Published in final edited form as:
Biometrika. 2013 ; 100(3): . doi:10.1093/biomet/ast008.

Survival analysis without survival data: connecting length-biased and case-control data

KWUN CHUEN GARY CHAN
Department of Biostatistics, University of Washington, Seattle, Washington 98195, U.S.A.
kcgchan@u.washington.edu

Summary

We show that relative mean survival parameters of a semiparametric log-linear model can be estimated using covariate data from an incident sample and a prevalent sample, even when there is no prospective follow-up to collect any survival data. Estimation is based on an induced semiparametric density ratio model for covariates from the two samples, and it shares the same structure as for a logistic regression model for case-control data. Likelihood inference coincides with well-established methods for case-control data. We show two further related results. First, estimation of interaction parameters in a survival model can be performed using covariate information only from a prevalent sample, analogous to a case-only analysis. Furthermore, propensity score and conditional exposure effect parameters on survival can be estimated using only covariate data collected from incident and prevalent samples.

Keywords

Accelerated failure time model; Biased sampling; Empirical likelihood; Prevalent cohort; Propensity score; Proportional mean residual life model

1. Introduction

Survival analysis methodologies are designed for analysing time-to-event data. Usually, a study records survival data as well as covariate information for incident cases over a certain period of time. In most situations, survival data are only partially observed subject to right censoring. Many common survival analysis methods are designed for analysing right-censored survival data, such as the Kaplan & Meier (1958) estimator and the proportional hazards model (Cox, 1972).

Collecting survival data from incident cases usually requires a long study period to gather enough events for meaningful analysis. Alternatively, one might sample from a disease-prevalent population cross-sectionally at a particular calendar time (Wang, 1991). A cross-sectional study that draws samples from a disease-prevalent population is more focused and economical than an incident study design (Wang, 1991; Wang et al., 1993). Cross-sectional sampling yields length-biased survival outcome when the disease incidence is stationary over time (Wang, 1991; Asgharian et al., 2002). Regression models for length-biased survival data have been discussed by Wang (1996), Bergeron et al. (2008), Shen et al. (2009), Chen (2010), Mandel & Ritov (2010), Qin & Shen (2010), Huang et al. (2012) and Chan et al. (2012) among others.

While regression models for length-biased survival data are widely studied, all existing methods require the survival time to be partially observable subject to right censoring. In contrast to the existing literature, the main results of this paper concern identifiability and estimation of relative mean survival parameters by comparing covariate distributions from

an unbiased and a length-biased sample, without follow-up to collect survival data. Studies can be designed to estimate survival parameters without a high cost of follow-up. In a typical regression analysis, the marginal distribution of covariates does not contain any information about regression parameters of interest when the sampling mechanism is unbiased. Under length-biased sampling, however, covariate values associated with longer survivors are preferentially sampled. We show that, in the absence of time-varying covariates, relative mean survival parameters in a class of semipara-metric log-linear models can be identified solely by comparing covariate distributions of an incident and a prevalent cohort.

2. Data model and likelihood inference

Suppose T is the survival outcome of interest and X is a p-vector of explanatory variables. We assume X are baseline variables that are not time-varying. The joint distribution of (T, X) in a population of interest follows a probability model

$$f_{T,X}(t,x;\theta) = f_{T|X}(t|x;\theta)\, f_X(x)$$

where $\theta \in \Theta \subset \mathbb{R}^p$ is a vector of parameters of interest. For a length-biased sample, the sampling distribution of (T, X) is

$$f^{\mathrm{S}}_{T,X}(t,x;\theta) = \frac{t f_{T,X}(t,x;\theta)}{\mu}$$

where $\mu = E(T)$. It follows that the sampling distribution of X is

$$f^{\mathrm{S}}_{X}(x) = \frac{E(T|X=x;\theta)\, f_X(x)}{\mu} \quad (1)$$

where $E(T \mid X = x;\ \theta) = \int t\, f_{T|X}(t \mid x;\ \theta)\, \mathrm{d}t$ (Bergeron et al., 2008; Chan & Wang, 2012). That is, the sampling distribution of covariates is proportional to the conditional mean of the survival outcome, which depends on regression parameters θ. Both Bergeron et al. (2008) and Chan & Wang (2012) considered right-censored length-biased data, and showed that (1) is the sampling distribution of covariate X in the presence of right censoring. Since X is a baseline variable and censoring happens only after an individual has been sampled, it is clear that the sampling distribution of X does not depend on the censoring distribution.

In standard regression analysis, it is usually optimal to maximize a conditional likelihood function for the outcome given covariates because the marginal likelihood function of covariates is typically strongly ancillary (Cox & Hinkley, 1974, pp. 31–5), since $f_X(x)$ does not involve any regression parameters of interest. Under length-biased sampling, however, the marginal sampling distribution of covariates involves the regression parameters, as seen in (1), because the sampling bias depends on the relationship between survival outcome and covariates. When the covariate distribution is not parameterized, the marginal likelihood of covariates under length-biased sampling is nonparametric weakly ancillary, meaning that the marginal likelihood is constant in θ after profiling $f_X(x)$. This can be shown using arguments similar to those of Wang (1989) and Wang et al. (1993). When the population covariate distribution $f_X(x)$ is parameterized, the joint likelihood of (T, X) under length-biased sampling can improve efficiency over the conditional likelihood in finite samples but not asymptotically, as shown by Bergeron et al. (2008). However, parameterizing the incident covariate distribution may be too restrictive when multivariate covariates are considered. In

the following discussions, we assume $f_X(x)$ to be nonparametric while estimating finite-dimensional parameters θ.

Suppose we collect independent data from an incident cohort and from a prevalent cohort. For example, the time from dementia onset to death was studied in the Canadian Study of Health and Aging (Wolfson et al., 2001). The study randomly sampled individuals throughout Canada and followed a prevalent subsample consisting of demented persons at the baseline visit. The maximal follow-up period was five years. Based on prevalent survival data, Wolfson et al. (2001) concluded that the survival time for female subjects were significantly longer than that of male subjects. Suppose that additional incident data, perhaps from a disease registry, can be collected from those who were free from dementia at baseline, but developed dementia during the five-year study period. Suppose that we observe that the proportion of women in the prevalent cohort is greater than that in the incident cohort. This information alone can lead to a conclusion that female subjects lived longer, even when survival endpoints are unobserved.

Let $(x_1, \ldots, x_p)$ denote independent and identically distributed covariate data in a prevalent sample having density $f_X^S(x)$ in (1), and let $(x_{p+1}, \ldots, x_n)$ denote independent and identically distributed covariate data from an incident sample. We assume the following log-linear model for population mean survival:

$$\log\{E(T|X)\} = \alpha^* + \beta^{\mathrm{T}} X. \quad (2)$$

Special cases of model (2) include an accelerated failure time model (Kalbfleisch & Prentice, 2002, pp. 44–5; Cox & Oakes, 1984, pp. 64–5), $\log T = \beta^{\mathrm{T}} X + \varepsilon$ where X and ε are independent and a proportional mean residual life model (Oakes & Dasu, 1990), $E(T - t \mid T \geq t, X = x) = m_0(t) \exp(\beta^{\mathrm{T}} x)$. Furthermore, model (2) can allow heteroscedastic errors, which is more general than the accelerated failure time model; existing rank-based estimation procedures (Tsiatis, 1990) cannot handle heteroscedastic errors. Models for marginal mean survival are usually nonidentifiable for right-censored survival data with a limited study period. However, relative mean survival parameters can be estimated in our setting without observing survival outcomes. We do not consider the proportional hazards model (Cox, 1972) in this paper, because its conditional mean survival function depends on both the baseline hazard and relative hazard parameters. It is unclear how the functional nuisance parameter can be eliminated without any observation of failure events.

Under the log-linear model (2), the covariate sampling distribution under length-biased sampling is

$$f_X^S(x) = \frac{\exp\left(\alpha^* + \beta^{\mathrm{T}} x\right) f_X(x)}{\mu} = \exp\left(\alpha + \beta^{\mathrm{T}} x\right) f_X(x) \quad (3)$$

where $\alpha = \alpha^* - \log(\mu)$. Model (3) is called an exponential tilted density ratio model (Qin & Zhang, 1997) and has been used as the basis of semiparametric estimation for case-control data. Let $D = \{0, 1\}$ be a case-control status and assume the logistic regression model

$$\mathrm{pr}(D=1|X) = \frac{\exp\left(\alpha^* + \beta^{\mathrm{T}} X\right)}{1 + \exp(\alpha^* + \beta^{\mathrm{T}} X)}. \quad (4)$$

It can be shown by applying Bayes' Theorem that $f(x \mid D = 1) = \exp(\alpha + \beta^{\mathrm{T}} x) f(x \mid D = 0)$ where $\alpha = \alpha^* - \log\{\mathrm{pr}(D = 1)/\mathrm{pr}(D = 0)\}$. Therefore, the probability structure incident and

prevalent data under model (2) is the same as case-control data under logistic regression model (4).

The likelihood function based on $(x_1, \ldots, x_n)$ is equivalent to the retrospective likelihood for logistic regression (Prentice & Pyke, 1979; Qin & Zhang, 1997). Following their arguments, we obtain a semi-parametric profile likelihood

$$\log L(\alpha, \beta) = \sum_{i=1}^{p} \left(\alpha + \beta^{\mathrm{T}} x_i\right) - \sum_{i=1}^{n} \left[1 + \{p/(n-p)\} \exp\left(\alpha + \beta^{\mathrm{T}} x_i\right)\right]. \quad (5)$$

This loglikelihood function is the same as that for case-control data under the logistic regression model (Prentice & Pyke, 1979). It was further shown by Prentice & Pyke (1979) and Qin & Zhang (1997) that the profile likelihood satisfies the usual properties of an ordinary likelihood function. Based on this likelihood equivalence and the equivalence of prospective and retrospective analysis for case-control data, the parameter β for the semiparametric log-linear survival model can be estimated by maximizing log L using commonly available software for logistic regression, as follows. Let $Y_i = 1$ for $i = 1, \ldots, p$ and $Y_i = 0$ for $i = p + 1, \ldots, n$. Maximizing the likelihood function for a logistic regression model treating Y as an outcome and X as explanatory variables is equivalent to maximizing (5). Standard logistic regression programs would give valid standard error estimates for $\widehat{\beta}$.

3. Estimation of interaction parameters from prevalent data

In medical applications, it is common that an effect of an exposure variable could be modified by other factors. Under gene-environment independence, a case-only design would yield consistent estimates for odds-ratio interaction parameters in a logistic regression model (Piegorsch et al., 1994). By exploiting the relationship between case-control and length-biased data, we will show that interaction parameters in a survival model can be estimated using data only from a prevalent sample.

We assume that

$$\log\{E(T|X_1, X_2)\} = \alpha^* + \beta_1 X_1 + \beta_2 X_2 + \beta_I X_1 X_2,$$

where X_1 is a binary exposure variable and X_2 can be discrete, continuous or a mixture of both. The main scientific interest is the estimation of β_I. If X_1 and X_2 are independent in the population, then it follows from (1) that

$$f^{\mathrm{S}}_{X_1, X_2}(x_1, x_2) = \exp(\alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_I x_1 x_2) f_{X_1}(x_1) f_{X_2}(x_2).$$

Furthermore,

$$f^{\mathrm{S}}_{X_2|X_1}(x_2|X_1 = 0) = \frac{\exp(\alpha + \beta_2 x_2) f_{X_2}(x_2)}{\int \exp(\alpha + \beta_2 x_2) f_{X_2}(x_2)\, \mathrm{d}x_2}$$

and

$$f^{\mathrm{S}}_{X_2|X_1}(x_2|X_1=1)=\frac{\exp\{\alpha+\beta_1+(\beta_2+\beta_I)x_2\}f_{X_2}(x_2)}{\int\exp\{\alpha+\beta_1+(\beta_2+\beta_I)x_2\}f_{X_2}(x_2)\,\mathrm{d}x_2}.$$

Therefore, the conditional covariate distribution for X_2 given X_1 follows a density ratio model

$$\frac{f^{\mathrm{S}}_{X_2|X_1}(x_2|X_1=1)}{f^{\mathrm{S}}_{X_2|X_1}(x_2|X_1=0)}=\exp(\tilde{\alpha}+\beta_I x_2)$$

and the conditional distribution of X_1 given X_2 follows a logistic regression model

$$f^{\mathrm{S}}_{X_1|X_2}(X_1=1|X_2=x_2)=\frac{\exp\left(\alpha^{\dagger}+\beta_I x_2\right)}{1+\exp(\alpha^{\dagger}+\beta_I x_2)}$$

where $\alpha^{\dagger}=\tilde{\alpha}+\log\{\mathrm{pr}(X=1)/\mathrm{pr}(X=0)\}$. Therefore, the interaction parameter can be estimated by fitting a logistic regression model using the prevalent sample only, treating X_1 as a binary outcome and X_2 as an explanatory variable.

The prevalent-only analysis has two advantages for estimating β_I . First, it does not require additional data collection from an incident population. Second, it has improved estimation efficiency compared to the estimation from maximizing (5) using both incident and prevalent samples. This is analogous to the improvement in efficiency for the estimation of odds-ratio interaction by case-only analysis (Piegorsch et al., 1994). The main drawback, similar to the case-only analysis, is that the estimator is biased when X_1 and X_2 are dependent. The bias-variance trade-off can be optimized by using empirical Bayes estimation for combining case-control and case-only estimators (Mukherjee & Chatterjee, 2008).

4. Propensity score and conditional treatment effect

Suppose A is a binary exposure variable. In an observational study, exposure is not randomized and the effect of A on survival is likely to be confounded by additional covariates X. The confounding relationship can be complex, and we assume that

$$\log\{E(T|A,X)\}=\alpha+\beta_A A+g(X), \quad (6)$$

where $g(\cdot)$ is an unspecified function and the parameter β_A is the main interest. When the confounding relationship is complex, so $g(\cdot)$ does not admit a known parametric form, an alternative way to estimate β_A is by propensity score subclassification or matching (Rosenbaum & Rubin, 1984). Under length-biased sampling and model (6), we establish the relationship between A and the propensity score $\pi(X)=\mathrm{pr}(A=1\mid X)$, and show that both β_A and propensity score parameters can be estimated without observing survival data. This contrasts with a recent paper by Cheng & Wang (2012) that shows a similar relationship, but their estimation requires the survival outcome to be observable.

The sampling distribution of (A, X) for a length-biased observation is

$$f^{\mathrm{S}}_{A,X}(a,x)=\frac{\exp\{\alpha+\beta_A a+g(x)\}\, f_{A|X}(a|x)\, f(x)}{\sum_{a=0,1}\int \exp\{\alpha+\beta_A a+g(x)\}\, f_{A|X}(a|x)\, f(x)\,\mathrm{d}x}.$$

The sample conditional probability of $A = 1$ given $X = x$ is

$$\mathrm{pr}^{\mathrm{S}}(A=1|X=x)=\frac{f^{\mathrm{S}}_{A,X}(1,x)}{f^{\mathrm{S}}_{A,X}(1,x)+f^{\mathrm{S}}_{A,X}(0,x)}=\frac{\exp(\beta_A)\, f_{A|X}(1|x)}{\exp(\beta_A)\, f_{A|X}(1|x)+f_{A|X}(0|x)}=\frac{\exp\{\beta_A+\gamma(x)\}}{1+\exp\{\beta_A+\gamma(x)\}} \quad (7)$$

where $\gamma(x) = \log[\pi(x)/\{1 - \pi(x)\}]$ is the log odds of the propensity score.

If the propensity score follows a logistic regression model

$$\mathrm{pr}(A=1|X=x)=\frac{\exp\left(\gamma_0+\gamma_X^{\mathrm{T}}x\right)}{1+\exp\left(\gamma_0+\gamma_X^{\mathrm{T}}x\right)}, \quad (8)$$

then $\gamma(x)=\gamma_0+\gamma_X^{\mathrm{T}}x$ and

$$\mathrm{pr}^{\mathrm{S}}(A=1|X=x)=\frac{\exp\left(\beta_A+\gamma_0+\gamma_X^{\mathrm{T}}x\right)}{1+\exp\left(\beta_A+\gamma_0+\gamma_X^{\mathrm{T}}x\right)}. \quad (9)$$

Expressions (7) and (9) lead to two methods for estimating β_A, depending on whether or not the propensity score is known. First, with known propensity score, it follows from (7) that β_A is the intercept term in a logistic regression model for A given X with an offset term $\gamma(X)$, using covariate information only from a prevalent sample. When $\pi(x)$ is unknown but is modelled by logistic regression model (8), β_A and $\pi(x)$ can be estimated simultaneously from a combination of incident and prevalent samples. Let Y be a prevalent sample status indicator, with $Y = 1$ corresponding to a prevalent observation and $Y = 0$ corresponding to an incident observation. Combining (8) and (9) we have

$$\mathrm{pr}(A=1|Y=y,X=x)=\frac{\exp\left(\gamma_0+\beta_A y+\gamma_X^{\mathrm{T}}x\right)}{1+\exp\left(\gamma_0+\beta_A y+\gamma_X^{\mathrm{T}}x\right)}$$

which again follows a logistic regression model and both β_A of model (6) and (γ_0, γ_X) of model (8) can be estimated simultaneously.

5. Simulation studies

We conducted simulation studies to examine the finite sample properties of the proposed estimators in § § 2–4. For each simulation scenario, 5000 independent datasets were generated. Each dataset consists of an incident cohort and a prevalent cohort both having n observations, with $n = 50, 100, 200$.

We considered the setting in § 2 in the first simulation study. We generated a $U(0, 1)$ variable X. In the first case, a homoscedastic error ε was generated from a centred Gaussian distribution with variance σ^2, where $\sigma = 0{\cdot}5$. In the second case, a heteroscedastic error ε

was generated from a centred Gaussian distribution with variance $X\sigma^2$. In both cases, the logarithmic survival time was $\log T = \beta_1^* X + \epsilon$ and the mean survival time followed a log-linear model $\log E(T \mid X) = \beta_1 X$, where $\beta_1 = \beta_1^*$ under homoscedasticity and $\beta_1 = \beta_1^* + \sigma^2/2$ under heteroscedasticity. We considered cases where $\beta_1 = 0$ and 1·5. Residual censoring time C, defined as the time from recruitment to censoring, was generated from a $U(0, 2)$ distribution. We compared the proposed estimator $\widehat{\beta}_{\mathrm{CC}}$ with the solution $\widehat{\beta}_{\mathrm{LR}}$ of a log-rank estimating equation using only incident survival data (Tsiatis, 1990). The log-rank estimating equation was expected to yield inconsistent estimates for β_1 or β_1^* when the error term was heteroscedastic. Table 1 shows that the proposed estimator had small bias and the log-rank estimating equation was biased under heteroscedasticity. We also performed Wald tests for testing the hypothesis $H_0 : \beta_1 = 0$ at 5% significance level. The test based on the proposed estimator had correct empirical Type I error and adequate power for both cases, while the test based on a log-rank estimating equation had incorrect size under heteroscedasticity.

To study the estimator of interaction parameters as discussed in § 3, we generated X_1 from a Bernoulli distribution with $p = 0{\cdot}1$ or 0·5, and X_2 from a standard normal distribution. The survival time T was generated from an exponential distribution with mean $\exp(\beta_I X_1 X_2)$, where $\beta_I = 1$. The residual censoring time was generated from a $U(0, 5)$ distribution. We compared three estimators for β_I : the proposed case-control estimator $\widehat{\beta}_{\mathrm{CC}}$ in § 2 using data from both incident and prevalent cohorts, the proposed case-only estimator $\widehat{\beta}_{\mathrm{CO}}$ in § 3 using data from the prevalent cohort, and the solution $\widehat{\beta}_{\mathrm{LR}}$ to a log-rank estimating equation using data from the incident cohort. Table 2 showed that the proposed case-only estimator gained efficiency by recognizing the independence relationship between X_1 and X_2. When $p = 0{\cdot}1$, $X_1 = 1$ is uncommon in the population and the estimate $\widehat{\beta}_{\mathrm{LR}}$ using only incident cohort data had low efficiency. Prevalent sampling preferentially samples individuals with $X = 1$ and estimators using information from prevalent data were more efficient than those using information only from incident data.

Next, we studied the performance of the propensity score methods proposed in § 4. Two covariates (X_1, X_2) were generated from $U(0, 1)$ and $N(0, 2)$ distributions. Exposure A was generated by a propensity score model $\mathrm{logit}\{\mathrm{pr}(A = 1 \mid X_1, X_2) = \gamma_0 + \gamma_1 X_1 + \gamma_2 X_2$, where $(\gamma_0, \gamma_1, \gamma_2) = (-1, 2, 2)$ and survival time followed $\log T = \beta_0 + \beta_A A - |X_1 X_2 - 0{\cdot}5| + \varepsilon$ where ε was standard normal. Censoring followed a $U(0, 2)$ distribution. We compared three estimators for β_A: the estimator $\widehat{\beta}_{\mathrm{LR,MIS}}$ which is the solution to a log-rank estimating equation based on a misspecified model $\log T = \beta_0 + \beta_A A + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$; the proposed estimator assuming the propensity score is known, $\widehat{\beta}_{\mathrm{known}}$; and the proposed estimator with unknown propensity score, β_{unknown}. Table 3 shows that the two propensity score methods had a small bias, while the estimator based on a misspecified regression model was biased.

Acknowledgments

The author thanks the editor, an associate editor, two reviewers, and Drs Ron Brookmeyer, Jim Hughes and Mary Lou Thompson for their helpful comments and suggestions, which greatly improved this paper. This research was partially supported by grants from the National Institutes of Health, U.S.A.

REFERENCES

Asgharian M, M'lan C, Wolfson DB. Length-biased sampling with right censoring: An unconditional approach. J. Am. Statist. Assoc. 2002; 97:201–10.

Bergeron PJ, Asgharian M, Wolfson DB. Covariate bias induced by length-biased sampling of failure times. J. Am. Statist. Assoc. 2008; 103:737–42.

Chan KCG, Chen YQ, Di C. Proportional mean residual life model for right-censored length-biased data. Biometrika. 2012; 99:995–1000. [PubMed: 23843676]

Chan KCG, Wang M-C. Estimating incident population distribution from prevalent data. Biometrics. 2012; 68:521–31. [PubMed: 22313264]

Chen YQ. Semiparametric regression in size-biased sampling. Biometrics. 2010; 66:149–58. [PubMed: 19432792]

Cheng Y-J, Wang M-C. Estimating propensity scores and causal survival functions using prevalent survival data. Biometrics. 2012; 68:707–16. [PubMed: 22834993]

Cox DR. Regression models and life-tables (with discussion). J. R. Statist. Soc. B. 1972; 34:187–220.

Cox, DR.; Hinkley, DV. Theoretical Statistics. Chapman & Hall; London: 1974.

Cox, DR.; Oakes, D. Analysis of Survival Data. Chapman & Hall; London: 1984.

Huang C-Y, Qin J, Follmann DA. A maximum pseudo-profile likelihood estimator for the Cox model under length-biased sampling. Biometrika. 2012; 99:199–210. [PubMed: 23843659]

Kalbfleisch, JD.; Prentice, RL. The Statistical Analysis of Failure Time Data. 2nd ed.. Wiley; New York: 2002.

Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. J. Am. Statist. Assoc. 1958; 53:457–81.

Mandel M, Ritov Y. The accelerated failure time model under biased sampling. Biometrics. 2010; 66:1306–8. [PubMed: 19995351]

Mukherjee B, Chatterjee N. Exploiting gene-environment independence for analysis of case–control studies: An empirical Bayes-type shrinkage estimator to trade-off between bias and efficiency. Biometrics. 2008; 64:685–94. [PubMed: 18162111]

Oakes D, Dasu T. A note on residual life. Biometrika. 1990; 77:409–10.

Piegorsch WW, Weinberg CR, Taylor JA. Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. Statist. Med. 1994; 13:153–62.

Prentice RL, Pyke R. Logistic disease incidence models and case-control studies. Biometrika. 1979; 66:403–11.

Qin J, Shen Y. Statistical methods for analyzing right-censored length-biased data under Cox model. Biometrics. 2010; 66:382–92. [PubMed: 19522872]

Qin J, Zhang B. A goodness-of-fit test for logistic regression models based on case-control data. Biometrika. 1997; 84:609–18.

Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. J. Am. Statist. Assoc. 1984; 79:516–24.

Shen Y, Ning J, Qin J. Analyzing length-biased data with semiparametric transformation and accelerated failure time models. J. Am. Statist. Assoc. 2009; 104:1192–202.

Tsiatis AA. Estimating regression parameters using linear rank tests for censored data. Ann. Statist. 1990; 18:354–72.

Wang M-C. A semiparametric model for randomly truncated data. J. Am. Statist. Assoc. 1989; 84:742–8.

Wang M-C. Nonparametric estimation from cross-sectional survival data. J. Am. Statist. Assoc. 1991; 86:130–43.

Wang M-C. Hazards regression analysis for length-biased data. Biometrika. 1996; 83:343–54.

Wang M-C, Brookmeyer R, Jewell NP. Statistical models for prevalent cohort data. Biometrics. 1993; 49:1–11. [PubMed: 8513095]

Wolfson C, Wolfson DB, Asgharian M, M'lan CE, Østbye T, Rockwood K, Hogan DB. A reevaluation of the duration of survival after the onset of dementia. New Engl. J. Med. 2001; 344:1111–6. [PubMed: 11297701]

Table 1

Comparisons between the proposed estimators and log-rank estimating equation under homoscedasticity and heteroscedasticity

		Homoscedasticity						Heteroscedasticity					
		$\hat{\beta}_{CC}$			$\hat{\beta}_{LR}$			$\hat{\beta}_{CC}$			$\hat{\beta}_{LR}$		
	n	**(a)**	**(b)**	**(c)**	**(a)**	**(b)**	**(c)**	**(a)**	**(b)**	**(c)**	**(a)**	**(b)**	**(c)**
$\beta = 0$	50	–15	825	5	–8	331	6	10	688	4	–257	221	20
	100	8	549	6	–7	225	5	4	574	5	–260	153	33
	200	17	374	4	3	154	5	5	567	6	–261	105	65
$\beta = 1{\cdot}5$	50	34	761	55	119	955	46	41	716	57	–478	448	38
	100	45	525	85	62	459	69	29	519	85	–546	238	46
	200	77	376	99	66	289	95	11	370	99	–578	149	72

(a) Bias $\times 10^3$.

(b) SSE $\times 10^3$, where SEE represents the sampling standard deviation.

(c) R%, which represents the proportion of Wald tests rejected for $H_0 : \beta_1 = 0$ at the 5% significance level.

Table 2

Comparisons among the proposed estimators and log-rank estimating equation for interaction parameters

		β_{CO}		β_{CC}		β_{LR}	
	n	**Bias × 10^3**	**SSE × 10^3**	**Bias × 10^3**	**SSE × 10^3**	**Bias × 10^3**	**SSE × 10^3**
$p = 0{\cdot}1$	50	79	164	103	360	−76	502
	100	22	104	58	197	51	328
	200	19	74	58	198	147	203
$p = 0{\cdot}5$	50	7	113	31	159	−38	139
	100	19	76	12	105	47	85
	200	7	52	10	72	−8	57

SSE, the sampling standard deviation.

Table 3

Comparisons among estimators under a propensity score model

	$\hat{\beta}_{\mathrm{LR,MIS}}$		$\hat{\beta}_{\mathbf{known}}$		$\hat{\beta}_{\mathbf{unknown}}$	
n	**Bias × 10^3**	**SSE × 10^3**	**Bias × 10^3**	**SSE × 10^3**	**Bias × 10^3**	**SSE × 10^3**
50	357	603	66	526	93	848
100	315	392	23	386	41	572
200	295	282	22	262	42	376

SSE, the sampling standard deviation.