected by prevalent sampling. Prevalent sampling design is often more focused and economical than incident study design for studying the survival distribution of a diseased population, but prevalent samples are biased by design. Subjects with longer survival time are more likely to be included in a prevalent cohort, and other baseline variables of interests that are correlated with survival time are also subject to sampling bias induced by the prevalent sampling scheme. Without recognition of the bias, applying empirical distribution function to estimate the population distribution of baseline variables can lead to serious bias. In this article, nonparametric and semiparametric methods are developed for distribution estimation of baseline variables using prevalent data.

Keywords

Accelerated failure time model; Cross-sectional sampling; Left truncation; Proportional hazards model

1. Introduction

An incident population in epidemiology is composed of individuals who experienced a disease incidence within a specified calendar time interval. In contrast to an incident population, a prevalent population is defined as a group of individuals who have experienced disease incidence but not failure event before a particular calendar time. A prevalent cohort is defined as a sample selected from a prevalent population. In a study of Alzheimer's disease where the failure time of interest is the time from disease onset to death, a prevalent cohort can be defined as a group of living patients who have been diagnosed with Alzheimer's disease before the time of recruitment. In an AIDS study where the incubation time between HIV infection and diagnosis of AIDS is the failure time of interest, a prevalent cohort can be defined as a group of patients who have been HIV infected but have not been diagnosed with AIDS at the time of recruitment.

In studies of the natural history of a disease, both the incident and prevalent sampling schemes can be adopted for data collection, though the probability structure of an incident population is the main scientific interest. In practice, incident sampling design could be

inefficient for studying disease history because it could take a long follow-up time to observe enough failure events for meaningful analysis. In contrast, a prevalent sampling design that draws samples from a disease prevalent population is more focused and economical (Brookmeyer and Gail, 1987; Wang, 1991; Wang, Brookmeyer, and Jewell, 1993). Nevertheless, standard methods in survival analysis are usually not directly applicable to prevalent cohort data because the sample is biased toward longer survival. When the variable of interest is time from disease diagnosis to a failure event, thereafter called failure time, statistical methods have been developed for nonparametric estimation (Vardi, 1989; Wang, 1991; Asgharian, M' Lan, and Wolfson, 2002, among others) and regression analysis (Lai and Ying, 1991; Wang et al., 1993, among others).

Very often, in addition to failure time, investigators are also interested in baseline variables associated with the failure time. For example, genetic markers associated with disease incidence and survival, age of onset among a group of patients with Alzheimer's disease, and other demographic variables that provide information about the disease burden in population are often studied and reported. Baseline variables of interests that are correlated with failure time are subject to bias induced by the prevalent sampling design. Failure to recognize this bias can lead to serious error in reporting the prevalence of risk factors among incident cases.

Although survival distribution is usually the main interest when employing the prevalent sampling design, covariate information is almost always reported in biomedical papers either as descriptive statistics for the study population, or combined with survival models for additional analysis, like in the calculation of population attributable risk (Chen, Hu, and Wang, 2006). For real data applications, under budget and other practical constraints, it may not be possible to augment the prevalent cohort with another incident cohort for the purpose of estimating covariate distributions. The other possible case is that, using an incident cohort design, the investigator decides to collect certain new covariates some time after the study has started. To ensure the availability of new covariates from study subjects, a prevalent cohort would then be artificially formed for statistical analysis.

In the statistical literature, most of the research on prevalent data focused on bias correction for failure time, and there was considerably less attention on bias correction for baseline variables. With focus on estimation of regression coefficients for survival models, the problem of covariate bias for prevalent survival data was considered in a recent paper by Bergeron, Asgharian, and Wolfson (2008) for parametric regression models under stationarity assumption. In an unpublished manuscript, Gürlér and Gijbels (1996) developed methods for nonparametric bivariate estimation of failure time and an associated variable for left-truncated and right-censored data that can be applied to prevalent data. In contrast with existing approaches, this article focuses on nonparametric and semiparametric estimation of marginal distribution of baseline variables of interest from prevalent survival data. In the semiparametric settings, our estimators are obtained based on semiparametric regression models of failure time. When survival distribution does not depend on calendar time of onset, the proposed estimator for the onset time distribution reduces to that proposed in Wang (1991). However, Wang (1991) did not consider estimation of incident covariate distribution, which is our main focus in this article.

The article is organized as follows. In Section 2, a probability model for prevalent sampling is introduced and baseline variables in a prevalent cohort are shown to be a biased sample from the incident population distribution. In Section 3, nonparametric bias corrected estimation of the distribution function of categorical variables will be discussed. In Section 4, bias corrected estimation of the distribution function of discrete or continuous variables will be studied under proportional hazards model and accelerated failure time model.

Simulations for the proposed nonparametric and semiparametric estimators and data analysis using the surveillance, epidemiology and end results (SEER)–Medicare linked data are included in Section 5. We will conclude this article with several remarks in Section 6.

2. Prevalent Population Distribution

In this section, we introduce a number of random variables defined in incident and prevalent populations, and describe the probability distributions of these variables. We first consider random variables defined in an incident population. Let X^0 be a failure time and Z^0 be a baseline random variable or vector of variables of interest; for instance, Z^0 could be the onset age of a disease, risk factors, or demographical variables. Also, let T^0 be the time from disease incidence to time of recruitment. Censoring is defined in the prevalent population for subjects who are eligible to be sampled at the recruitment time, we denote C' be the time from recruitment to the potential censoring time. For convenience of discussion, we extend definition of censoring time from prevalent population to incident population and define the time from disease diagnosis to time of censoring as $C^0 = T^0 \oplus C'$, the convolution of T^0 and C' . We assume in this article that the baseline variables are collected at the sampling time and are observed regardless of whether the failure events are eventually censored.

Under the prevalent sampling, an individual would be qualified to be included in the sampling population at the recruitment time only if $X^0 - T^0$ is satisfied; that is, observation of failure events are left truncated. A graphical illustration of prevalent sampling is provided in Figure 1.

We define random variables in the prevalent population by (T, Z, X, C) , and note that the probability distribution of (T, Z, X, C) is the same as the probability distribution of (T^0, Z^0, X^0, C^0) conditional on $X^0 - T^0$. The observed data include independent and identically distributed (i.i.d.) copies of (T, Z, Y, δ) , where $Y = \min\{X, C\}$ and $\delta = I(X \leq C)$. We assume a independent censoring and truncation assumption throughout the article, where X^0 is independent of (T^0, C^0) conditioning on Z^0 . This is a common assumption in regression analysis for left-truncated and right-censored data.

For convenience of exposition, we assume the existence of probability density functions for random variables or vectors in discussion. For any random variable W , we denote f_W , F_W , and S_W as density function, distribution function, and survival function for W , respectively. We also assume that

$$\inf\{t: F_{T^0}(t) > 0\} < \inf\{x: F_{X^0}(x) > 0\} \quad \text{and} \\ \sup\{t: S_{T^0}(t) > 0\} < \sup\{x: S_{X^0}(x) > 0\}. \quad (1)$$

These are technical assumptions for identifiability in nonparametric estimation of survival distribution, see, for example, Wang (1989, 1991) among others. In some practical situations, (1) may not be satisfied. In those cases, we can redefine the problem to be estimation of a conditional version of the distribution of interest, as in Gross and Lai (1996).

By decomposing the full likelihood of the observed data into product of conditional and marginal likelihood functions, we are going to obtain a selection bias function for baseline variables under prevalent sampling.

The likelihood function of the observed data (T, Z, Y, δ) can be decomposed into the product of a conditional likelihood L_C and a marginal likelihood L_M

$$L = L_C \times L_M = f_{Y,\delta|T,Z}(y, \delta|t, z) \times f_{T,Z}(t, z), \quad (2)$$

where the conditional likelihood L_C can be expressed as

$$\begin{aligned} L_C &\propto \frac{f_{X^0|T^0,Z^0}(y|t,z)^\delta S_{X^0|T^0,Z^0}(y|t,z)^{1-\delta}}{S_{X^0|T^0,Z^0}(t|t,z)} \\ &= \frac{f_{X^0|Z^0}(y|z)^\delta S_{X^0|Z^0}(y|z)^{1-\delta}}{S_{X^0|Z^0}(t|z)}. \end{aligned} \quad (3)$$

The equality follows from the independent censoring and truncation condition. Furthermore, the marginal likelihood L_M can be expressed as a selection-biased probability density function (Vardi, 1985):

$$L_M = \frac{S_{X^0|Z^0}(t|z) f_{T^0,Z^0}(t, z)}{\int S_{X^0|Z^0}(t^*|z^*) f_{T^0,Z^0}(t^*, z^*) dt^* dz^*}. \quad (4)$$

Here $S_{X^0|Z^0}(t|z)$ plays the role as a covariate-specific selection bias function. The conditional likelihood (3) and the marginal likelihood (4) will be the base for our proposed estimation procedures in the following sections. For notational simplicity, we will write $S^0(\cdot|z)$ for $S_{X^0|Z^0}(\cdot|z)$ in the subsequent sections.

3. Nonparametric Estimation

In this section, we will discuss nonparametric estimation of the distribution of baseline variables Z^0 , when Z^0 is a categorical random variable. Estimation of general discrete or continuous variables will be discussed in the next section under a semiparametric framework.

In Section 2, it was shown that the marginal likelihood of (T, Z) has the form of a selection-biased likelihood. From (4), we have

$$\int \frac{f_{T,Z}(t,z)}{S^0(t|z)} dt dz = \left\{ \int S^0(t|z) f_{T^0,Z^0}(t, z) dt dz \right\}^{-1}$$

and

$$f_{T^0,Z^0}(t, z) = \left\{ \int \frac{f_{T,Z}(t,z)}{S^0(t|z)} dt dz \right\}^{-1} \frac{f_{T,Z}(t, z)}{S^0(t|z)}.$$

If the covariate-specific selection bias function $S^0(\cdot|z)$ is known, it follows from Vardi (1985) that the nonparametric maximum likelihood estimator (NPMLE) of the distribution of (T^0, Z^0) is

$$\widehat{F}_{T^0,Z^0}(t, z; S^0) = \left\{ \sum_{i=1}^n \frac{1}{S^0(T_i|Z_i)} \right\}^{-1} \sum_{i=1}^n \frac{I(T_i \leq t, Z_i \leq z)}{S^0(T_i|Z_i)}.$$

Hence, the NPMLE of the marginal distribution of Z^0 is

$$\begin{aligned}\widehat{F}_{Z^0}(z; S^0) &= \int \widehat{F}_{T^0, Z^0}(dt, z; S^0) \\ &= \left\{ \sum_{i=1}^n \frac{1}{S^0(T_i|Z_i)} \right\}^{-1} \sum_{i=1}^n \frac{I(Z_i \leq z)}{S^0(T_i|Z_i)}.\end{aligned}$$

Note that for one sample selection bias problem, the NPMLE is a weighted empirical distribution with weights inversely proportional to the bias function (Vardi, 1985). It should not be confused with NPMLE for bivariate estimation of censored outcome, which is usually very complicated. In our case, Z is not subject to censoring and the NPMLE has a simpler form. In general, $S^0(\cdot|z)$ is unknown, but we could estimate the distribution function of Z^0 by $\widehat{F}_{Z^0}(z; \widehat{S})$ where $\widehat{S}$ is a consistent estimate of S . When Z is a categorical variable, $S^0(I\{Z_i = z\})$ can be estimated by truncation product limit estimator $\widehat{S}(I\{Z_i = z\})$ using data in the category $\{Z_i = z\}$ (Tsai et al., 1987). That is,

$$\widehat{S}(t|z) = \prod_{s_{(l)} \leq t} \left\{ 1 - \frac{\sum_i I(Y_i = s_{(l)}, \delta_i = 1, Z_i = z)}{\sum_i I(T_i \leq s_{(l)} \leq Y_i, Z_i = z)} \right\}$$

where $s_{(l)}$ are ordered failure times. The proposed estimator of $F_{Z^0}(z)$ is then

$$\widehat{F}_{Z^0}(z; \widehat{S}) = \left\{ \sum_{i=1}^n \frac{1}{\widehat{S}(T_i|Z_i)} \right\}^{-1} \sum_{i=1}^n \frac{I(Z_i \leq z)}{\widehat{S}(T_i|Z_i)}. \quad (5)$$

The estimator (5) is useful when Z^0 contains only a few categories and the sample size for every categories are large. When Z^0 is a continuous variable, we cannot estimate $S^0(\cdot|z)$ consistently by product limit estimator using data at each z , because there is only one observation at each z . In principle, smoothing techniques could be used, but we do not pursue this approach here. Instead, to handle continuous covariates, we introduce a semiparametric model for $S^0(\cdot|z)$. This semiparametric approach is developed in the next section. The semiparametric model can also be usefully applied in the setting of multivariate categorical Z^0 , where the non-parametric estimator is difficult to apply to due to the curse of dimensionality.

4. Semiparametric Estimation under Failure Time Regression Models

In Section 2 we showed that the bias of Z^0 in a prevalent cohort depends on the conditional survival function $S^0(x|z)$. In practice, it is very often that there is also a scientific interest in quantifying the relationship between failure time and baseline variables of interest using a regression model, which specifies $S^0(x|z)$ through a set of (possibly infinite dimensional) parameters. In such cases, estimation efficiency of the distribution of Z^0 can be improved by borrowing strength from the posited regression model. Also, a limitation for the proposed nonparametric estimator is that the variable of interest must be a categorical variable with only a few categories. Semiparametric methods developed in this section are applicable to continuous and multivariate variables. In this section, we consider semiparametric bias corrected estimators under proportional hazards model and accelerated failure time model. The proposed semiparametric methods use regression models for dimensional reduction

(see, e.g., Carroll et al., 1997) and are a practical way for estimation involving multidimensional Z .

4.1 Proportional Hazards Model

Let $h(x; z)$ be the hazard function at time x following an initial event for an individual with covariate value z . In this subsection, we assume the validity of the proportional hazards model:

$$h(x; z) = h_0(x) \exp(\beta z)$$

where $h_0(t)$ is an unspecified baseline hazard function. Let $H_0(t) = \int_0^t h_0(u) du$ be the cumulative baseline hazard. Under this model, the covariate-specific selection bias function is

$$S^0(t|z; H, \beta) = \exp\{-H(t) \exp(\beta z)\}.$$

On the basis of the conditional likelihood L_C and under independent censoring and truncation assumption, left truncation techniques were developed as an extension of the partial likelihood method (Andersen et al., 1993; Wang et al., 1993). The parameter β can be estimated by

$$\hat{\beta} = \arg\max_{\beta} L_p(\beta)$$

where

$$L_p(\beta) = \prod_{i=1}^n \left\{ \frac{\exp(\beta Z_i)}{\sum_{j=1}^n I(T_j \leq Y_i \leq Y_j) \exp(\beta Z_j)} \right\}^{\delta_i}$$

is the partial likelihood. The baseline hazard function can be estimated by a Breslow-type estimator

$$\hat{H}(x; \hat{\beta}) = \sum_{i=1}^n \frac{\delta_i I(Y_i \leq x)}{\sum_j I(T_j \leq Y_i \leq Y_j) \exp(\hat{\beta} Z_j)}.$$

After estimating (β, H) , the joint distribution of (T^0, Z^0) can be estimated by

$$\widehat{F}_{T^0, Z^0}(t, z; \widehat{H}, \widehat{\beta}) = \left[\sum_{i=1}^n \frac{1}{\exp\{-\widehat{H}(T_i) \exp(\widehat{\beta} Z_i)\}} \right]^{-1} \times \sum_{i=1}^n \frac{I(T_i \leq t, Z_i \leq z)}{\exp\{-\widehat{H}(T_i) \exp(\widehat{\beta} Z_i)\}},$$

and the distribution of Z^0 can be estimated by

$$\begin{aligned} \widehat{F}_{Z^0}(z; \widehat{H}, \widehat{\beta}) &= \int \widehat{F}_{T^0, Z^0}(dt, z; \widehat{H}, \widehat{\beta}) \\ &= \left[\sum_{i=1}^n \frac{1}{\exp\{-\widehat{H}(T_i) \exp(\widehat{\beta} Z_i)\}} \right]^{-1} \times \sum_{i=1}^n \frac{I(Z_i \leq z)}{\exp\{-\widehat{H}(T_i) \exp(\widehat{\beta} Z_i)\}}. \end{aligned} \quad (6)$$

The following results summarize the large sample properties of $\widehat{F}_{Z^0}(z; \widehat{H}, \widehat{\beta})$

Theorem 1—Under regularity conditions that guarantee weak convergence of $\sqrt{n}(\widehat{H}(\cdot) - H(\cdot), \widehat{\beta} - \beta)$ (see, e.g., chapter 7 in Andersen et al., 1993) and assuming that (T^0, Z^0) has a finite support, $\sqrt{n}(\widehat{F}_{Z^0}(z; \widehat{H}, \widehat{\beta}) - F_{Z^0}(z; H, \beta))$ converges to a mean zero Gaussian process with a covariance function defined in the appendix.

4.2 Accelerated Failure Time Model

Another popular regression model for failure time data is the accelerated failure time model (Kalbfleisch and Prentice, 1980)

$$\ln X^0 = \gamma Z^0 + \varepsilon,$$

where ε 's are i.i.d. random variables from an unspecified distribution with survival distribution $S^\varepsilon(\cdot)$. Lai and Ying (1991) proposed estimators for γ and S^ε using prevalent cohort data. The target parameter γ can be estimated by solving a log-rank type estimating equation $\xi(\gamma) = 0$, where

$$\xi(\gamma) = \frac{1}{n} \sum_i \delta_i \times \left[Z_i - \frac{\sum_j Z_j I\{T_j \exp(-\gamma Z_j) \leq Y_i \exp(-\gamma Z_i) \leq Y_j \exp(-\gamma Z_j)\}}{\sum_j I\{T_j \exp(-\gamma Z_j) \leq Y_i \exp(-\gamma Z_i) \leq Y_j \exp(-\gamma Z_j)\}} \right].$$

Under this model, the bias function is

$$S^0(t|z; \gamma, S^\varepsilon) = S^\varepsilon\{t \exp(\gamma z)\}.$$

Following the arguments similar to the previous two sections, the distribution of Z^0 can be estimated by

$$\widehat{F}_{z^0}(z) = \left[\sum_{i=1}^n \frac{1}{\widehat{S}^\varepsilon\{T_i \exp(\widehat{\gamma} Z_i)\}} \right]^{-1} \sum_{i=1}^n \frac{I(Z_i \leq z)}{\widehat{S}^\varepsilon\{T_i \exp(\widehat{\gamma} Z_i)\}}.$$

This estimate is constructed without any parametric assumptions on (T^0, Z^0) .

When sampling is done at a particular time, T^0 is the distribution of disease incidence over calendar time and we might sometimes have scientific evidence to model the distribution of T^0 . We next focus on a stationary disease model, which assumes that the occurrence of disease incidence is a stationary Poisson process over time. Stationary models have been studied for prevalent survival data with or without censoring by Wang (1996), Asgharian et al. (2002), Bergeron et al. (2008), and Shen, Ning, and Qin (2009), among others. Interestingly, we can improve the efficiency of distributional estimation of Z_0 with stationarity introduced into the model.

Under the stationary disease incidence assumption, the marginal density of Z is

$$f_z(z) = \frac{E(X|Z=z)f_{z^0}(z)}{\int E(X|Z=z^*)f_{z^0}(z^*)dz^*}, \quad (7)$$

which is also a likelihood function of a selection bias model with bias function $E(X|Z=z)$. Under the accelerated failure time model, we have the following factorization

$$E(X|Z=z) = \exp(\gamma z) \times \int_0^\infty S^\varepsilon(t) dt. \quad (8)$$

Thus,

$$f_z(z) = \frac{\exp(\gamma z)f_{z^0}(z)}{\int \exp(\gamma z)f_{z^0}(z) dz}.$$

Under the stationary disease incidence assumption, we can estimate the distribution of Z^0 by

$$\widehat{F}_{z^0}^S(z; \widehat{\gamma}) = \left(\sum_{i=1}^n \frac{1}{\exp(\widehat{\gamma} Z_i)} \right)^{-1} \sum_{i=1}^n \frac{I(Z_i \leq z)}{\exp(\widehat{\gamma} Z_i)}. \quad (9)$$

Note that (9) is free from nonparametric components because the bias function (8) factorizes into product of a parametric and a nonparametric part, and only the parametric part is a function of z . The nonparametric functions are canceled out, leaving only finite-dimensional parameters γ in the estimator (9). In other models like the semiparametric proportional hazards model, there is no such good factorization property like (8) and we are not able to use (7) to derive estimators for distribution of Z^0 when we have limited study period. It is because the bias function under stationary disease model is a conditional expectation of the failure time involving infinite-dimensional parameters, which may not be identified from incomplete follow-up data, unless strong parametric assumptions on the tail of the survival distribution are assumed.

We have the following large sample properties for $\widehat{F}_{Z^0}^S(z;\widehat{\gamma})$.

Theorem 2—Under regularity conditions that guarantee weak convergence of $\sqrt{n}(\widehat{\gamma}-\gamma)$ (see, e.g., Theorem 1 in Lai and Ying, 1991) and assuming that Z^0 has a finite support, the stochastic process $\sqrt{n}(\widehat{F}_{Z^0}^S(z;\widehat{\gamma})-F_{Z^0}(z;\gamma))$ converges weakly to a mean zero Gaussian process with a covariance function defined in the appendix.

5. Numerical Examples

5.1 Simulations

We examine the validity and finite sample properties of the proposed estimators through a Monte Carlo simulation. To evaluate the validity of the estimators, data are generated 1000 times in the simulation, and each simulated data set consists of 200 or 400 untruncated observations. Pointwise 95% bootstrap confidence intervals for the cumulative distribution function (cdf) are produced based on 2000 bootstrap replications.

We consider scenarios where disease incidence rate is either constant or increasing over time. The incident covariate Z^0 is generated from Uniform(0, 1) distribution. Conditioning on $Z^0 = z$, we generate T^0 from the density

$$f_{T^0}(t) = \begin{cases} 0.5 + \theta z - t\theta z, & 0 \leq t \leq 2 \\ 0, & \text{otherwise} \end{cases}.$$

When $\theta = 0$, the scenario corresponds to stable disease incidence over calendar time. When $\theta > 0$, the disease incidence rate is increasing over calendar time and the rate of increase depends on the covariate $Z^0 = z$. Note that the calendar date of disease incidence is the calendar date of recruitment minus T^0 , so a larger T^0 corresponds to earlier incidence in calendar time. We consider two cases, $\theta = 0$ and 0.5. The survival time X^0 is generated from a Weibull distribution with shape=2 and scale= $\exp(0.125 + 0.5z)$. This corresponds to both an accelerated failure time model with $\gamma = -0.5$ and a proportional hazards model with $\beta = 1$. We also perform simulations for cases corresponding to $\gamma = -0.25, 0.25, 0.5$ ($\beta = 0.5, -0.5, -1$). The conclusions for these cases are similar and we only report detailed results for one case ($\gamma = -0.5, \beta = 1$). We independently generate (ρ^0, Z^0) , and exclude the triplet (ρ^0, Z^0, X^0) if $X^0 < \rho^0$. For each simulated data set, we generate the data until n data points (t, z, x) are included as untruncated data. Then, we generate C' following a uniform (0, 5) distribution and define $y_i = \min(x_i, t_i + c'_i)$ and $\delta_i = I(x_i < t_i + c'_i)$. The observations are $(\delta_i, y_i, t_i, x_i)$, $i = 1, \dots, n$. We consider $n = 200$ and $n = 400$. In each scenario, we compare three estimators: the empirical distribution (EMP), the proposed semiparametric estimator based on proportional hazards model (PH), and the proposed semiparametric estimator based on accelerated failure time model (AFT). The simulation results are shown in Tables 1 and 2.

In Table 1, we consider a case where disease incidence rate is stationary over time. Both proposed estimators show small empirical bias whereas the empirical distribution is biased because of the correlation between survival time and covariate Z^0 . Between the two proposed estimators, the semiparametric estimator based on accelerated failure time model has smaller standard error. A possible reason is that the estimation of the baseline hazard function in proportional hazards model inflated the error in estimating the covariate distribution.

In Table 2, we consider a case where disease incidence was increasing over time. In these cases, the proposed semiparametric estimator under the accelerated failure time model is not consistent in general, because the disease incidence is not stable over time. However, the simulations show that the AFT estimator still has a small empirical bias. The other proposed estimator based on the PH model has negligible bias because their validity does not depend on the stationary disease incidence assumption.

5.2 Data Analysis

We apply the proposed methods to analyze the ovarian cancer cases in the SEER–Medicare linked data (Warren et al., 2002). The SEER data consist of information from cancer patients diagnosed between 1973 and 2002. The Medicare data contain medical cost information between 1986 and 2004. The failure time is defined as the time from diagnosis to death. The SEER–Medicare linked data set was obtained from the National Cancer Institute, and the linked data only include information from SEER subjects who were entitled to Medicare after 1986.

The linkage of the two data sets creates an incident cohort as well as a prevalent cohort of underlying population at different time frames. Patients diagnosed with cancer before 1986 forms a prevalent cohort, because only those patients surviving through 1986 (the start for collection of Medicare data) were included. Patients diagnosed with cancer after 1986 form an incident cohort, because those patients were included at the onset of disease. The prevalent cohort consists of left-truncated and right-censored data and a combination of the two cohorts are also left truncated and right censored. Therefore, the proposed methods can be applied to both cases. However, we note that the two cohorts refer to different underlying population at different time frames and hence the covariate distributions may not be identical.

Although medical cost is not used in our analysis, SEER–Medicare data provide valuable covariate information not covered in SEER and have better quality of survival data than the SEER data (Bach et al., 2002). For instance, in the original SEER data, the date of birth is not recorded and the age of diagnosis is recorded by categories with 5-year increment. With the additional information from the Medicare data, exact age at diagnosis (up to month) can be obtained in the SEER–Medicare data and such information is valuable for data analysis. Also, the date-of-death variable from SEER data is determined from different sources and is less reliable than the date of death collected in the Medicare system.

We first illustrate the degree of bias for the sample proportion of patients diagnosed with local, regional, and distant stages of ovarian cancer. In terms of severity, local stage is the least serious and distant stage is the most severe. We stratify the time of cancer diagnosis into six 5-year intervals and analyzed the corresponding survival data, which are observed subject to right censoring and possibly left truncation (left truncation applicable only to the first three 5-year intervals). The comparisons are shown in Table 3. We can see a huge bias in the prevalent cohorts where a larger proportion of subjects collected are in local and regional stages, because survival performance for subjects in these two stages is much better than subjects in distant stage. After applying bias correction using the nonparametric estimator (5), the estimated proportions for the prevalent cohorts became much similar to the incident cohorts.

We also estimate the joint distribution of ovarian cancer age of diagnosis and stage at diagnosis for elderly persons aged 65 or above in the SEER population between 1973 and 1985. Exploratory analysis suggested that ovarian cancer diagnosis was stable over calendar time. We consider the two semiparametric estimators (6) and (9), respectively, obtained from proportional hazards model and accelerated failure time model. We include age at

diagnosis and two dummy variables corresponding to regional and distant stages at diagnosis as covariates. The results are shown in Figure 2 and Table 4. As in the previous example, there is a huge bias in estimation from the empirical distribution because stage at diagnosis is correlated with disease severity and survival time. Preliminary data analysis reveals that the proportional hazard model does not fit the data well. Thus, the semiparametric estimator based on proportional hazards model may not yield a satisfactory result of bias correction.

6. Conclusion

Prevalent sampling design, although often more practical than incident sampling design for studying survival distribution, is a biased sampling scheme. In this article, we discussed non-parametric and semiparametric estimation of the distribution of baseline variables correcting the bias induced by prevalent sampling design. Nonparametric estimation is proposed that does not rely on any parametric assumption of the conditional survival function but is only applicable to categorical random variables. When a regression model for failure time is specified, semiparametric estimation can be done by borrowing strength from the regression model of scientific interest. For semiparametric estimation, we considered semiparametric proportional hazards model and accelerated failure time model. When stationary disease conditions are satisfied and under the accelerated failure time model, the proposed estimator only involves finite-dimensional parameters, which simplify the estimation and give a better precision. Unless baseline variables are independent of failure time, naive estimation using empirical distribution function yields incorrect estimate for the underlying incident population in general. Although we focused on estimating the marginal distribution function of Z_0 , functionals of the distribution function such as population mean and variance can be readily obtained from the proposed estimators. For instance, $E(Z^0)$ can be estimated by $\widehat{\mu}_{Z^0} = \int z \widehat{F}_{Z^0}(dz)$.

The validity of the semiparametric estimators depend on a suitably chosen model. In the literature of survival analysis, model checking methods have been discussed mainly for right-censored data, though some of the methodological ideas can be readily extended to handle left-truncated right-censored data. Specifically, if the methods are risk set based, we can apply similar methodological steps simply with replacement of risk sets for right-censored data by modified risk sets for left-truncated and right-censored data, under either the proportional hazards model or the accelerated failure time model. Examples of such approaches include the generalized residuals in the sense of Cox and Snell (1968) for the proportional hazards model or the accelerated failure time model, which was adopted by Crowley and Hu (1977) to analyze the Stanford heart transplant data, and the martingale residual technique of Schoenfeld (1980) under the proportional hazards model.

There are two classes of commonly used model checking techniques for survival models. In the first class, two estimating equations with different weight functions are constructed and the two corresponding estimators are compared. The two estimators would converge to the same limit under the null hypothesis that the model is correctly specified. Such methods have been discussed in Lin (1991) for proportional hazards model and in Chen and Jewell (2001) for a general class of hazard regression models that include both proportional hazards and accelerated failure time models. The second type of methods are based on martingale residuals as discussed in Therneau, Grambsch, and Fleming (1990) and in Lin, Wei, and Ying (1993), which focused on checking the proportional hazard assumptions. Martingales can be constructed for accelerated failure time models (Lai and Ying, 1991; Chen and Jewell, 2001), and therefore similar methods can be applied to check the model assumptions.

A reviewer mentioned that sometimes covariate information is collected in a cross-sectional sample but there may not be any follow-up to collect survival data. We note that it is only possible to estimate the incident distribution of the baseline variables based on this type of cross-sectional data if the conditional survival function is known, or under some additional assumptions like a constant rate of incidence and constant survival distribution over calendar time. The work of Keiding (1991) assumes that the disease incidence rate is constant, where our method does not require this assumption. Also, the age-specific mortality rate is not estimated from the cross-sectional data in Keiding (1991); it was treated as a fixed function known from external data. Keiding (2006) also mentioned estimation of age distribution using current status data under stationarity conditions. The estimate in that paper is cube-root- n consistent. Our estimator uses the information from survival time is square-root- n consistent.

We used the SEER–Medicare data as an illustration of the proposed method, and the data set includes a prevalent cohort due to linkage of two different data sets. Linking two or more data sets had become popular in public health research. The method developed is general and can be applicable to cases where a survival data set is linked to a cross-sectional data set containing other covariate information. The covariate information from a cross-sectional data is biased for estimation of incident covariate distribution and the proposed method is applicable to remove bias in estimating incident covariate distribution.

Acknowledgments

The authors thank Prof David Zucker, an associate editor, and two anonymous referees for their helpful comments on this article. The first author was partially supported by the National Institutes of Health grant R01 AI 089341. The second author was partially supported by the National Institutes of Health grants R01 AI 078835 and P50 CA 098252.

References

- Andersen, P.; Borgan, O.; Gill, R.; Keiding, N. Statistical Models Based on Counting Process. New York: Springer-Verlag; 1993.
- Asgharian M, M'Lan C, Wolfson D. Length-biased sampling with right censoring: An unconditional approach. *Journal of the American Statistical Association*. 2002; 97:201–210.
- Bach P, Guadagnoli E, Schrag D, Schussler N, Warren J. Patient demographic and socioeconomic characteristics in the SEER-Medicare database: Applications and limitations. *Medical Care*. 2002; 40:IV-19–IV25.
- Bergeron P, Asgharian M, Wolfson D. Covariate bias induced by length-biased sampling of failure times. *Journal of the American Statistical Association*. 2008; 103:737–742.
- Brookmeyer R, Gail M. Biases in prevalent cohorts. *Biometrics*. 1987; 43:739–749. [PubMed: 3427161]
- Carroll R, Fan J, Gijbels I, Wand M. Generalized partially linear single-index models. *Journal of the American Statistical Association*. 1997; 92:477–489.
- Chen Y, Jewell N. On a general class of semiparametric hazards regression models. *Biometrika*. 2001; 88:687–702.
- Chen Y, Hu C, Wang Y. Attributable risk function in the proportional hazards model for censored time-to-event. *Biostatistics*. 2006; 7:515–529. [PubMed: 16478758]
- Cox D, Snell E. A general definition of residuals. *Journal of the Royal Statistical Society, Series B (Methodological)*. 1968; 30:248–275.
- Crowley J, Hu M. Covariance analysis of heart transplant survival data. *Journal of the American Statistical Association*. 1977; 72:27–36.
- Gross S, Lai T. Nonparametric estimation and regression analysis with left-truncated and right-censored data. *Journal of the American Statistical Association*. 1996; 91:1166–1180.

- Gürler, U.; Gijbels, I. Technical Report. Catholique de Louvain Institut de Statistique; 1996. A bivariate distribution function estimator and its variance under left truncation and right censoring.
- Huang Y, Wang C. Cox regression with accurate covariates unascertainable: A nonparametric-correction approach. *Journal of the American Statistical Association*. 2000; 95:1209–1219.
- Kalbfleisch, J.; Prentice, R. *The Statistical Analysis of Failure Time Data*. New York: Wiley; 1980.
- Keiding N. Age-specific incidence and prevalence: A statistical perspective. *Journal of the Royal Statistical Society, Series A (Statistics in Society)*. 1991; 154:371–412.
- Keiding N. Event history analysis and the cross-section. *Statistics in Medicine*. 2006; 25:2343–2364. [PubMed: 16708345]
- Lai T, Ying Z. Rank regression methods for left-truncated and right-censored data. *The Annals of Statistics*. 1991; 19:531–556.
- Lin D. Goodness-of-fit analysis for the Cox regression model based on a class of parameter estimators. *Journal of the American Statistical Association*. 1991; 86:725–728.
- Schoenfeld D. Chi-squared goodness-of-fit tests for the proportional hazards regression model. *Biometrika*. 1980; 67:145–153.
- Shen Y, Ning J, Qin J. Analyzing length-biased data with semiparametric transformation and accelerated failure time models. *Journal of the American Statistical Association*. 2009; 104:1192–1202. [PubMed: 21057599]
- Therneau T, Grambsch P, Fleming T. Martingale-based residuals for survival models. *Biometrika*. 1990; 77:147–160.
- Tsai WY, Jewell N, Wang MC. A note on the product-limit estimator under right censoring and left truncation. *Biometrika*. 1987; 74:883–886.
- van der Varrrt, A.; Wellner, J. *Weak Convergence and Empirical Processes*. New York: Springer; 1996.
- Vardi Y. Empirical distributions in selection bias models. *The Annals of Statistics*. 1985; 13:178–203.
- Wang M. A semiparametric model for randomly truncated data. *Journal of the American Statistical Association*. 1989; 84:742–748.
- Ying Z. A large sample study of rank estimation for censored regression data. *The Annals of Statistics*. 1993; 21:76–99.
- Vardi Y. Multiplicative censoring, renewal processes, de-convolution and decreasing density: Nonparametric estimation. *Biometrika*. 1989; 76:751–761.
- Wang M. Nonparametric estimation from cross-sectional survival data. *Journal of the American Statistical Association*. 1991; 86:130–143.
- Wang M. Hazards regression analysis for length-biased data. *Biometrika*. 1996; 83:343–354.
- Wang M, Brookmeyer R, Jewell N. Statistical models for prevalent cohort data. *Biometrics*. 1993; 49:1–11. [PubMed: 8513095]
- Warren J, Klabunde C, Schrag D, Bach P, Riley G. Overview of the SEER-Medicare data: Content, research applications, and generalizability to the United States elderly population. *Medical Care*. 2002; 40:3–18.

APPENDIX

Proof of Theorem 1

Suppose the following regularity conditions hold:

- i. $(T_i, Z_i, Y_i, \delta_i)$ are i.i.d.
- ii. $\inf\{t: F_{T^0}(t) > 0\} < \inf\{x: F_{X^0}(x) > 0\}$ and $\sup\{t: S_{T^0}(t) > 0\} < \sup\{x: S_{X^0}(x) > 0\}$.
- iii. $P(C_i - \tau) > 0, i = 1, \dots, n$, where t is a predetermined constant.
- iv. The support of Z is finite.

v. $E[\int_0^\tau \{Z - \frac{s^{(1)}(t)}{s^{(0)}(t)}\}^{\otimes 2} I(C \geq t \geq T) \exp(\beta Z) h_0(t) dt]$ is finite, where

$$s^{(k)}(t) = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n I(T_i \leq t \leq C_i) Z_i^{\otimes k} \exp(\beta Z_i),$$

$k = 0, 1, 2$, and for any column vector a , $a^{\otimes 0} = 1$, $a^{\otimes 1} = a$, and $a^{\otimes 2} = aa^T$.

Let $\widehat{F}_{T,Z}(t, z) = \sum_{i=1}^n I(T_i \leq t, Z_i \leq z)/n$ and $\varphi = \int \int \exp(H(t) \exp(\beta z)) F_{T,Z}(dt, dz)$.

First, consider

$$\begin{aligned} & \frac{n^{-1/2} \sum_{i=1}^n \exp(\widehat{H}(T_i) \exp(\widehat{\beta} Z_i)) I(Z_i \leq z)}{n^{-1} \sum_{i=1}^n \exp(\widehat{H}(T_i) \exp(\widehat{\beta} Z_i))} - \frac{n^{-1/2} \sum_{i=1}^n \exp(H(T_i) \exp(\beta Z_i)) I(Z_i \leq z)}{n^{-1} \sum_{i=1}^n \exp(H(T_i) \exp(\beta Z_i))} \\ &= \frac{n^{-1/2} \sum_{i=1}^n (\exp(\widehat{H}(T_i) \exp(\widehat{\beta} Z_i)) - \exp(H(T_i) \exp(\beta Z_i))) I(Z_i \leq z)}{n^{-1} \sum_{i=1}^n \exp(\widehat{H}(T_i) \exp(\widehat{\beta} Z_i))} \\ &+ \left(\frac{1}{n^{-1} \sum_{i=1}^n \exp(\widehat{H}(T_i) \exp(\widehat{\beta} Z_i))} - \frac{1}{n^{-1} \sum_{i=1}^n \exp(H(T_i) \exp(\beta Z_i))} \right) \\ &\quad \times n^{-1/2} \sum_{i=1}^n \exp(H(T_i) \exp(\beta Z_i)) I(Z_i \leq z) \\ &= \frac{\sqrt{n} \int \int I(z' \leq z) (\exp(\widehat{H}(t') \exp(\widehat{\beta} z')) - \exp(H(t') \exp(\beta z'))) \widehat{F}_{T,Z}(dt', dz')}{\sqrt{n} \int \int (\exp(\widehat{H}(t') \exp(\widehat{\beta} z')) - \exp(H(t') \exp(\beta z'))) \widehat{F}_{T,Z}(dt', dz')} \\ &\quad - \frac{\varphi^2}{\varphi^2} \varphi F_{Z^0}(z) + o_p(1) \\ &= \int \int [I(z' \leq z) - F_{Z^0}(z)] \exp(\beta z') \sqrt{n} (\widehat{H}(t') - H(t')) + z' \exp(\beta z') H(t') \sqrt{n} (\widehat{\beta} - \beta) F_{T,Z}(t', z') + o_p(1). \end{aligned} \tag{A.1}$$

Also,

$$\begin{aligned} & \frac{n^{-1/2} \sum_{i=1}^n \exp(H(T_i) \exp(\beta Z_i)) I(Z_i \leq z)}{n^{-1} \sum_{i=1}^n \exp(H(T_i) \exp(\beta Z_i))} - F_{Z^0}(z) \\ &= n^{-1/2} \sum_{i=1}^n [\varphi^{-1} \exp(H(T_i) \exp(\beta Z_i)) I(Z_i \leq z) - F_{Z^0}(z)] + o_p(1). \end{aligned} \tag{A.2}$$

By arguments as in chapter 7 of Andersen et al. (1993) and Theorems 1–2 in Huang and Wang (2000), $\widehat{\beta}$ and $\widehat{H}(t)$ each admits an asymptotic i.i.d. representation, denote by

$\sqrt{n}(\widehat{\beta} - \beta) = n^{-1/2} \sum_{i=1}^n \xi_{\beta,i} + o_p(1)$ and $\sqrt{n}(\widehat{H}(t) - H(t)) = n^{-1/2} \sum_{i=1}^n \xi_{H,i}(t) + o_p(1)$, respectively.

Because $\sqrt{n}(\widehat{F}_{Z^0}(z; \widehat{\beta}) - F_{Z^0}(z))$ equals the sum of (A.1) and (A.2), so the i.i.d. representation follows with

$$d_i(z) = \left\{ \int \exp(\beta z') [I(z' \leq z) - F_{z^0}(z)] \xi_{H,i}(t') + z' H(t') \xi_{\beta,i} F_{T^0, z^0}(dt', dz') + \varphi^{-1} \exp(H(T_i) \exp(\beta Z_i)) I(Z_i \leq z) - F_{z^0}(z) \right\}.$$

Because $E(\xi_{\beta,i}) = 0$, $E(\xi_{H,i}(t)) = 0$ and $\varphi^{-1} E(\exp(H(T_i) \exp(\beta Z_i)) I(Z_i \leq z)) = F_{z^0}(z)$, the process $n^{-1} \sum_{i=1}^n d_i(z)$ has mean zero. Also, because the i.i.d. representation can be expressed as a linear combination of monotone functions of z , weak convergence follows from the example 2.11.16 of van der Vaart and Wellner (1996). The covariance function of the limiting Gaussian process is $E(d_1(z_1) d_1(z_2))$, for z_1, z_2 within the support of Z^0 .

Proof of Theorem 2

Suppose the following regularity conditions hold:

- i. $(T_i, Z_i, Y_i, \delta_i)$ are i.i.d.
- ii. $\inf\{t: F_{T^0}(t) > 0\} < \inf\{x: F_{X^0}(x) > 0\}$ and $\sup\{t: S_{T^0}(t) > 0\} < \sup\{x: S_{X^0}(x) > 0\}$.
- iii. The random variable ε is continuously distributed, with a bounded density function f with bounded derivative, and

$$\int \{f'(t)/f(t)^2\} f(t) dt < \infty.$$

- iv. Censoring time C has a bounded density function.
- v. The support of Z is finite.
- vi. $\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n E\{Z_i^{\otimes r} P(\ln T_i - \gamma Z_i \leq s \leq \ln C_i - \gamma Z_i | Z_i)\}$ and $\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n P(\ln T_i - \gamma Z_i \leq \ln C_i - \gamma Z_i < s)$ exist for all s , for $r = 0, 1, 2$.

Let $\eta = E(e^{-\gamma Z})$ and $\widehat{F}_Z(z) = (\sum_{i=1}^n I(Z_i \leq z))/n$. First, consider

$$\begin{aligned}
& \frac{n^{-1/2} \sum_{i=1}^n \exp(-\widehat{\gamma} Z_i) I(Z_i \leq z)}{n^{-1} \sum_{i=1}^n \exp(-\widehat{\gamma} Z_i)} \\
& - \frac{n^{-1/2} \sum_{i=1}^n \exp(-\gamma Z_i) I(Z_i \leq z)}{n^{-1} \sum_{i=1}^n \exp(-\gamma Z_i)} \\
& = \frac{n^{-1/2} \sum_{i=1}^n (\exp(-\widehat{\gamma} Z_i) - \exp(-\gamma Z_i)) I(Z_i \leq z)}{n^{-1} \sum_{i=1}^n \exp(-\widehat{\gamma} Z_i)} \quad (\text{A.3}) \\
& + \left(\frac{1}{n^{-1} \sum_{i=1}^n \exp(-\widehat{\gamma} Z_i)} - \frac{1}{n^{-1} \sum_{i=1}^n \exp(-\gamma Z_i)} \right) \\
& \quad \times n^{-1/2} \sum_{i=1}^n \exp(-\gamma Z_i) I(Z_i \leq z) \\
& = \frac{\sqrt{n} \int I(z' \leq z) (\exp(-\widehat{\gamma} z') - \exp(-\gamma z')) \widehat{F}_z(dz')}{\frac{\eta}{\sqrt{n} \int (\exp(-\widehat{\gamma} z') - \exp(-\gamma z')) \widehat{F}_z(dz')} \eta F_{z^0}(z) + o_p(1)} \\
& = -\sqrt{n}(\widehat{\gamma} - \gamma) \int z' (I(z' \leq z) - F_{z^0}(z)) F_{z^0}(dz') + o_p(1).
\end{aligned}$$

Also,

$$\frac{n^{-1/2} \sum_{i=1}^n \exp(-\gamma Z_i) I(Z_i \leq z)}{n^{-1} \sum_{i=1}^n \exp(-\gamma Z_i)} - F_{z^0}(z) = n^{-1/2} \sum_{i=1}^n [\eta^{-1} \exp(-\gamma Z_i) I(Z_i \leq z) - F_{z^0}(z)] + o_p(1) \quad (\text{A.4})$$

By arguments as in Theorems 1 and 2 in Lai and Ying (1991) and Theorems 1 and 2 in Ying (1993), γ follows an asymptotic i.i.d. representation $\sqrt{n}(\widehat{\gamma} - \gamma) = n^{-1/2} \sum_{i=1}^n \xi_{\gamma,i} + o_p(1)$.

Because $\sqrt{n}(\widehat{F}_{z^0}^S(z; \widehat{\gamma}) - F_{z^0}(z))$ equals the sum of (A.3) and (A.4), the i.i.d. representation follows with

$$e_i(z) = \left[- \left(\int z' (I(z' \leq z) - F_{z^0}(z)) F_{z^0}(dz') \right) \xi_{\gamma,i} + \eta^{-1} \exp(-\gamma Z_i) I(Z_i \leq z) - F_{z^0}(z) \right].$$

Because $E(\xi_{\gamma,i}) = 0$ and $\eta^{-1} E(\exp(-\gamma Z) I(Z \leq z)) = F_{z^0}(z)$, the process $n^{-1} \times \sum_{i=1}^n e_i(z)$ has mean zero. Also, because the i.i.d. representation involves a linear combination of monotone functions of z , weak convergence follows from the example 2.11.16 of van der Vaart and Wellner (1996). The covariance function of the limiting Gaussian process is $E(e_1(z_1) e_1(z_2))$, for z_1, z_2 within the support of Z^0 .

Figure 1.

A graphical illustration of a prevalent sample. Each line represents an individual from disease diagnosis to death. Subjects alive at the time of recruitment ($X_0 - T_0$) is collected, indicated by bold lines.

Figure 2.

Estimation of the joint distribution of age at diagnosis and stage at diagnosis. EMP (—) represents the empirical distribution estimate, PH (—) represents the proposed semiparametric estimator based on proportional hazards model, and AFT (—) represents the proposed estimator based on accelerated failure time model.

Table 1

Simulation results of estimating the cdf under stable disease incidence ($\theta = 0$). EMP represents the empirical distribution estimate, PH represents the proposed semiparametric estimator based on proportional hazards model, AFT represents the proposed estimator based on accelerated failure time model. Bias represents the empirical bias, SSE represents the sampling standard error, and C.I.L and C.I.U represent the averaged lower and upper limits of 95% confidence intervals, respectively.

Estimator		0.2	0.4	0.6	0.8
(a) n = 200					
EMP	Bias	-0.036	-0.054	-0.055	-0.036
	Bias	0.002	0.000	0.000	-0.002
	SSE	0.076	0.078	0.065	0.043
	C.I.L	0.103	0.271	0.477	0.708
	C.I.U	0.358	0.565	0.732	0.878
AFT	Bias	0.003	0.006	0.005	0.005
	SSE	0.033	0.037	0.037	0.028
	C.I.L	0.141	0.329	0.530	0.747
	C.I.U	0.268	0.482	0.677	0.857
(b) n = 400					
EMP	Bias	-0.035	-0.054	-0.054	-0.036
	Bias	0.001	0.001	0.001	0.001
	SSE	0.044	0.051	0.046	0.032
	C.I.L	0.130	0.308	0.508	0.729
	C.I.U	0.299	0.509	0.692	0.859
AFT	Bias	0.001	0.002	0.002	0.002
	SSE	0.022	0.027	0.026	0.020
	C.I.L	0.159	0.349	0.549	0.760
	C.I.U	0.245	0.455	0.653	0.841

\$watermark-text

\$watermark-text

\$watermark-text

Table 2

Simulation results of estimating the cdf under nonstable disease incidence($\theta = 0.5$). EMP represents the empirical distribution estimate, PH represents the proposed semiparametric estimator based on proportional hazards model. AFT represents the proposed estimator based on accelerated failure time model. Bias represents the empirical bias, SSE represents the sampling standard error, and C.I.L and C.I.U represent the averaged lower and upper limits of 95% confidence intervals, respectively.

Estimator		0.2	0.4	0.6	0.8
(a) $n = 200$					
EMP	Bias	-0.053	-0.079	-0.078	-0.051
	Bias	-0.001	0.001	0.003	0.001
	SSE	0.082	0.081	0.065	0.040
	C.I.L	0.101	0.278	0.491	0.722
	C.I.U	0.347	0.557	0.724	0.874
AFT	Bias	-0.020	-0.025	-0.021	-0.012
	SSE	0.031	0.039	0.038	0.029
	C.I.L	0.118	0.297	0.503	0.730
	C.I.U	0.242	0.453	0.655	0.846
(b) $n = 400$					
EMP	Bias	-0.054	-0.080	-0.078	-0.052
	Bias	0.000	0.000	-0.001	-0.001
	SSE	0.058	0.056	0.043	0.028
	C.I.L	0.122	0.308	0.518	0.745
	C.I.U	0.316	0.518	0.690	0.853
AFT	Bias	-0.019	-0.023	-0.019	-0.010
	SSE	0.022	0.027	0.026	0.021
	C.I.L	0.137	0.323	0.529	0.748
	C.I.U	0.225	0.431	0.633	0.832

\$watermark-text

\$watermark-text

\$watermark-text

Table 3
Estimation of population proportion of stage at diagnosis for elderly ovarian cancer patients

Year	Sample proportion			Bias adjusted estimate (95% CI)		
	Localized (%)	Regional (%)	Distant (%)	Localized (%)	Regional (%)	Distant (%)
1973–1977	72	14	15	22 (18–27)	10 (7–13)	68 (61–73)
1978–1982	27	14	29	18 (16–21)	8 (7–10)	73 (70–76)
1983–1987	23	8	69	18 (17–19)	7 (6–8)	74 (72–75)
1988–1992	21	7	72		—	
1993–1997	20	7	73		—	
1998–2002	15	7	78		—	

Table 4

Estimation of the joint distribution of age at diagnosis and stage at diagnosis using the proposed methods. EMP represents the empirical distribution estimate, PH represents the proposed semiparametric estimator based on proportional hazards model, and AFT represents the proposed estimator based on accelerated failure time model

Stage/Age		65–74		75–84		85+	
		Estimate	95% C.I.	Estimate	95% C.I.	Estimate	95% C.I.
Localized	EMP	0.299	(0.278–0.319)	0.145	(0.130, 0.161)	0.018	(0.013, 0.025)
	PH	0.222	(0.189–0.252)	0.129	(0.108–0.151)	0.027	(0.014–0.045)
	AFT	0.153	(0.131–0.174)	0.101	(0.085–0.118)	0.019	(0.012–0.028)
Regional	EMP	0.073	(0.061, 0.084)	0.039	(0.031, 0.047)	0.004	(0.002, 0.007)
	PH	0.064	(0.049–0.079)	0.042	(0.029–0.061)	0.010	(0.002–0.026)
	AFT	0.051	(0.040–0.063)	0.034	(0.024–0.044)	0.005	(0.001–0.009)
Distant	EMP	0.270	(0.251, 0.289)	0.121	(0.107, 0.136)	0.030	(0.022, 0.037)
	PH	0.267	(0.236–0.296)	0.176	(0.133–0.217)	0.062	(0.017–0.112)
	AFT	0.353	(0.319–0.386)	0.217	(0.186–0.246)	0.066	(0.046–0.091)