Published in final edited form as:

Biometrics. 2018 June ; 74(2): 575–583. doi:10.1111/biom.12772.

Analysis of Restricted Mean Survival Time for Length-Biased Data

Chi Hyun Lee*, Jing Ning, and Yu Shen

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

Summary

In clinical studies with time-to-event outcomes, the restricted mean survival time (RMST) has attracted substantial attention as a summary measurement for its straightforward clinical interpretation. When the data are subject to length-biased sampling, which is frequently encountered in observational cohort studies, existing methods to estimate the RMST are not applicable. In this paper, we consider nonparametric and semiparametric regression methods to estimate the RMST under the setting of length-biased sampling. To assess the covariate effects on the RMST, a semiparametric regression model that directly relates the covariates and the RMST is assumed. Based on the model, we develop unbiased estimating equations to obtain consistent estimators of covariate effects by properly adjusting for informative censoring and length bias. Stochastic process theories are used to establish the asymptotic properties of the proposed estimators. We investigate the finite sample performance through simulations and illustrate the methods by analyzing a prevalent cohort study of dementia in Canada.

Keywords

Length-biased data; nonparametric estimation; semiparametric regression method; restricted mean survival time

1. Introduction

In clinical studies with time-to-event outcomes, the restricted mean survival time (RMST) is a useful summary measurement. It captures the expected survival time within a specific time period τ , which is evaluated by the area under the survival function over $[0, \tau]$. This measurement can serve as an alternative to the overall mean or the median survival time. The RMST has attracted great attention for its transparent and intuitive interpretation, especially in two-sample problems. Recent literature has pointed out that a hazard ratio, which is the foundation of conventional survival analysis when comparing multiple groups or assessing treatment effects, may be difficult to interpret in clinical settings, regardless of the validity of the proportional hazards assumption (Rudser, LeBlanc, and Emerson, 2012; Tian, Zhao, and

---

* clee9@mdanderson.org.

Supplementary Materials

Web Appendices referenced in Sections 3, 4.2, 4.3, 5.1, 5.2, and 7, and the R codes implementing the proposed methods are available with this paper at the Biometrics website on Wiley Online Library.

Wei, 2014; Uno et al., 2014). While the RMST difference or ratio describes the absolute magnitude of risk or benefit in survival, the hazard ratio characterizes the relative risk of an event between groups. The interpretation of the hazard ratio becomes more complicated when the proportional hazards assumption is violated. Royston and Parmar (2011), Zhao et al. (2012), and Royston and Parmar (2013) proposed using the RMST in place of the hazard ratio in comparative clinical studies when the proportional hazards assumption was in doubt. Furthermore, if the event of interest is rare, the results based on the proportional hazards model may not be reliable, while the estimator of the RMST difference or ratio is robust. Trinquart et al. (2016) compared the performance of the hazard ratio and the RMST ratio in 54 randomized controlled trials. The measures based on the RMST were found to be more clinically relevant than those based on the hazard ratio, and were robust to assumptions and the number of observed events.

Extensive studies have been conducted on the RMST, first introduced by Irwin (1949), for conventional survival data. In Karrison (1997), the RMST was used as an index for comparing survival in two groups. Recently, Zhao et al. (2016) examined the inferential procedures for the RMST curve over a specific time duration based on simultaneous confidence intervals, and expanded them to evaluate the difference between two groups. The RMST was also extended to incorporate covariates under the Cox regression model (Karrison, 1997; Zucker, 1998; Chen and Tsiatis, 2001; Zhang and Schaubel, 2011). To assess the covariate effects on the RMST directly, Andersen, Hansen, and Klein (2004) proposed using pseudo observations based on generalized linear models with a robust variance estimator. The mean of the pseudo observations was constructed to be equal to the RMST. More recently, Tian et al. (2014) assumed the RMST to have a linear relationship with the covariates through a link function that allowed for covariate-specific RMST prediction. They followed the least squares principle with an inverse probability censoring weight to estimate the covariate effects.

Despite the progression of statistical tools for inference on the RMST, existing methods cannot be readily used to analyze length-biased right-censored data. Length-biased data are often encountered in prevalent cohort studies, in which subjects who have a specific disease or condition at the time of enrollment are followed prospectively for a failure event. Because of this unique sampling mechanism, subjects in the sampled cohort are not representative of the target population. In addition, informative censoring induced by the sampling scheme needs to be properly addressed. While estimation based on the Cox proportional hazards model under the setting of length-biased sampling is well-established (Wang, 1996; Tsai, 2009; Qin and Shen, 2010; Qin et al., 2011), there is no formal analytic tool in the literature for checking the validity of the proportional hazards assumption. Analysis based on the RMST would be desirable, especially when one lacks confidence in the proportional hazards assumption.

In this article, we generalize the nonparametric estimation and the semiparametric regression analysis of the RMST to analyze length-biased survival data. The remainder of this paper is organized as follows. In Section 2, we briefly introduce length-biased data with notations. In Section 3, we present the nonparametric estimation of the RMST and the large sample properties. In Section 4, we propose a semiparametric regression method based on a model

that directly relates the baseline covariates and the RMST. The asymptotic properties of the estimators are established therein. In Section 5, we investigate the finite sample properties through simulation studies under various settings. We apply the proposed estimation methods to a prevalent cohort study of dementia in Canada in Section 6. We conclude with a discussion of our methods.

2. Length-Biased Data and Notations

Assume that patients with a specific disease are enrolled into a prevalent cohort study and followed prospectively until either the occurrence of a failure event or the end of the study. Denote $\tilde{T}$ as the duration from the disease onset to the failure event (referred to as the “unbiased failure time”), with the density function $f_{\tilde{T}}(t) = dF_{\tilde{T}}(t)/dt$ and the corresponding survival function $S_{\tilde{T}}(t)$. Let $\tilde{A}$ be the duration from the initiation of the disease until study enrollment (referred to as the “backward recurrence time”). Under this study design, the failure time, $\tilde{T}$, is not randomly sampled from the target population and is only observable for patients sampled with bias. Specifically, a patient who experienced the failure event prior to study enrollment cannot be sampled. Thus, patients must have $\tilde{A} < \tilde{T}$ to be enrolled in the cohort. The sampled data are subject to length bias since longer failure times are more likely to be sampled from the target population. We denote the sampled failure time as T (referred to as the “biased failure time”) and the corresponding backward recurrence time as A . For the sampled patients, we measure the duration from enrollment to the failure event, which is denoted as V , and have $T = A + V$. Under the setting of length-biased sampling, the density function of the biased failure time T is $f(t) = dF(t)/dt = tf_{\tilde{T}}(t)/\mu$, where $\mu = \int_0^\infty S_{\tilde{T}}(u)du$ and the corresponding survival function $S(t)$. Since V is subject to right censoring, we observe event time $Y = \min(T, A + C)$ and censoring indicator $\delta = I(T \leq A + C)$, where C is the residual censoring time (i.e., the duration from enrollment to censoring). The length-biased data consist of independent and identically distributed (i.i.d.) (Y_i, A_i, δ_i) for $i = 1, \dots, n$. The censoring time C and (A, V) are assumed to be independent. Nevertheless, failure times are subject to dependent censoring induced by length-biased sampling since $\text{Cov}(T, A + C) = \sigma_A^2(1 + \rho_{A,V}\sigma_V/\sigma_A) > 0$ except for trivial cases, where $\sigma_A^2 = \text{Var}(A)$, $\sigma_V^2 = \text{Var}(V)$, and $\rho_{A,V} = \text{corr}(A, V)$ (Asgharian and Wolfson, 2005).

We define τ as the restricted time point of interest from an initiating event. Let the restricted survival time be $\tilde{T}_\tau = \min(\tilde{T}, \tau)$ for a fixed time point τ . Then, the RMST is defined as $\mu_\tau = E(\tilde{T}_\tau) = \int_0^\tau S_{\tilde{T}}(u)du$. Note that we use the subscript τ to distinguish the RMST from the overall mean survival time μ . While we are interested in the restricted survival time in the target population, only the biased failure times are observable. The observed restricted event time is $\min(Y, \tau)$. Unless otherwise specified, we use these notations throughout this paper.

3. Nonparametric Estimation of RMST

In the analysis of conventional survival data, the RMST can be easily estimated by integrating the Kaplan–Meier estimator of the survival distribution over the time period $[0, \tau]$. However, the Kaplan–Meier estimator for the length-biased data results in overestimation

of the survival distribution of $\tilde{T}$ (Wolfson et al., 2001; Asgharian and Wolfson, 2005). Thus, naively applying the Kaplan–Meier estimator in RMST estimation would lead to biased estimates.

Under the setting of length-biased sampling, the likelihood of the observed data is proportional to

$$\prod_{i=1}^n \frac{\{f_U(Y_i)\}^{\delta_i} \{S_U(Y_i)\}^{1-\delta_i}}{\mu} \propto \prod_{i=1}^n \{dF(Y_i)\}^{\delta_i} \prod_{i=1}^n \left\{ \int_{Y_i}^{\infty} \frac{dF(u)}{u} \right\}^{1-\delta_i}.$$

Vardi (1989) proposed an expectation-maximization algorithm to compute the nonparametric maximum likelihood estimation of the biased distribution function, $F(t)$. The estimator for $S_U(t)$ (referred to as “Vardi’s estimator”) then can be derived through the relationship between $S_U(t)$ and $S(t)$, $\hat{S}_U(t) = \int_t^{\infty} u^{-1} d\hat{S}(u) / \int_0^{\infty} u^{-1} d\hat{S}(u)$, where $\hat{S}(t) = 1 - \hat{F}(t)$ and $\hat{F}(t)$ is the estimator of $F(t)$. The asymptotic properties of Vardi’s estimator were established in Asgharian, M’Lan, and Wolfson (2002) and Asgharian and Wolfson (2005).

Using Vardi’s estimator, the RMST can be estimated by $\hat{\mu}_{\tau} = \int_0^{\tau} \hat{S}_U(t) dt$. We summarize the weak convergence of the estimator, $\hat{\mu}_{\tau}$, as follows.

Theorem 1

Under regularity conditions (A₁) and (A₂) listed in Web Appendix A.1, $\sqrt{n}(\hat{\mu}_{\tau} - \mu_{\tau})$ converges weakly to a normal distribution with mean zero and variance $\sigma^2 = \mu^2 E \left[\mu_{\tau} \int_0^{\infty} u^{-1} dW(u) - \int_0^{\tau} \int_t^{\infty} u^{-1} dW(u) dt \right]^2$.

The definition of $W(t)$ and a detailed proof are provided in Web Appendix A.2.

When a two-sample comparison is of primary interest, the RMST can be used to describe the difference between the two groups. The difference in the RMST between two groups can be estimated by the area between the survival curves, that is,

$$\hat{\mu}_{\tau}^{(1)} - \hat{\mu}_{\tau}^{(2)} = \int_0^{\tau} \{ \hat{S}_U^{(1)}(u) - \hat{S}_U^{(2)}(u) \} du, \text{ where superscripts (1) and (2) are the group indicators.}$$

A statistical test can be constructed based on the estimator accordingly.

Note that we gain some efficiency by assuming that $\tilde{A}$ follows a uniform distribution (i.e., the stationarity assumption) in the length-biased sampling design. When the stationarity assumption is violated but the distribution of $\tilde{A}$ is known, we may use the estimator discussed by Mandel (2007) instead of Vardi’s estimator for RMST estimation. For the setting of more general left-truncation, the nonparametric method can be established by computing the area under the product-limit estimator that maximizes the conditional likelihood (Tsai, Jewell, and Wang, 1987; Wang, 1991). Since the conditional product-limit estimator does not use the distribution information on truncation time, our estimator is expected to be more efficient relative to the method for general left-truncated data. We make empirical comparison between the two methods in Section 5.1.

4. Semiparametric Regression Analysis of RMST

4.1 Model

Let the covariates of interest be denoted by a $p \times 1$ vector $\mathbf{X}$. Given the covariates, the density function of the unbiased failure time $\tilde{T}$ is denoted by $f_U(t|\mathbf{x})$ and the corresponding survival function by $S_U(t|\mathbf{x})$. It follows that the covariate-specific length-biased sampling density is $f(t|\mathbf{x}) = tf_U(t|\mathbf{x})/\mu(\mathbf{x})$ where $\mu(\mathbf{x}) = \int_0^\infty S_U(u|\mathbf{x})du$. For $i = 1, \dots, n$, we observe i.i.d. $(Y_i, A_i, \delta_i, X_i)$. The residual censoring C is assumed to be independent of (A, V) . While it can be easily relaxed to the conditional independence given the covariates, we keep this stringent assumption for the sake of simplicity. In Section 4.2, we briefly describe how we can relax this assumption and account for dependency between C and the covariates.

The RMST adjusted for the covariates can be defined as $\mu_\tau(\mathbf{x}) = E(\tilde{T}_\tau|\mathbf{x})$. Following the model specified for conventional survival data in Tian et al. (2014), we assume that the RMST of the failure times in the target population has the following relationship with the covariates for a fixed τ that satisfies $\Pr(Y > \tau) > 0$:

$$\eta\{\mu_\tau(\mathbf{x})\} = \mathbf{Z}^\top \boldsymbol{\beta}_\tau \quad (1)$$

where $\eta(\cdot)$ is a smooth, strictly increasing link function, $\boldsymbol{\beta}_\tau$ is a $(p+1) \times 1$ coefficient vector specific to τ , and $\mathbf{Z} = (1, \mathbf{X}^\top)^\top$. Under this model, the RMST is directly related to the covariates. The model only postulates that the relationship between the covariates and the transformed RMST via the link function is linear. Some special forms of link functions including $\eta(\cdot) = \log(\cdot)$ and $\eta(m) = \log\{m/(\tau - m)\}$ are commonly used in applications. Using these specific links, the exponential of the estimated regression coefficient of a treatment indicator can be interpreted as the ratio and the odds ratio of the RMST, respectively, for a two-sample case. A simple alternative is an identity link function, in which the RMST itself is linearly related to the covariates. In this case, the coefficient estimates can be translated directly to the increase in the RMST per unit increase in the corresponding covariate with all other covariates fixed. Although the interpretation is straightforward, one should be cautious to use the identity link function because it may lead to a negative RMST prediction.

4.2 Estimating Procedure

The estimating approach by Tian et al. (2014) for conventional survival data is not directly applicable for length-biased data because the observed biased failure times do not follow the model that is assumed for the target population. We generalize their method to length-biased data by proposing unbiased estimating equations that use the relationship between distributions of the underlying failure time and the biased failure time.

We first assume there is no censored observation in the data. The joint densities of (A, V) and (A, T) are shown to be equal given the covariates as follows (Asgharian and Wolfson, 2005):

$$f_{A,V}(a, v | \mathbf{X} = \mathbf{x}) = f_{A,T}(a, t | \mathbf{x}) = f_U(a + v | \mathbf{x})I(a > 0, v > 0)/\mu(\mathbf{x}). \quad (2)$$

There are three exclusive possibilities: (i) an event occurs before the restricted time point τ (i.e., $Y \leq \tau$); (ii) an event occurs after τ and the backward recurrence time is shorter than τ (i.e., $A < \tau < Y$); and (iii) the backward recurrence time is longer than or equal to τ (i.e., $A \geq \tau$). An example of these exclusive cases is illustrated in Figure 1.

By using the joint distribution of (A, V) for the observed data in equation (2), we have the following expectations for the first two cases:

$$E[I(Y \leq \tau) | \mathbf{X} = \mathbf{x}] = \int_0^\tau \int_0^y \frac{f_U(a + v | \mathbf{x})I(a > 0, v > 0)}{\mu(\mathbf{x})} da dv = \int_0^\tau \frac{y f_U(a + v | \mathbf{x})}{\mu(\mathbf{x})} dy, \quad (3)$$

$$\begin{aligned} E[I(A < \tau < Y) | \mathbf{X} = \mathbf{x}] &= \int_\tau^\infty \int_0^\tau \frac{f_U(a + v | \mathbf{x})I(a > 0, v > 0)}{\mu(\mathbf{x})} da dv \\ &= \int_\tau^\infty \frac{\tau f_U(y | \mathbf{x})}{\mu(\mathbf{x})} dy = \frac{\tau S_U(\tau | \mathbf{x})}{\mu(\mathbf{x})}. \end{aligned} \quad (4)$$

Under the third case ($A \geq \tau$), we can derive the following expectation by the density function of the backward recurrence time A given the covariates, $f_A(a | \mathbf{x}) = S_U(a | \mathbf{x})/\mu(\mathbf{x})$:

$$E[\tau I(A \geq \tau) | \mathbf{X} = \mathbf{x}] = \int_\tau^\infty \frac{\tau S_U(a | \mathbf{x})}{\mu(\mathbf{x})} da = \frac{\tau \{\mu(\mathbf{x}) - \mu_\tau(\mathbf{x})\}}{\mu(\mathbf{x})}. \quad (5)$$

It can be easily observed that the sum of equations (3) and (4) is

$$\int_0^\tau \frac{y f_U(y | \mathbf{x})}{\mu(\mathbf{x})} dy + \frac{\tau S_U(\tau | \mathbf{x})}{\mu(\mathbf{x})} = \frac{\mu_\tau(\mathbf{x})}{\mu(\mathbf{x})}.$$

Under Model (1), it can be shown that the following expectation for length-biased data without right censoring equals zero.

$$E\left[I(Y \leq \tau) + I(A < \tau < Y) + \tau I(A \geq \tau) - \frac{(1-\tau)}{\mu(\mathbf{X})} \eta^{-1} (\mathbf{Z}^\top \boldsymbol{\beta}_\tau) - \tau \mid \mathbf{X} = \mathbf{x}\right] = 0.$$

Thus, we can construct the following estimating equation for $\boldsymbol{\beta}_\tau$ when there is no right censoring:

$$\sum_{i=1}^n z_i \left\{ I(y_i \leq \tau) + I(a_i < \tau < y_i) + \tau I(a_i \geq \tau) - \frac{(1-\tau)}{\mu(\mathbf{x}_i)} \eta^{-1} (\mathbf{z}_i^\top \boldsymbol{\beta}_\tau) - \tau \right\} = 0.$$

When the length-biased data are right-censored, the informative censoring induced by the sampling scheme must be handled properly. Assuming that the censoring time C is independent of the covariates, we have

$$\Pr(Y \in (y, y + dy), A \in (a, a + da), C \geq y - a | \mathbf{X} = \mathbf{x}) = f_{A, V}(a, y - a | \mathbf{x}) S_C(y - a) da dy \quad (6)$$

where S_C is the survival function of the residual censoring time C . We again consider the three possible cases. By applying equation (6), we can derive the following probabilities for the first two cases. For $Y > \tau$,

$$\Pr(Y \in (y, y + dy), \delta = 1 | \mathbf{X} = \mathbf{x}) = \frac{f_U(y | \mathbf{x})}{\mu(\mathbf{x})} \int_0^y S_C(v) dv dy.$$

When $A < \tau < Y$,

$$\Pr(Y \in (y, y + dy), \delta = 1 | \mathbf{X} = \mathbf{x}) = \frac{f_U(y | \mathbf{x})}{\mu(\mathbf{x})} \int_0^\tau S_C(y - a) da dy = \frac{f_U(y | \mathbf{x})}{\mu(\mathbf{x})} \int_{y-\tau}^y S_C(v) dv dy.$$

Thus, the following expectations can be derived.

$$E \left[\frac{Y \delta I(Y \leq \tau)}{w(Y)} \middle| \mathbf{X} = \mathbf{x} \right] = \int_0^\tau \frac{y f_U(y | \mathbf{x})}{w(y) \mu(\mathbf{x})} \int_0^y S_C(v) dv dy = \int_0^\tau \frac{y f_U(y | \mathbf{x})}{\mu(\mathbf{x})} dy \quad (7)$$

$$E \left[\frac{\tau \delta I(A < \tau < Y)}{w_\tau(Y)} \middle| \mathbf{X} = \mathbf{x} \right] = \tau \int_\tau^\infty \frac{f_U(y | \mathbf{x})}{w_\tau(y) \mu(\mathbf{x})} \int_{y-\tau}^y S_C(v) dv dy = \frac{\tau S_C(\tau | \mathbf{x})}{\mu(\mathbf{x})}, \quad (8)$$

where $w(y) = \int_0^y S_C(v) dv$ and $w_\tau(y) = \int_{y-\tau}^y S_C(v) dv$. Note that the observed restricted event time is equal to τ when $A < \tau < Y$. When $A \leq \tau$, the same expectation as in equation (5) can be used. By combining equations (5), (7), and (8), we can construct an unbiased estimating equation for $\boldsymbol{\beta}_\tau$ as follows.

$$\sum_{i=1}^n z_i \left\{ \frac{y_i \delta_i I(y_i \leq \tau)}{w(y_i)} + \frac{\tau \delta_i I(a_i < \tau < y_i)}{w_\tau(y_i)} + \tau I(a_i \geq \tau) - \frac{(1-\tau)}{\mu(\mathbf{x}_i)} \eta^{-1} (\mathbf{z}_i^\top \boldsymbol{\beta}_\tau) - \tau \right\} = 0. \quad (9)$$

There are three unknown quantities in equation (9): the weight functions $w(t)$ and $w_\tau(t)$, and the covariate-specific overall mean survival time $\mu(\mathbf{x})$. Under the independent censoring assumption, the weight function can be consistently estimated by $\hat{w}(t) = \int_0^t \hat{S}_C(v) dv$ and $\hat{w}_\tau(t) = \int_{t-\tau}^t \hat{S}_C(v) dv$, where $\hat{S}_C$ is the Kaplan–Meier estimator of the residual censoring time. $\hat{S}_C$ can be estimated by treating $Y_i - A_i$ as the event time and $1 - \delta_i$ as the event indicator. Model (1) implies that the regression model for the covariate-specific overall mean survival time is $\mu(\mathbf{x}) = \eta^{-1}(\mathbf{Z}^\top \boldsymbol{\alpha})$ when $\tau \rightarrow \infty$. To derive a consistent estimator of $\boldsymbol{\alpha}$, we construct another estimating equation based on the following expectation:

$$E\left[\frac{\delta}{w(Y)} \middle| X = \mathbf{x}\right] = \int_0^\infty \frac{f_U(y|\mathbf{x})}{w(y)\mu(\mathbf{x})} \int_0^y S_C(v) dv dy = \frac{1}{\mu(\mathbf{x})}. \quad (10)$$

After plugging the estimated weight functions $\hat{w}(t)$ and $\hat{w}_\tau(t)$ into equation (9) and the estimating equation constructed based on equation (10), we have

$$\mathbf{U}_1(\boldsymbol{\alpha}) = \sum_{i=1}^n z_i \left[\frac{\delta_i}{\hat{w}(y_i)} - \frac{1}{\eta^{-1}(\mathbf{z}_i^\top \boldsymbol{\alpha})} \right] = 0 \quad (11)$$

$$\begin{aligned} \mathbf{U}_2(\boldsymbol{\theta}) &= \sum_{i=1}^n z_i \left\{ \frac{y_i \delta_i I(y_i \leq \tau)}{\hat{w}(y_i)} + \frac{\tau \delta_i I(a_i < \tau < y_i)}{\hat{w}_\tau(y_i)} + \tau I(a_i \geq \tau) - \frac{(1-\tau)}{\eta^{-1}(\mathbf{z}_i^\top \boldsymbol{\alpha})} \eta^{-1}(\mathbf{z}_i^\top \boldsymbol{\beta}_\tau) - \tau \right\} \\ &= 0, \end{aligned} \quad (12)$$

where $\boldsymbol{\theta} = (\boldsymbol{\alpha}^\top, \boldsymbol{\beta}_\tau^\top)^\top$. Equations (11) and (12) can be solved jointly to obtain $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\alpha}}^\top, \hat{\boldsymbol{\beta}}_\tau^\top)^\top$.

In the estimation procedure, we assume that censoring is independent of the covariates. We may relax this assumption by replacing $S_C(\cdot)$ with $S_C(\cdot | \mathbf{x})$, and $w(t)$ and $w_\tau(t)$ with $w(t|\mathbf{x}) = \int_0^t S_C(v|\mathbf{x}) dv$ and $w_\tau(t|\mathbf{x}) = \int_{t-\tau}^t S_C(v|\mathbf{x}) dv$, respectively. When the covariates are discrete, $S_C(\cdot | \mathbf{x})$ can be consistently estimated nonparametrically through the covariate-specific Kaplan–Meier estimator. Otherwise, we can adopt the local Kaplan–Meier estimator of $S_C(\cdot | \mathbf{x})$, as shown in Wang and Wang (2014). Let $\{B_j(\mathbf{x}), i = 1, \dots, n\}$ be a sequence of weights satisfying $B_j \geq 0$ and $\sum_{i=1}^n B_i = 1$. We can estimate $S_C(t|\mathbf{x})$ by

$$\hat{S}_C(t|\mathbf{x}) = \prod_{j=1}^n \left\{ 1 - \frac{B_j(\mathbf{x})}{\sum_{k=1}^n I(Y_k - A_k \geq Y_j - A_j) B_k(\mathbf{x})} \right\}^{I(Y_j - A_j \leq t, \delta_j = 0)}.$$

A common choice of $B_i(\mathbf{x})$ is the Nadaraya–Watson-type kernel estimates. As pointed out by Wang and Wang (2014), when $B_i(\mathbf{x}) = 1/n$ for $i = 1, \dots, n$, $S_C(\cdot | \mathbf{x})$ reduces to $S_C(\cdot)$. Alternatively, we may postulate a semiparametric regression model for $S_C(\cdot | \mathbf{x})$.

Regarding the efficiency of the proposed estimating equations, we derived the semiparametrically efficient estimating equations among a class of linear combinations of the original set of estimating functions; see Web Appendix C. We note that the derived efficient estimating equations rely on the form of the original estimating equations and may not reach the semiparametric efficiency bound in general.

4.3 Asymptotic Properties

The consistency and weak convergence of $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}_\tau$ can be established under the regularity conditions listed in Web Appendix B.1. The set of equations $\{\mathbf{U}_1(\boldsymbol{\alpha})^\top, \mathbf{U}_2(\boldsymbol{\theta})^\top\}^\top = \mathbf{0}$ have a unique solution $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\alpha}}^\top, \hat{\boldsymbol{\beta}}_\tau^\top)^\top$ (for proof, see Web Appendix B.2). In Web Appendix B.3, we show that the unique solution $\hat{\boldsymbol{\theta}}$ converges to a constant in probability, even under a misspecified model assumption. We denote the true values as $\boldsymbol{\theta}^0 = (\boldsymbol{\alpha}^{0\top}, \boldsymbol{\beta}_\tau^{0\top})^\top$. The weak convergence of the estimator for $\boldsymbol{\theta}$ is summarized in the following theorem.

Theorem 2—Under regularity conditions (B₁) – (B₄) in Web Appendix B.1, $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0)$ converges weakly to a normal distribution with mean zero and variance-covariance matrix $\boldsymbol{\Gamma}^{-1}\boldsymbol{\Sigma}\{\boldsymbol{\Gamma}^{-1}\}^\top$, where $\boldsymbol{\Gamma}$ and $\boldsymbol{\Sigma}$ are defined in Web Appendix B.4.

Define

$$\begin{aligned}\hat{\phi}_i &= z_i \left\{ \frac{\delta_i}{\hat{w}(y_i)} - \frac{1}{\eta^{-1}(z_i^\top \hat{\boldsymbol{\alpha}})} \right\} + \int_0^{\tau_C} \frac{\hat{G}(t) d\hat{M}_{C_i}(t)}{\hat{\pi}(t)} \\ \hat{\psi}_i &= z_i \left\{ \frac{y_i \delta_i I(y_i \leq \tau)}{\hat{w}(y_i)} + \frac{\tau \delta_i I(a_i < \tau < y_i)}{\hat{w}_\tau(y_i)} + \tau I(a_i \geq \tau) - \frac{(1-\tau)}{\eta^{-1}(z_i^\top \hat{\boldsymbol{\alpha}})} \eta^{-1}(z_i^\top \hat{\boldsymbol{\beta}}_\tau) - \tau \right\} \\ &\quad + \int_0^{\tau_C} \frac{[\hat{H}(t) + \hat{H}_\tau(t)] d\hat{M}_{C_i}(t)}{\hat{\pi}(t)},\end{aligned}$$

where

$$\hat{G}(t) = n^{-1} \sum_{i=1}^n \frac{z_i \delta_i I(y_i \geq t)}{\hat{w}^2(y_i)} \int_t^{y_i} \hat{S}_C(v) dv$$

$$\hat{H}(t) = n^{-1} \sum_{i=1}^n \frac{z_i y_i \delta_i I(y_i \leq \tau) I(y_i \geq t)}{\hat{w}_{\tau}^2(y_i)} \int_t^{y_i} \hat{S}_C(v) dv$$

$$\hat{H}_{\tau}(t) = n^{-1} \sum_{i=1}^n \frac{z_i \tau \delta_i I(a_i < \tau < y_i)}{\hat{w}_{\tau}^2(y_i)} \left\{ I(y_i \geq t) \int_t^{y_i} \hat{S}_C(v) dv - I(y_i - \tau \geq t) \int_t^{y_i - \tau} \hat{S}_C(v) dv \right\}$$

$$\hat{M}_{Ci}(t) = I(Y_i - A_i \leq t, \delta_i = 0) - \int_0^t I(Y_i - A_i \geq u) d\hat{\Lambda}_C(u)$$

$$\hat{\pi}(t) = \sum_{i=1}^n I(Y_i - A_i \geq t) / n,$$

in which $\tau_C = \sup\{t: S_C(t) > 0\}$ and $\hat{\Lambda}_C(u)$ is the Nelson–Aalen estimator for the cumulative hazard function of the censoring times. Let $\mathbf{a}^{\otimes 2} = \mathbf{a}\mathbf{a}^{\top}$. Then, $\mathbf{\Gamma}$ and $\mathbf{\Sigma}$ can be consistently

estimated by $\hat{\mathbf{\Gamma}} = -n^{-1} [\{ \mathbf{U}_1(\hat{\mathbf{a}}) / \mathbf{\theta} \}^{\top}, \{ \mathbf{U}_2(\hat{\mathbf{\theta}}) / \mathbf{\theta} \}^{\top}]^{\top}$ and $\hat{\mathbf{\Sigma}} = \begin{pmatrix} \hat{\Sigma}_{11} & \hat{\Sigma}_{12} \\ \hat{\Sigma}_{12} & \hat{\Sigma}_{22} \end{pmatrix}$,

respectively, where

$$\partial \mathbf{U}_1(\hat{\mathbf{a}}) / \partial \mathbf{\theta} = \left(- \sum_{i=1}^n z_i \frac{\partial}{\partial \mathbf{\alpha}} \frac{1}{\eta^{-1}(\mathbf{z}_i^{\top} \mathbf{\alpha})} \bigg|_{\mathbf{\alpha} = \hat{\mathbf{a}}}, \mathbf{0}_{(p+1) \times (p+1)} \right)$$

$$\partial \mathbf{U}_2(\hat{\mathbf{\theta}}) / \partial \mathbf{\theta} = \left(- \sum_{i=1}^n z_i \frac{\partial}{\partial \mathbf{\alpha}} \frac{(1-\tau)}{\eta^{-1}(\mathbf{z}_i^{\top} \mathbf{\alpha})} \eta^{-1}(\mathbf{z}_i^{\top} \hat{\mathbf{\beta}}_{\tau}) \bigg|_{\mathbf{\alpha} = \hat{\mathbf{a}}}, - \sum_{i=1}^n z_i \frac{(1-\tau)}{\eta^{-1}(\mathbf{z}_i^{\top} \hat{\mathbf{a}})} \frac{\partial \eta^{-1}(\mathbf{z}_i^{\top} \hat{\mathbf{\beta}}_{\tau})}{\partial \mathbf{\beta}_{\tau}} \bigg|_{\mathbf{\beta}_{\tau} = \hat{\mathbf{\beta}}_{\tau}} \right),$$

and $\hat{\Sigma}_{11} = \frac{1}{n} \sum_{i=1}^n \hat{\phi}_i^{\otimes 2}$, $\hat{\Sigma}_{22} = \frac{1}{n} \sum_{i=1}^n \hat{\psi}_i^{\otimes 2}$, and $\hat{\Sigma}_{12} = \frac{1}{n} \sum_{i=1}^n \hat{\phi}_i \hat{\psi}_i^{\top}$. A detailed proof for Theorem 2 can be found in Web Appendix B.4.

5. Simulation Studies

5.1 Nonparametric RMST Estimation

We conducted a series of simulation studies to evaluate the performance of the proposed estimation procedures under various settings. For each setting, 1000 replications were carried out with a sample size of $n = 200$ or 400 . We generate unbiased failure times $\tilde{T}$ from a Weibull distribution with the shape parameter equal to 2 and the scale equal to $\exp(1.5)$.

The residual censoring times are randomly generated from a uniform distribution to allow for various levels of censoring (i.e., censoring rates of 15%, 30%, and 45%, respectively). The RMST is estimated for specific time points $\tau = 3, 4$, and 5. Since the asymptotic variance requires solving intractable integral equations, the variance estimation is approximated by bootstrap resampling, as commonly adopted in the literature (Asgharian, M'Lan, and Wolfson, 2002; Huang and Qin, 2011). Table 1 summarizes the results. Across all settings, the point estimates appear to be unbiased. The empirical standard deviations are somewhat larger than the averaged bootstrap standard errors for sample size of 200. However, the underestimation of the standard error disappears with increased sample size of 400, and the corresponding coverage probabilities are reasonably close to the 95% confidence level. Note that the nonparametric estimation of the RMST performs consistently well over various levels of censoring.

We conducted another set of simulations to compare the efficiencies of the proposed nonparametric method based on Vardi's estimator with those of the method that uses the conditional product-limit estimator for general left-truncated data, with sample size $n = 400$. The results are summarized in Table S1 in Web Appendix E. The method based on the conditional estimator presents unbiased point estimates, but is less efficient than the proposed estimator, as expected.

5.2 Estimation of Coefficients in Semiparametric Regression Model on RMST

To evaluate the finite sample performance of the semiparametric regression method, we assume that the target population follows Model (1) with $\eta(\cdot) = \log(\cdot)$. To mimic real data, we consider sample sizes of $n = 400$ and 600. We randomly assign each patient to three groups, A, B, and C, with unequal proportions of 50%, 30%, and 20%, and set the two covariates X_1 and X_2 as dummy variables indicating groups A and B, respectively. Group C is treated as the reference. Let $\mathbf{Z} = (1, X_1, X_2)^\top$. Then, our assumed model is $\mu_\tau(\mathbf{x}) = E[\tilde{T}_\tau | \mathbf{X} = \mathbf{x}] = \exp(\beta_{\tau 0} + \beta_{\tau 1}x_1 + \beta_{\tau 2}x_2)$. The failure times are generated from a distribution with hazard function $h(t | \mathbf{x}) = \exp\{-2(\mathbf{z}^\top \boldsymbol{\gamma}) + \mathbf{z}^\top \boldsymbol{\zeta} \log(2t)\}$, where $\boldsymbol{\gamma} = (0.5, 1, 1)^\top$ and $\boldsymbol{\zeta} = (1, -0.3, 0)^\top$. It is obvious that the proportional hazards assumption does not hold since the hazard ratio of group A versus C varies by time. While group A has a survival rate that is similar to that of group B at an earlier time, it deviates from B afterward. The simulation setting is illustrated in Figure S1 in Web Appendix F. In Table 2, we summarize the simulation results. The point estimates are virtually unbiased, and the empirical standard deviations are close to the asymptotic standard errors in all scenarios. The empirical standard deviations and standard errors tend to increase slightly as the censoring rate increases. While the coverage probabilities are close to the 95% confidence level overall, slight undercoverage can be observed in the estimation of the intercept parameter, $\beta_{\tau 0}$, when the level of censoring is light to moderate. However, the estimation of the covariate effects, which is the main interest, is reliable.

To further assess the performance of the proposed regression method when the stationarity assumption is violated, additional simulation studies were conducted. We generated truncation variable $\tilde{A}$ from a truncated exponential distribution with parameter $\lambda = 0.01, 0.1$, and 0.5 while maintaining all other settings the same. As shown in Figure S2 in Web

Appendix F, λ indicates the extent of departure from uniformity of the truncation time. The simulation results are summarized in Table S2 in Web Appendix E. The bias in point estimates increases as the true truncation variable departs more from a uniform distribution, whereas the empirical standard error estimates are reasonably close to the asymptotic standard errors across all scenarios. In particular, when τ is small, the point estimates for $\beta_{\tau 1}$ and $\beta_{\tau 2}$ are more biased compared to $\beta_{\tau 0}$. Significant undercoverage is observed for larger λ when τ is small. However, the proposed semiparametric method is fairly robust to violation of the stationarity assumption when τ is large.

In addition, we investigated the potential efficiency gain from using the proposed regression method compared to the general left-truncation method derived in Web Appendix D. The results of adopting the general left-truncation method can be found in Table S3 in Web Appendix E. As expected, the estimator from the proposed regression method is 1.50 – 1.75 times more efficient than that from the general left-truncation method when the truncation time follows a uniform distribution. If the stationarity assumption is violated, the estimation method for general left-truncated data may outperform the proposed regression method since it is robust to the stationarity model assumption.

6. A Real Data Example

We apply the estimation procedures to the Canadian dementia data. The data set was collected from the Canadian Study of Health and Aging, a large prevalent cohort study with time to death outcomes for subjects who were identified as having dementia. Over 14,000 subjects, who were 65 years old or older, were randomly chosen for a health survey throughout Canada. Among the subjects who agreed to participate (10,263 subjects), 1,132 were diagnosed with dementia (Wolfson et al., 2001). Between 1991 and 1996, the individuals with dementia were followed prospectively until death or end of the study. The dementia diagnosis dates were confirmed retrospectively through medical records. The data included three subtypes of dementia, which were probable Alzheimer's disease, possible Alzheimer's disease, and vascular dementia, as well as the date of disease diagnosis, the enrollment date, and the date of the failure event or censoring. After excluding patients with missing entries, 818 patients remained in the data set. For this data set, it was verified that the incidence of dementia did not change over time (Addona and Wolfson, 2006). The stationarity assumption holds for the data, which consists of length-biased samples from the target population. Patients who had shorter survival times due to worse prognosis were unlikely to be sampled to the cohort.

In the data set, 638 (78%) among 818 patients experienced failure events and 180 (22%) were censored by the end of the study. Among all patients, 393 (48%) were diagnosed with probable Alzheimer's disease, 252 (31%) with possible Alzheimer's disease, and 173 (21%) with vascular dementia. We use the group indicators as the covariates in our analysis. The group of patients with vascular dementia is used as the baseline. The nonparametric method is applied to estimate the RMST for the three subtypes of dementia. Figure 2 shows the survival function estimated by Vardi's method and the nonparametric RMST estimates over varying values of τ for the three groups. While not much difference is detected overall in the survival curve, the RMST curves for possible and probable Alzheimer's disease begin to

deviate from the RMST curve for vascular dementia after approximately $\tau = 5$ years. The estimated RMST for patients with possible Alzheimer's disease is 3.67 (standard error=0.16, computed by the bootstrap method with 200 resamples); for those with probable Alzheimer's disease, it is 3.58 (0.18); and for those with vascular dementia, it is 3.31 (0.17) at $\tau = 5$. Obviously, the three curves remain apart after $\tau = 10$ years. When $\tau = 15$, the RMST estimated for individuals with possible Alzheimer's disease becomes 0.95 years longer than that for individuals with vascular dementia (possible Alzheimer's disease: 5.19 (0.29); vascular dementia: 4.24 (0.28)). The results of the regression analysis with $\eta(\cdot) = \log(\cdot)$ and the identity link are shown in Table 3; these results are consistent with our findings from nonparametric estimation. Based on the nonparametric estimation and the regression analysis, we observe that no significant survival difference is detected between different subtypes during the initial 5 years post diagnosis of dementia. However, when patients are followed for a longer period of time, those with possible Alzheimer's disease tend to live longer than patients with vascular dementia (using the semiparametric method with the identity link, the estimated RMST difference is 0.80 years when $\tau = 20$).

7. Remarks

The semiparametric regression method considered in this paper does not postulate any restrictive structural assumption such as the proportional hazards assumption other than the linearity between the covariates and the transformed RMST via the link function. The model allows for flexibility through various choices of link functions in a fashion similar to that of the generalized linear models, and provides a robust way to model the association between the covariates and the RMST. The link function can be selected according to the scientific question of interest. For instance, one may choose the log link or the identity link to estimate the RMST ratio or difference, respectively. Note that the linearity assumption in our model can be relaxed by using sufficiently flexible functions, such as polynomial functions or regression splines.

When the study design does not satisfy the stationarity assumption, we should consider the general setting of left-truncation. The nonparametric estimation method can be generalized as described in Section 3. The generalization of the semiparametric regression method, however, can be more challenging. The estimation of the weight functions adjusted for left-truncated data would involve estimating the survival function of the failure time and the distribution of the truncation variable as well as the survival function of the residual censoring time. We provide the main steps of how our semiparametric regression method can be extended for general left-truncated data in Web Appendix D. Further development of this approach is certainly warranted.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The work was partially supported by the U.S. National Institutes of Health Grants CA193878 and CA016672. The authors thank Professor Asgharian and the investigators from the Canadian Study of Health and Aging for

generously sharing the dementia data. The data reported in this article were collected as part of the Canadian Study of Health and Aging. The core study was funded by the Seniors Independence Research Program, through the National Health Research and Development Program (NHRDP) of Health Canada Project 6606-3954-MC(S). Additional funding was provided by Pfizer Canada Incorporated through the Medical Research Council/Pharmaceutical Manufacturers Association of Canada Health Activity Program, NHRDP Project 6603-1417-302(R), Bayer Incorporated, and the British Columbia Health Research Foundation Projects 38 (93-2) and 34 (96-1). The study was coordinated through the University of Ottawa and the Division of Aging and Seniors, Health Canada. The authors also acknowledge the Texas Advanced Computing Center (TACC) at The University of Texas at Austin for providing HPC resources that have contributed to the research results reported within this paper.

References

- Addona V, Wolfson DB. A formal test for the stationarity of the incidence rate using data from a prevalent cohort study with follow-up. *Lifetime Data Analysis*. 2006; 12:267–284. [PubMed: 16917734]
- Andersen PK, Hansen MG, Klein JP. Regression analysis of restricted mean survival time based on pseudo-observations. *Lifetime Data Analysis*. 2004; 10:335–350. [PubMed: 15690989]
- Asgharian M, M'LAN CE, Wolfson DB. Length-biased sampling with right censoring: an unconditional approach. *Journal of American Statistical Association*. 2002; 97:201–209.
- Asgharian M, Wolfson DB. Asymptotic behavior of the unconditional NPMLE of the length-biased survivor function from right censored prevalent cohort data. *The Annals of Statistics*. 2005; 33:2109–2131.
- Chen PY, Tsiatis AA. Causal inference on the difference of the restricted mean lifetime between two groups. *Biometrics*. 2001; 57:1030–1038. [PubMed: 11764241]
- Huang C-Y, Qin J. Nonparametric estimation for length-biased and right-censored data. *Biometrika*. 2011; 98:177–186. [PubMed: 23049126]
- Irwin JO. The standard error of an estimate of expectation of life, with special reference to expectation of tumourless life in experiments with mice. *The Journal of Hygiene*. 1949; 47:188–189. [PubMed: 15406758]
- Karrison TG. Use of Irwin's restricted mean as an index for comparing survival in different treatment groups--interpretation and power considerations. *Controlled Clinical Trials*. 1997; 18:151–167. [PubMed: 9129859]
- Mandel M. Nonparametric estimation of a distribution function under biased sampling and censoring. *The Institute of Mathematical Statistics Lecture Notes--Monograph Series*. 2007; 54:224–238.
- Qin J, Shen Y. Statistical methods for analyzing right-censored length-biased data under Cox model. *Biometrics*. 2010; 66:382–392. [PubMed: 19522872]
- Qin J, Ning J, Liu H, Shen Y. Maximum likelihood estimations and EM algorithms with length-biased data. *Journal of the American Statistical Association*. 2011; 106:1434–1449. [PubMed: 22323840]
- Royston P, Parmar MKB. The use of restricted mean survival time to estimate the treatment effect in randomized clinical trials when the proportional hazards assumption is in doubt. *Statistics in Medicine*. 2011; 30:2409–2421. [PubMed: 21611958]
- Royston P, Parmar MKB. Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. *BMC Medical Research Methodology*. 2013; 13:152. [PubMed: 24314264]
- Rudser KD, LeBlanc ML, Emerson SS. Distribution-free inference on contrasts of arbitrary summary measures of survival. *Statistics in Medicine*. 2012; 31:1722–1737. [PubMed: 22362470]
- Tian L, Zhao L, Wei LJ. Predicting the restricted mean event time with the subject's baseline covariates in survival analysis. *Biostatistics*. 2014; 15:222–233. [PubMed: 24292992]
- Trinquart L, Jacot J, Conner SC, Porcher R. Comparison of treatment effects measured by the hazard ratio and by the ratio of restricted mean survival times in oncology randomized controlled trials. *Journal of Clinical Oncology*. 2016; 34:1813–1819. [PubMed: 26884584]
- Tsai WY. Pseudo-partial likelihood for proportional hazards models with biased-sampling data. *Biometrika*. 2009; 96:601–615. [PubMed: 22422175]
- Tsai W-Y, Jewell NP, Wang M-C. A note on the product-limit estimator under right censoring and left truncation. *Biometrika*. 1987; 74:883–886.

- Uno H, Claggett B, Tian L, Inoue E, Gallo P, Miyata T, et al. Moving beyond the hazard ratio in quantifying the between-group difference in survival analysis. *Journal of Clinical Oncology*. 2014; 32:2380–2385. [PubMed: 24982461]
- Vardi Y. Multiplicative censoring, renewal processes, deconvolution and decreasing density: nonparametric estimation. *Biometrika*. 1989; 76:751–761.
- Wang HJ, Wang L. Quantile regression analysis of length-biased survival data. *Stat*. 2014; 3:31–47.
- Wang M-C. Nonparametric estimation from cross-sectional survival data. *Journal of the American Statistical Association*. 1991; 84:742–748.
- Wang M-C. Hazards regression analysis for length-biased data. *Biometrika*. 1996; 83:343–354.
- Wolfson C, Wolfson DB, Asgharian M, M'Lan CE, Ostbye T, Rockwood K, et al. A reevaluation of the duration of survival after the onset of dementia. *The New England Journal of Medicine*. 2001; 344:1111–1116. [PubMed: 11297701]
- Zhang M, Schaubel DE. Estimating differences in restricted mean lifetime using observational data subject to dependent censoring. *Biometrics*. 2011; 67:740–749. [PubMed: 21039400]
- Zhao L, Claggett B, Tian L, Uno H, Pfeffer MA, Solomon SD, et al. On the restricted mean survival time curve in survival analysis. *Biometrics*. 2016; 72:215–221. [PubMed: 26302239]
- Zhao L, Tian L, Uno H, Solomon SD, Pfeffer MA, Schindler JS, et al. Utilizing the integrated difference of two survival functions to quantify the treatment contrast for designing, monitoring, and analyzing a comparative clinical study. *Clinical Trials*. 2012; 9:570–577. [PubMed: 22914867]
- Zucker DM. Restricted mean life with covariates: modification and extension of a useful survival analysis method. *Journal of the American Statistical Association*. 1998; 93:702–709.

Case

(i) Patient 1

(ii) Patient 2

(iii) Patient 3

Enrollment

Figure 1.

An illustration of the three exclusive possibilities, (i) $Y < \tau$, (ii) $A < \tau < Y$, and (iii) $A > \tau$.

Figure 2.

Vardi's estimator of $S_{\mathcal{L}}(t)$ (left panel) and the nonparametric RMST estimator (right panel) for each subtype of dementia. The solid line represents possible Alzheimer's disease; the dashed line, probable Alzheimer's disease; the dotted line, vascular dementia.

Table 1

Summary statistics of simulations for nonparametric estimation of the RMST. True RMST (True), Monte Carlo mean of the estimates (Est.), the empirical standard deviation (SD), average of bootstrap standard errors (SE), and coverage probability (CP) for $\tau = 3, 4, 5$ with sample sizes of $n = 200, 400$ and censoring rates 15%, 30%, and 45%.

Censoring Rate													
15%							30%						
n	τ	True	Est.	SD	SE	CP	Est.	SD	SE	CP	Est.	SD	SE
200	3	2.606	2.604	0.125	0.094	0.887	2.602	0.139	0.096	0.881	2.601	0.149	0.099
	4	3.150	3.147	0.164	0.131	0.903	3.143	0.181	0.134	0.908	3.143	0.192	0.138
	5	3.517	3.512	0.192	0.159	0.917	3.508	0.210	0.163	0.917	3.507	0.222	0.169
400	3	2.606	2.606	0.072	0.068	0.912	2.606	0.073	0.069	0.915	2.605	0.074	0.071
	4	3.150	3.150	0.098	0.094	0.925	3.150	0.099	0.095	0.930	3.149	0.101	0.099
	5	3.517	3.517	0.119	0.114	0.932	3.517	0.120	0.116	0.930	3.516	0.122	0.120

Table 2

Summary statistics of simulations for estimation of the regression coefficients. True $\boldsymbol{\beta}_\tau = (\beta_{\tau 0}, \beta_{\tau 1}, \beta_{\tau 2})$ and $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \alpha_2)$ (True), Monte Carlo mean of the estimates (Est.), the empirical standard deviation (SD), average of the standard errors (SE), and coverage probability (CP) for $\tau = 3, 5, 7$ with sample sizes of $n = 400, 600$ and censoring rates 15%, 30%, and 45%.

Censoring Rate														
15%														
τ	True	Est.	SD	SE	CP	30%			45%					
		Est.	SD	SE	CP	Est.	SD	SE	Est.	SD	SE	CP		
$n = 400$														
3	β_0	0.369	0.380	0.144	0.124	0.925	0.381	0.151	0.137	0.935	0.379	0.166	0.155	0.942
	β_1	0.620	0.609	0.169	0.148	0.936	0.630	0.184	0.180	0.962	0.684	0.242	0.250	0.930
	β_2	0.589	0.579	0.169	0.157	0.941	0.577	0.188	0.183	0.951	0.591	0.234	0.240	0.964
5	β_0	0.379	0.391	0.142	0.121	0.917	0.392	0.145	0.130	0.935	0.390	0.153	0.142	0.942
	β_1	0.982	0.971	0.157	0.137	0.938	0.978	0.165	0.154	0.952	0.994	0.190	0.184	0.952
	β_2	0.878	0.867	0.161	0.144	0.933	0.866	0.171	0.160	0.940	0.873	0.190	0.186	0.955
7	β_0	0.379	0.392	0.141	0.120	0.917	0.393	0.143	0.128	0.928	0.391	0.152	0.139	0.940
	β_1	1.156	1.145	0.155	0.135	0.940	1.147	0.160	0.148	0.948	1.155	0.179	0.169	0.959
	β_2	0.972	0.961	0.160	0.142	0.933	0.960	0.167	0.155	0.937	0.965	0.184	0.175	0.955
	α_0	0.379	0.390	0.133	0.117	0.910	0.391	0.135	0.123	0.921	0.391	0.144	0.132	0.937
	α_1	1.298	1.288	0.147	0.132	0.936	1.288	0.150	0.141	0.948	1.290	0.162	0.154	0.954
	α_2	1.000	0.990	0.151	0.138	0.939	0.989	0.156	0.148	0.937	0.992	0.170	0.163	0.953
$n = 600$														
3	β_0	0.369	0.375	0.122	0.107	0.931	0.380	0.133	0.118	0.936	0.380	0.147	0.132	0.950
	β_1	0.620	0.614	0.138	0.126	0.946	0.628	0.157	0.153	0.957	0.678	0.210	0.203	0.915
	β_2	0.589	0.585	0.149	0.133	0.944	0.580	0.166	0.154	0.953	0.595	0.201	0.194	0.943
5	β_0	0.379	0.386	0.120	0.104	0.923	0.390	0.128	0.112	0.935	0.387	0.136	0.124	0.951
	β_1	0.982	0.975	0.131	0.117	0.932	0.976	0.142	0.131	0.956	0.992	0.165	0.158	0.948
	β_2	0.878	0.874	0.140	0.123	0.941	0.870	0.152	0.136	0.938	0.877	0.168	0.158	0.951
7	β_0	0.379	0.386	0.119	0.103	0.919	0.390	0.126	0.110	0.931	0.391	0.136	0.119	0.945

Censoring Rate												
τ	True	15%			30%			45%			SE	CP
		Est.	SD	SE	Est.	SD	SE	Est.	SD	SE		
β_{τ_1}	1.156	1.149	0.130	0.115	0.937	0.139	0.126	1.151	0.157	0.143	0.143	0.948
β_{τ_2}	0.972	0.968	0.137	0.121	0.940	0.148	0.132	0.965	0.162	0.148	0.148	0.946
α_0	0.379	0.385	0.114	0.102	0.917	0.116	0.107	0.388	0.124	0.113	0.113	0.939
α_1	1.298	1.293	0.124	0.113	0.940	0.127	0.120	1.291	0.139	0.131	0.131	0.951
α_2	1.000	0.996	0.132	0.119	0.939	0.134	0.126	0.995	0.147	0.138	0.138	0.942

Table 3
The regression coefficient estimates (SE) for the values of τ , with the log and identity link functions assumed.

τ (years)	Log Link			Identity Link		
	Intercept	Probable Alzheimer's Disease	Possible Alzheimer's Disease	Intercept	Probable Alzheimer's Disease	Possible Alzheimer's Disease
3	0.93 (0.10)	-0.01 (0.12)	-0.17 (0.21)	2.55 (0.24)	-0.02 (0.32)	-0.41 (0.48)
5	1.19 (0.07)	0.06 (0.10)	0.01 (0.10)	3.30 (0.24)	0.19 (0.33)	0.02 (0.35)
7	1.31 (0.07)	0.08 (0.09)	0.06 (0.10)	3.70 (0.25)	0.31 (0.35)	0.21 (0.37)
11	1.39 (0.06)	0.12 (0.09)	0.12 (0.09)	4.04 (0.26)	0.49 (0.37)	0.50 (0.39)
15	1.41 (0.06)	0.12 (0.09)	0.16 (0.09)	4.11 (0.26)	0.50 (0.38)	0.72 (0.40)
20	1.42 (0.06)	0.12 (0.09)	0.18 (0.09)	4.16 (0.27)	0.50 (0.38)	0.80 (0.40)
Overall	1.44 (0.06)	0.11 (0.08)	0.17 (0.09)	4.24 (0.27)	0.48 (0.38)	0.79 (0.39)