HHS Public Access

Author manuscript

Lifetime Data Anal. Author manuscript; available in PMC 2020 January 01.

Published in final edited form as:

Lifetime Data Anal. 2019 January ; 25(1): 79–96. doi:10.1007/s10985-018-9422-y.

Model diagnostics for the proportional hazards model with length-biased data

Chi Hyun Lee, **Jing Ning**, and **Yu Shen**

Department of Biostatistics, The University of Texas MD Anderson Cancer Center, 1400 Pressler Street Unit 1411, Houston, TX 77030, USA

Abstract

Length-biased data are frequently encountered in prevalent cohort studies. Many statistical methods have been developed to estimate the covariate effects on the survival outcomes arising from such data while properly adjusting for length-biased sampling. Among them, regression methods based on the proportional hazards model have been widely adopted. However, little work has focused on checking the proportional hazards model assumptions with length-biased data, which is essential to ensure the validity of inference. In this article, we propose a statistical tool for testing the assumed functional form of covariates and the proportional hazards assumption graphically and analytically under the setting of length-biased sampling, through a general class of multiparameter stochastic processes. The finite sample performance is examined through simulation studies, and the proposed methods are illustrated with the data from a cohort study of dementia in Canada.

Keywords

Dementia; Length-biased data; Model diagnostics; Proportional hazards model; Stochastic processes

1 Introduction

In prevalent cohort studies, a group of patients who have experienced an initial event (i.e., disease onset) at the time of recruitment are followed prospectively until an event of interest (i.e., failure) or censoring occurs. Under the stationarity assumption, in which the initiating events are assumed to follow a stationary Poisson process, the right-censored time-to-event data are subject to length-biased sampling and are referred to as "length-biased data." These data arise in various applications such as cancer screening trials (Zelen and Feinleib, 1969) and studies of unemployment (Lancaster, 1979; de Una-Alvarez et al, 2003). In such data, subjects with longer failure times are more likely to be sampled. Thus, the observed data may not be representative of the target population, which poses additional challenges in the analysis.

Correspondence to: Chi Hyun Lee.

A number of studies have been conducted to estimate the association between covariates and the survival outcome under the setting of length-biased sampling. Specifically, methods based on the proportional hazards model (Cox, 1972) have been extensively studied. Wang (1996) constructed the pseudo-likelihood by using a bias-adjusted risk set to address length-biased failure time data without right censoring. Tsai (2009) generalized the pseudo-partial likelihood approach for biased sampling data under a different right-censoring schema. Qin and Shen (2010) proposed two inversely weighted estimating equations for observed length-biased data subject to right-censoring. Their methods can be easily implemented with readily available software, which is especially appealing. Qin et al (2011) and Huang and Qin (2012) proposed more efficient full likelihood and composite partial likelihood approaches, respectively, to estimate the covariate effects on the underlying survival outcome.

Under the proportional hazards model, the functional form of covariates and the proportional hazards assumption are fundamental components that need to be adequately verified for valid inferential results. Among many model checking tools for survival data, the method proposed by Lin et al (1993), for which the building block is the cumulative sum of martingale residuals, has been frequently adapted in the literature for various types of complex survival data. Spiekerman and Lin (1996) extended the method to handle correlated failure time in clustered data. Analogously, Huang et al (2011) developed the model checking technique for recurrent gap time data, where events may be observed multiple times for each patient, using the averaged martingale-like processes. More recently, a model checking tool has been developed for survival data observed in nested case-control studies (Borgan and Zhang, 2015; Lu et al, 2014). However, there is no such model checking tool available for length-biased data. We note that the model checking technique established by Lin et al (1993) cannot be directly applied because the martingale residual processes are unavailable for observed length-biased data. Extensions of that technique also cannot be utilized because of the unique structure of length-biased data.

We face some complications when checking the proportional hazards model assumptions with length-biased right-censored data because of the sampling mechanism. First of all, the proportional hazards model assumed for the target population may not fit the observed biased data. Thus, checking the proportional hazards model under length-biased sampling requires proper adjustment. In addition, the failure time can depend on the censoring time or the duration from the initiating event to censoring. In this paper, we propose a statistical tool for checking the proportional hazards model with length-biased data. In Section 2, we introduce the data structure and the inferential procedure based on the proportional hazards model. We outline the diagnostic method in Section 3, derive the asymptotic distribution under the null hypothesis, and establish a computationally efficient resampling method. We demonstrate the performance of the proposed model checking tools through simulations under various settings in Section 4. In Section 5, we apply the method to a real data set collected from a large prevalent cohort study on dementia in Canada. In Section 6, we conclude with some remarks on the proposed method.

2 Length-Biased Data and Statistical Inference

2.1 Notation and Model

Let $\tilde{T}$ be the duration from an initiating event to failure; $\tilde{A}$ be the duration from the initiating event to enrollment into the study; and V be the duration from enrollment to failure. We let a $p\times 1$ vector $\boldsymbol{Z}$ denote the baseline covariates. Assume that the failure time $\tilde{T}$ follows the proportional hazards model

$$\lambda(t \mid z) = \lambda_0(t) \exp\ (\boldsymbol{\beta}_0^{\top} z) \quad (1)$$

where β_0 is a $p \times 1$ vector of unknown coefficients for $\boldsymbol{Z}$ and $\lambda_0(t)$ is an unspecified baseline hazard function. Note that we only observe failure time $T = \tilde{T}$ when $\tilde{A} < \tilde{T}$ due to length-biased sampling. Thus, the biased failure time T is $A + V$, where A is the observed backward recurrence time or truncation variable. Since V, the residual survival time or the forward recurrence time, is subject to right censoring, the observed survival time is $Y = \min(T, A + C)$ and the censoring indicator is $\delta = I(T \le A + C)$, where C is the residual censoring time. For n independent subjects, the data consist of $(Y_i, A_i, \delta_i, \boldsymbol{Z}_i)$, $i = 1, \ldots, n$. We assume that C is independent of (A, V) given $\boldsymbol{Z}$, and the distribution of C is independent of $\boldsymbol{Z}$. Conditioning on $\boldsymbol{Z}$, even under the independent censoring assumption on C, there exists dependence between the failure time and censoring time (i.e., duration from an initiating event to a censoring event) because $\mathrm{Cov}(T, A + C \mid \boldsymbol{Z}) = \mathrm{Var}(A \mid \boldsymbol{Z}) + \mathrm{Cov}(A, V \mid \boldsymbol{Z}) > 0$. Also, model (1) is postulated for the underlying (unbiased) failure time while the observed survival time is subject to length bias. Therefore, the observed data may not follow the model structure assumed for the target population.

2.2 Estimation of the Covariate Effects

Among many approaches developed under the proportional hazards model with length-biased data, we briefly review the generalized estimating equation method of Qin and Shen (2010). Following the conventional counting process notation, we define $N_i(t) = I\{Y_i \le t, \delta_i = 1\}$ and $R_i(t) = I\{Y_i \ge t, \delta_i = 1\}$ for subject i. Let the weight function $w_C(t) = \int_0^t S_C(u)\mathrm{d}u$, where $S_C(t) = \Pr(C > t)$ is the survival function of the residual censoring variable C. Let $\mathbf{a}^0 = 1$, $\mathbf{a}^1 = \mathbf{a}$, and $\mathbf{a}^2 = \mathbf{a}\mathbf{a}^{\top}$ for any vector $\mathbf{a}$. We define

$$S^{(k)}(\boldsymbol{\beta}, t) = n^{-1} \sum_{i=1}^{n} w_C(t) R_i(t) \{w_C(Y_i)\}^{-1} \mathbf{Z}_i^k \exp\ (\boldsymbol{\beta}^{\top} \mathbf{Z}_i)$$

for $k = 0, 1,$ and 2. Let $E(\boldsymbol{\beta}, t) = S^{(1)}(\boldsymbol{\beta}, t)/S^{(0)}(\boldsymbol{\beta}, t)$, and denote its limit as $e(\boldsymbol{\beta}, t)$. Note that $w_C(t)$ can be consistently estimated by $\hat{w}_C(t) = \int_0^t \hat{S}_C(u)\mathrm{d}u$, where $\hat{S}_C(t)$ is the Kaplan–Meier estimator of the residual censoring survival function. Replacing $w_C(\cdot)$ with its consistent estimator, $\hat{w}_C(\cdot)$, we have

$$\hat{S}^{(k)}(\boldsymbol{\beta}, t) = n^{-1} \sum_{i=1}^{n} \hat{w}_C(t) R_i(t) \{\hat{w}_C(Y_i)\}^{-1} \mathbf{Z}_i^k \exp(\boldsymbol{\beta}^\top \mathbf{Z}_i)$$

for $k = 0, 1,$ and 2. Then the estimator $\hat{\boldsymbol{\beta}}$ for the true covariate effects $\boldsymbol{\beta}_0$ can be obtained by solving the following unbiased estimating equation

$$\hat{U}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \int_0^{\tau} \left\{ \mathbf{Z}_i - \hat{E}(\boldsymbol{\beta}, u) \right\} \mathrm{d}N_i(u) = 0,$$

where τ satisfies $\Pr(Y \geq \tau) > 0$ and $\hat{E}(\boldsymbol{\beta}, t) = \hat{S}_{(1)}(\boldsymbol{\beta}, t) / \hat{S}_{(0)}(\boldsymbol{\beta}, t)$.

3 Model Diagnostic Methods

By following the counting process and martingale framework, a stochastic process for length-biased data can be constructed as

$$M_i(t) = N_i(t) - \int_0^t w_C(u) R_i(u) \{w_C(Y_i)\}^{-1} \exp(\boldsymbol{\beta}_0^\top \mathbf{Z}_i) \mathrm{d}\Lambda_0(u), \quad (2)$$

for $i = 1, \ldots, n$, where $\Lambda_0(t) = \int_0^t \lambda_0(s) \mathrm{d}s$ is the cumulative baseline hazard function. The stochastic process can be interpreted as the difference between the observed number of events and the expected number of events under the assumed model until time t. When model (1) is correctly specified, this becomes a mean zero stochastic process. Thus, the process is informative for detecting violations of model assumptions. The stochastic process (2) can be estimated by

$$\hat{M}_i(t) = N_i(t) - \int_0^t \hat{w}_C(u) R_i(u) \{\hat{w}_C(Y_i)\}^{-1} \exp(\hat{\boldsymbol{\beta}}^\top \mathbf{Z}_i) \mathrm{d}\hat{\Lambda}_0(\hat{\boldsymbol{\beta}}, u)$$

for $i = 1, \ldots, n$, where

$$\hat{\Lambda}_0(\hat{\boldsymbol{\beta}}, t) = \int_0^t \frac{\sum_{i=1}^n \mathrm{d}N_i(u)}{n \hat{S}^{(0)}(\hat{\boldsymbol{\beta}}, u)}.$$

The stochastic process $\hat{M}_i(t)$ performs in a fashion similar to martingale residuals. However, the process (2) does not satisfy the martingale definition due to the length-biased data structure. Therefore, $\hat{M}_i(t)$ is different from the ordinary martingale residuals. To test the model assumptions, we consider a general form of the multiparameter stochastic process using cumulative summation, in a manner similar to the approach of Lin et al (1993),

$$\boldsymbol{G}(t,z) = \sum_{i=1}^{n} f(\boldsymbol{Z}_i) I(\boldsymbol{Z}_i \le z) \widehat{M}_i(t), \quad (3)$$

where $f(\cdot)$ is a prespecified smooth and bounded function, and $I(\boldsymbol{Z}_i \le \boldsymbol{z}) = I(Z_{i1} \le z_1, \ldots, Z_{ip} \le z_p)$ with $\boldsymbol{Z}_i = (Z_{i1}, \ldots, Z_{ip})^\top$ and $\boldsymbol{z} = (z_1, \ldots, z_p)^\top$. As pointed out by Lin et al (1993), approximating the distribution of (3) is more accessible than that of an individual stochastic process, $\hat{M}_i(t)$. Furthermore, individual stochastic processes evaluated with censored survival outcomes are equal to zero, which makes it disadvantageous to detect and measure model departure with an individual stochastic process. Therefore, we establish our model diagnostic procedure on the basis of (3) instead of $\hat{M}_i(t)$. If the assumed model (1) is true, the process (3) will fluctuate randomly around zero.

The null distribution of the multiparameter stochastic process (3) under model (1) needs to be studied to construct our test procedures. By applying the Taylor series expansion and empirical process approximation techniques, we derive a stochastic process asymptotically equivalent to (3) and approximate its distribution to obtain critical values for test statistics later. First, define

$$S_Z^{(l)}(\boldsymbol{\beta}, t, z) = n^{-1} \sum_{i=1}^{n} f(\boldsymbol{Z}_i) I(\boldsymbol{Z}_i \le z) w_C(t) R_i(t) \{w_C(Y_i)\}^{-1} \boldsymbol{Z}_i^l \exp(\boldsymbol{\beta}^\top \boldsymbol{Z}_i)$$

for $l = 0, 1$. Let $E_Z(\boldsymbol{\beta}, t, z) = S_Z^{(0)}(\boldsymbol{\beta}, t, z)/S^{(0)}(\boldsymbol{\beta}, t)$, and the corresponding limit be $e_Z(\boldsymbol{\beta}, t, \boldsymbol{z})$. We denote the limit of the first derivatives of $\int_0^t E_Z(\boldsymbol{\beta}, u, z) \mathrm{d}N_i(u)$ and $\int_0^\tau \{\boldsymbol{Z}_i - E(\boldsymbol{\beta}, u)\} \mathrm{d}N_i(u)$ with respect to $\boldsymbol{\beta}$ as

$$\Gamma_Z(\boldsymbol{\beta}, t, z) = \mathrm{E}\left\{ \int_0^t \left[\frac{S_{\boldsymbol{Z}}^{(1)}(\boldsymbol{\beta}, u, z)}{S^{(0)}(\boldsymbol{\beta}, u)} - \frac{S_{\boldsymbol{Z}}^{(0)}(\boldsymbol{\beta}, u, z) S^{(1)}(\boldsymbol{\beta}, u)}{\{S^{(0)}(\boldsymbol{\beta}, u)\}^2} \right] \mathrm{d}N_i(u) \right\}$$

and

$$\Gamma(\boldsymbol{\beta}) = -\mathrm{E}\left\{ \int_0^\tau \left[\frac{S^{(2)}(\boldsymbol{\beta}, u)}{S^{(0)}(\boldsymbol{\beta}, u)} - \left\{ \frac{S^{(1)}(\boldsymbol{\beta}, u)}{S^{(0)}(\boldsymbol{\beta}, u)} \right\}^2 \right] \mathrm{d}N_i(u) \right\},$$

respectively. For $i = 1, \ldots, n$, we define

$$\boldsymbol{G}_i^*(t,z)=\int_0^t\{f(\boldsymbol{Z}_i)I(\boldsymbol{Z}_i\le z)-e_Z(\boldsymbol{\beta}_0,u)\}\,\mathrm{d}M_i(u)+\int_0^t H(\boldsymbol{\beta}_0,u)\frac{\mathrm{d}M_{C_i}(u)}{\pi(u)}+\Gamma_Z(\boldsymbol{\beta}_0,t,z \quad (4)$$

$$)\{\Gamma(\boldsymbol{\beta}_0)\}^{-1}\int_0^\infty\{\boldsymbol{Z}_i-e(\boldsymbol{\beta}_0,u)\}\,\mathrm{d}M_i(u)$$

where

$$H(\boldsymbol{\beta},t)=\lim_{n\to\infty}\sum_{i=1}^n\sum_{k=1}^n\frac{f(\boldsymbol{Z}_k)I(\boldsymbol{Z}_k\le z)w_C(Y_i)R_k(Y_i)\exp(\boldsymbol{\beta}^\top\boldsymbol{Z}_k)\{w_C(Y_k)\}^{-2}h_k(t)}{n^2S^{(0)}(\boldsymbol{\beta},Y_i)},$$

$$M_{C_i}(t)=I(V_i\le t,\delta_i=0)-\int_0^t I(V_i\ge u)\mathrm{d}\Lambda_C(u),$$

$$h_k(t)=I(Y_k\ge t)\int_t^{Y_k}S_C(u)\mathrm{d}u,$$

$$\pi(t)=S_C(t)S_V(t),$$

in which $\Lambda_C(\cdot)$ is the cumulative hazard function for the residual censoring time and $S_V(\cdot)$ is the survival function of the residual survival time. It is obvious that given the covariates, (4) is a mean zero stochastic process under model (1). We summarize the asymptotic properties of (3) in the following theorem.

Theorem 1—*Under model* (1) *and the regularity conditions, the stochastic process* (3) *can be approximated by* $n^{-1/2}\boldsymbol{G}(t,z)=n^{-1/2}\sum_{i=1}^n\boldsymbol{G}_i^*(t,z)+o_p(1)$. *The process* $n^{-1/2}\boldsymbol{G}(t,\boldsymbol{z})$ *converges weakly to a mean zero Gaussian process with covariance* $E\{\boldsymbol{G}_i^*(t_1,z_1)\boldsymbol{G}_i^*(t_2,z_2)^\top\}$ *as* $n\to\infty$.

The list of regularity conditions and the proof of Theorem 1 are provided in Appendices A and B, respectively. It is noteworthy that the stochastic process (3) can be approximately represented by the sum of independent and identically distributed (i.i.d.) mean zero processes as $\boldsymbol{G}(t,z)\approx\sum_{i=1}^n\boldsymbol{G}_i^*(t,z)$. While the asymptotic presentation may bear some similarity with that of Lin et al (1993) and its extensions, the approximation is in fact different because of the extra weight terms introduced to adjust for potential dependent censoring. Note that the representation (4) has a second term that accounts for the uncertainty induced additionally by $\hat{w}_C(\cdot)$. Since the distributional form of the stochastic process (2) is unknown and the covariance structure is complicated in the asymptotic distribution of (3), it is quite challenging to analytically evaluate the limiting distribution. As an alternative method, we approximate the asymptotic distribution through Monte Carlo simulation, which has been widely adopted in the literature (Lin et al, 1993; Spiekerman and Lin, 1996).

We can estimate the i.i.d. stochastic processes (4) for $i = 1, \ldots, n$ by replacing the components with their respective consistent estimators as follows.

$$\widehat{G}_i^*(t,z) = \int_0^t \left\{ f(\mathbf{Z}_i) I(\mathbf{Z}_i \le z) - \hat{E}_Z(\hat{\boldsymbol{\beta}}, u, z) \right\} \mathrm{d}\widehat{M}_i(u) + \int_0^t \widehat{H}(\hat{\boldsymbol{\beta}}, u) \frac{\mathrm{d}\widehat{M}_{C_i}(u)}{\hat{\pi}(u)} + \hat{\Gamma}_Z(\hat{\boldsymbol{\beta}}, t, z$$

$$)\{\hat{\Gamma}(\hat{\boldsymbol{\beta}})\}^{-1} \int_0^\tau \left\{ \mathbf{Z}_i - \hat{E}(\hat{\boldsymbol{\beta}}, u) \right\} \mathrm{d}\widehat{M}_i(u),$$

where

$$\hat{S}_Z^{(l)}(\boldsymbol{\beta}, l, z) = n^{-1} \sum_{i=1}^n f(\mathbf{Z}_i) I(\mathbf{Z}_i \le z) \hat{w}_C(t) R_i(t) \{\hat{w}_C(Y_i)\}^{-1} \mathbf{Z}_i^l \exp(\boldsymbol{\beta}^\top \mathbf{Z}_i)$$

for $l = 0$, $\hat{E}_Z(\boldsymbol{\beta}, u, z) = \hat{S}_Z^{(0)}(\boldsymbol{\beta}, t, z) / \hat{S}^{(0)}(\boldsymbol{\beta}, t)$,

$$\widehat{H}(\boldsymbol{\beta}, t) = \sum_{i=1}^n \sum_{k=1}^n \frac{f(\mathbf{Z}_k) I(\mathbf{Z}_k \le z) \hat{w}_C(Y_i) R_k(Y_i) \exp(\boldsymbol{\beta}^\top \mathbf{Z}_k) \{\hat{w}_C(Y_k)\}^{-2} \hat{h}_k(t)}{n^2 \hat{S}^{(0)}(\boldsymbol{\beta}, Y_i)},$$

$$\widehat{M}_{C_i}(t) = I(V_i \le t, \delta_i = 0) - \int_0^t I(V_i \ge u) \mathrm{d}\hat{\Lambda}_C(u),$$

$$\hat{h}_k(t) = I(Y_k \ge t) \int_t^{Y_k} \hat{S}_C(u) \mathrm{d}u,$$

$$\hat{\pi}(t) = \hat{S}_C(t) \hat{S}_V(t),$$

in which $\hat{\Lambda}_C(\cdot)$ is the Nelson-Aalen estimator for the residual censoring time and $\hat{S}_V(\cdot)$ is the Kaplan–Meier estimator of the residual survival time. Also, the consistent estimators

$$\hat{\Gamma}_Z(\boldsymbol{\beta}, t, z) = n^{-1} \sum_{i=1}^n \int_0^t \left[\frac{\hat{S}_{\mathbf{Z}}^{(1)}(\boldsymbol{\beta}, u, z)}{\hat{S}^{(0)}(\boldsymbol{\beta}, u)} - \frac{\hat{S}_{\mathbf{Z}}^{(0)}(\boldsymbol{\beta}, u, z) \hat{S}^{(1)}(\boldsymbol{\beta}, u)}{\{\hat{S}^{(0)}(\boldsymbol{\beta}, u)\}^2} \right] \mathrm{d}N_i(u)$$

and

$$\hat{\Gamma}(\boldsymbol{\beta}) = -n^{-1} \sum_{i=1}^n \int_0^\tau \left[\frac{\hat{S}^{(2)}(\boldsymbol{\beta}, u)}{\hat{S}^{(0)}(\boldsymbol{\beta}, u)} - \left\{ \frac{\hat{S}^{(1)}(\boldsymbol{\beta}, u)}{\hat{S}^{(0)}(\boldsymbol{\beta}, u)} \right\}^2 \right] \mathrm{d}N_i(u)$$

can be used in place of Γ_Z and Γ, respectively. Define $\tilde{\boldsymbol{G}}_m(t,z) = \sum_{i=1}^n \hat{\boldsymbol{G}}_i^*(t,z)V_{mi}$, where V_{mi}, $i = 1, \ldots, n$, are independent random variables sampled from a standard normal distribution for $m = 1, \ldots, M$. According to the following theorem, the limiting distribution of the multiparameter stochastic process (3) can be approximated via Monte Carlo simulation.

Theorem 2—*Under model (1) and the regularity conditions, $n^{-1/2}\tilde{\boldsymbol{G}}_m(t, \boldsymbol{z})$ converges weakly to the same Gaussian process with the asymptotic distribution of $n^{-1/2}\sum_{i=1}^n \boldsymbol{G}_i^*(t,z)$ as $n \to \infty$.*

The proof of Theorem 2 is provided in Appendix C. For a large M, the simulated realizations $\tilde{\boldsymbol{G}}_m(t, \boldsymbol{z}), m = 1, \ldots, M$, approximate the limiting distribution of (3). Given a fixed pair $(t, \boldsymbol{z})$, it follows that the estimated variance of $\tilde{\boldsymbol{G}}_m(t, \boldsymbol{z})$ is $n^{-1}\sum_{i=1}^n \{\hat{\boldsymbol{G}}_i^*(t,z)\}^2$ conditional on the observed data, which shares the same limiting variance with (3). Based on the Monte Carlo simulation, the model assumptions can be diagnosed graphically. We can compare the pattern of the observed stochastic process (3) with the simulated realizations of the limiting distribution under the assumed model by plotting a few of them. When the observed stochastic process deviates from the group of simulated processes, it possibly indicates a sign of model misspecification.

To develop formal test procedures, we construct test statistics using the supremum test, $\sup_{t,z}|\boldsymbol{G}(t, \boldsymbol{z})|$. The general form $\boldsymbol{G}(t, \boldsymbol{z})$ in (3) can be adjusted for checking two aspects of the model assumptions. When testing the functional form of the jth component of the covariates, we set $f(\cdot) = 1$, $t = \tau$, and $z_k = \infty$ for all $k \neq j$ in (3). We construct a test statistic $T_1^j = \sup_z |\boldsymbol{G}_1^j(z)|$, where $\boldsymbol{G}_1^j(z) = \sum_{i=1}^n I(Z_{ij} \leq z)\hat{M}_i(\tau)$. To test the proportional hazards assumption for the jth component of the covariates, we set $f(Z_{ij}) = Z_{ij}$ and $\boldsymbol{z} = \infty$, which is also a special case of the general class of multiparameter processes (3). The test statistic $T_2^j = \sup_t |\boldsymbol{G}_2^j(t)|$ can be considered, where $\boldsymbol{G}_2^j(t) = \sum_{i=1}^n Z_{ij}\hat{M}_i(t)$. When it is of interest to check the overall proportionality of hazards for all covariates, the global test statistic $T_2 = \sup_t \sum_{j=1}^p |\boldsymbol{G}_2^j(t)|$ can be considered. In general, an individual test for the jth covariate is expected to have greater power than the global test on the overall proportionality of hazards. We note that the function $f(\boldsymbol{Z}_i)$ is not selected for the purpose of improving efficiency of the test. The critical values for test statistics can be obtained by computing $\sup_{t,z}|\tilde{\boldsymbol{G}}_m(t, \boldsymbol{z})|$ for $m = 1, \ldots, M$. The p-values are estimated empirically by calculating the proportion of critical values greater than the proposed test statistic.

The multiparameter stochastic process can be constructed based on alternative estimation approaches with mean zero estimating functions. For the purpose of testing the proportional hazards model assumptions, we consider the simplicity of the estimating equation formulation as the most important feature. Hence, the proposed test procedures are established based on the estimating equation proposed by Qin and Shen (2010), which allows the test statistics to be implemented using existing software for conventional right-censored data.

4 Simulations

To assess the performance of the proposed model diagnostic method, we conducted simulation studies with 1000 replications. We considered various scenarios for testing the functional form of covariates and/or the proportional hazards assumption.

4.1 Testing the Functional Form of Covariates

We assumed that the failure times follow the hazard functions $\lambda(t \mid \mathbf{Z}) = 2t \exp\{\beta \cdot g(Z_1) - Z_2\}$, where $\beta = 1.5$ and $\mathbf{Z} = (Z_1, Z_2)$, of which Z_1 was generated from a uniform distribution on [0, 4], and Z_2 from an independent Bernoulli distribution with probability 0.5. We generated failure times with sample sizes of 200 and 400. To ensure that the generated data were subject to length-biased sampling, we sampled truncation variables from uniform distributions, allowing for a truncation rate close to 94%. The residual censoring times were also randomly generated from uniform distributions that satisfy censoring rates of 30% and 45%.

Under the null hypothesis, we assumed that the failure times followed the proportional hazards model by setting $g(Z_1) = Z_1$. To investigate the power of our diagnostic method, we considered five alternative functional forms of the first covariate: (a) indicator function, $g(Z_1) = I(Z_1 > 1.5)$; (b) quadratic function, $g(Z_1) = Z_1^2$; (c) square root function, $g(Z_1) = \sqrt{Z_1}$; (d) log function, $g(Z_1) = -\log(Z_1)$; and (e) exponential function $g(Z_1) = \exp(Z_1)$. We computed the test statistic T_1^1 and its corresponding critical values from 1000 resampled statistics to test the functional form of the first covariate, Z_1. Table 1 summarizes the estimated type I error rate under the null hypothesis and the power of the proposed test. When the functional form of the first covariate is linear, the rejection rates range between 0.04 ~ 0.05 and 0.10 ~ 0.14 at significance levels of 5% and 10%, respectively, indicating the proposed test procedure controls the type I error rate reasonably well. Overall, the power of the test increases as the sample size increases, and decreases as the censoring rate increases. We observe that testing the functional form of covariates under alternative (c) has the lowest power. This is likely that the relationship between the first covariate Z_1 and its alternative form $\sqrt{Z_1}$ is relatively close to linear over the support of [0, 4], which eventually results in lower power.

To further investigate the sensitivity of the power of the proposed test to the magnitude of the coefficients, we conducted additional simulations with varying βs under the alternatives (b)–(e). The results are presented in Table 2. The power estimated under alternative (c) is found to be more sensitive to the magnitude of the coefficient (i.e., β), while other alternatives present robust power estimates over different values of β. We note that model departure is more detectable for a functional form of the square root with a larger β. Another interesting observation is that under (e), the power slightly decreases as β increases, which is opposite to the pattern observed under other alternatives. This may be explained by the curve of the exponential function becoming more linear as the exponential form of the covariate stretches vertically by the factor β (see Figure S1, which is available online as supplementary material). Our simulation studies imply that the power of the proposed test on

the functional form of the covariates highly depends on the shape of the distribution of the covariates.

4.2 Testing the Proportional Hazards Assumption

Under the null hypothesis, we generated failure times from the hazard function $\lambda(t \mid \mathbf{Z}) = 2t \exp(Z_1 - Z_2)$, generating Z_1 from a uniform distribution on [0, 1], and Z_2 from an independent Bernoulli distribution with probability 0.5. To evaluate the power of detecting violations of the proportional hazards assumption, we generated failure times from the proportional hazards model with time-dependent covariate effects $\lambda(t \mid \mathbf{Z}) = \exp\{g(t,\mathbf{Z})\}$ as alternatives. We considered two alternative hypotheses: (a) time-dependent model I, $g(t,\mathbf{Z}) = \{1 + 2\log(t)\}Z_1 - Z_2$; and (b) time-dependent model II, $g(t,\mathbf{Z}) = Z_1 - I(t > 0.7)Z_2$. We conducted tests based on the global test statistic T_2. The p-values were computed on the basis of 1000 resampled statistics. The simulation results are summarized in Table 3. Under the null hypothesis, the estimated type I error rates are slightly greater than the nominal levels at significance levels of 5% and 10% when $n = 200$. However, when the sample size increases, they range between 0.05 ~ 0.06 and 0.11 ~ 0.12 at significance levels of 5% and 10%, respectively. The proposed diagnostic tool has adequate power under the two alternatives. Similar to the findings in Table 1, the power increases with the sample size, and decreases for increased censoring rate.

In addition, we examined the power of the test when using $f(Z_{ij}) = Z_{ij}^2$ in (3) for the jth component of the covariates. The additional simulation results are summarized in Table S1 in the supplementary material. We observe that the powers under the alternative hypotheses slightly decrease compared to the test where $f(Z_{ij}) = Z_{ij}$ (results shown in Table 3), but are fairly robust to the choice of $f(\mathbf{Z}_i)$.

5 Applications

The dementia data were collected in Canada from a large prevalent cohort of individuals who were 65 years of age or older, as part of the Canadian Study of Health and Aging (Asgharian et al, 2002; Wolfson et al, 2001). Among individuals who agreed to participate, 1,132 who were confirmed to have dementia at the time of enrollment were followed prospectively from 1991 to 1996 until death. The data consist of the date of disease diagnosis, which was ascertained through medical records; the enrollment date; the date of death or censoring; and the subtype of dementia. The subtypes include probable Alzheimer's disease, possible Alzheimer's disease, and vascular dementia. A total of 818 patients remained in the data cohort after excluding patients with missing entries. In the data set, 22% of the patients were alive and censored at the end of the study. Among all, 393 (48%) were classified as having probable Alzheimer's disease, 252 (31%) as having possible Alzheimer's disease, and 173 (21%) as having vascular dementia. The stationarity assumption was verified (Asgharian et al, 2006), and thus the data were confirmed to be subject to length-biased sampling.

We used the estimating equations of Qin and Shen (2010) to assess the effects of dementia subtypes on the overall survival time under the proportional hazards model. We defined two

indicator variables for probable and possible Alzheimer's disease, respectively, with vascular dementia as the baseline. The estimated coefficient for probable Alzheimer's disease was −0.07 (SE=0.10, p-value=0.47), and that for possible Alzheimer's disease was −0.21 (SE=0.11, p-value=0.05). The results suggest that individuals with possible Alzheimer's disease had longer survival times compared to those with vascular dementia; whereas little difference was detected between individuals with probable Alzheimer's disease and those with vascular dementia. Note that this inferential result is only meaningful when the underlying model assumptions are not violated. We applied the proposed diagnostic method to evaluate whether the proportional hazards model fit the data well. Specifically, we tested the proportional hazards assumption for each covariate, indicating the subtypes of dementia using test statistics T_2^j, $j = 1, 2$. The corresponding critical values were obtained on the basis of 1000 resampled statistics. To check the model assumption graphically, we randomly chose 20 simulated processes among 1000, and plotted them along with the observed process. The results are shown in Figure 1. We can see that both probable and possible Alzheimer's disease satisfy the proportional hazards assumption because the black solid lines derived from the data fluctuate around 0 and lie within the grey lines, which are the 20 randomly selected processes. This is confirmed via the formal test, which gives p-values of 0.35 and 0.92 for probable and possible Alzheimer's disease, respectively. As an alternative, we may assess the proportional hazards model assumptions by comparing the distributions estimated by Vardi's estimator in each subgroup and the distributions based on the Cox model (see Figure S2 in the supplementary material). While this approach may be intuitive, it can only serve as a tool for exploratory data analysis and cannot provide a formal test statistic. Thus, the proposed test procedure is desirable.

6 Discussion

To the best of our knowledge, no model diagnostic method that examines the adequacy of the proportional hazards model has been studied for length-biased data. In this paper, we proposed model checking tools based on the cumulative sums of mean zero stochastic processes, which is similar to the approach of Lin et al (1993). Model checking is an essential step that needs to be implemented for valid inference. The proposed diagnostic method provides both graphical plots and formal analytical tests to detect model departure. When the functional form of the continuous covariates is found to deviate from the assumed linear functional form, one may consider the proper transformation of the covariates. If the proportional hazards assumption fails, it is worthwhile to examine other models such as the accelerated failure time model or the semiparametric transformation model (Shen et al, 2009).

We assume that censoring is independent of covariates in the construction of a stochastic process in equation (2), and can relax this assumption by revising the weight function. In application, one can first check if the censoring times are independent of covariates as in conventional survival analyses. When the censoring distribution depends on covariates, the weight function $w_C(t)$ needs to be replaced by $w_C(t \mid z) = \int_0^t S_C(u \mid z)\mathrm{d}u$, where $S_C(t \mid \boldsymbol{z})$ is the conditional survival distribution of the residual censoring time given the covariates. The conditional weight function $w_C(t \mid \boldsymbol{z})$ can be consistently estimated by

$\hat{w}_C(t \mid z) = \int_0^t \hat{S}_C(u \mid z)\mathrm{d}u$, where $\hat{S}_C(t \mid \boldsymbol{z})$ is estimated by the covariate-specific Kaplan–Meier estimator when the covariates are discrete. For continuous covariates, we can either postulate a regression model for the censoring distribution or adopt the local Kaplan–Meier estimator (Wang and Wang, 2014). We conducted a set of simulation studies to examine the robustness of the proposed model checking method to misspecification of the censoring distribution. The results show that the proposed model checking method is reasonably robust to the violation of the covariate-independent censoring assumption. In addition, we evaluated the performance of the proposed test generalized to account for covariate-dependent censoring. The generalized test provides adequate levels of power under the alternative hypotheses. The simulation settings and results are summarized in the supplementary material (see Table S2).

While the proposed model diagnostics require approximating the limiting distribution of a stochastic process, we find it easy and fast to achieve via the computationally efficient resampling method, proposed in Section 3. Unlike the method of Lin et al (1993) or its extensions, our method has an extra layer of complexity to adjust for sampling bias. Specifically, the constructed stochastic processes have a term induced from the uncertainty of the weight function. However, the additional term does not introduce much computational burden when the resampling method is adopted.

For data subject to length bias, various semiparametric models have been considered; however, model checking has been less represented in the literature. It is certainly warranted to develop test procedures to test the regression models based on the cumulative sum of residuals for length-biased data. Under one special case, where the proportional mean residual life model is assumed for the target population, the model checking method for conventional survival data (Lin et al, 1993) can be directly applied to the length-biased data because the length-biased subpopulation would follow the proportional hazards model (Chan et al, 2012).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This 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 at The University of Texas at Austin for providing HPC resources that contributed to the research results reported within this paper.

References

Asgharian M, M'Lan CE, Wolfson DB. 2002; Length-biased sampling with right censoring: an unconditional approach. Journal of American Statistical Association. 97:201–209.

Asgharian M, Wolfson DB, Zhang X. 2006; Checking stationarity of the incidence rate using prevalent cohort survival data. Statistics in Medicine. 25:1751–1767. [PubMed: 16220462]

Borgan O, Zhang Y. 2015; Using cumulative sums of martingale residuals for model checking in nested case-control studies. Biometrics. 71:696–703. [PubMed: 25854648]

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

Cox DR. 1972; Regression models and life-tables (with discussion). Journal of Royal Statistical Society Series B (Methodological). 34:187–220.

Huang CY, Qin J. 2012; Composite partial likelihood estimation under length-biased sampling, with application to a prevalent cohort study of dementia. Journal of American Statistical Association. 107:946–957.

Huang CY, Luo X, Follmann DA. 2011; A model checking method for the proportional hazards model with recurrent gap time data. Biostatistics. 12:535–547. [PubMed: 21138876]

Kosorok, MR. Introduction to Empirical Processes and Semiparametric Inference. New York: Springer; 2008.

Lancaster T. 1979; Econometric methods for the duration of unemployment. Econometrica. 47:939–956.

Lin DY, Wei LJ, Ying ZL. 1993; Checking the cox model with cumulative sums of martingale-based residuals. Biometrika. 80:557–572.

Lu W, Liu M, Chen YH. 2014; Testing goodness-of-fit for the proportional hazards model based on nested case-control data. Biometrics. 70:845–851. [PubMed: 25298193]

Pepe MS, Fleming TR. 1991; Weighted kaplan-meier statistics: large sample and optimality considerations. Journal of Royal Statistical Society B. 53:341–352.

Qin J, Shen Y. 2010; Statistical methods for analyzing right-censored length-biased data under cox model. Biometrics. 66:382–392. [PubMed: 19522872]

Qin J, Ning J, Liu H, Shen Y. 2011; Maximum likelihood estimations and em algorithms with length-biased data. Journal of American Statistical Association. 106:1434–1449.

Shen Y, Ning J, Qin J. 2009; Analyzing length-biased data with semiparametric transformation and accelerated failure time models. Journal of American Statistical Association. 104:1192–1202.

Spiekerman CF, Lin DY. 1996; Checking the marginal cox model for correlated failure time data. Biometrika. 83:143–156.

Tsai WY. 2009; Pseudo-partial likelihood for proportional hazards models with biased-sampling data. Biometrika. 96:601–615. [PubMed: 22422175]

de Una-Alvarez J, Otero-Giraldez MS, Alvarez-Llorente G. 2003; Estimation under length-bias and right-censoring: an application to unemployment duration analysis for married women. Journal of Applied Statistics. 30:283–291.

Wang HJ, Wang L. 2014; Quantile regression analysis of length-biased survival data. Stat. 3:31–47.

Wang MC. 1996; Hazards regression analysis for length-biased data. Biometrika. 83:343–354.

Wolfson C, Wolfson DB, Asgharian M, M'Lan CE, Ostbye T, Rockwood K, Hogan DB. for the clinical progression of dementia study group. 2001; A reevaluation of the duration of survival after the onset of dementia. New England Journal of Medicine. 344:1111–1116. [PubMed: 11297701]

Zelen M, Feinleib M. 1969; On the theory of screening for chronic diseases. Biometrika. 56:601–614.

Appendix A: Regularity Conditions

We assume the following regularity conditions for the large sample properties:

1. $(Y_i, A_i, \delta_i, \boldsymbol{Z}_i)$ are independent and identically distributed for $i = 1, \ldots, n$.
2. The parameters $\boldsymbol{\beta}_0$ belong to an interior of a known compact set.
3. The covariates $\boldsymbol{Z}$ are bounded, and $\boldsymbol{\alpha} = 0$ almost surely if $\boldsymbol{\alpha}^\top \boldsymbol{Z} = 0$ with probability one.

4. The differentiable baseline cumulative hazard function $\Lambda_0(\tau) < \infty$ where τ satisfies $\Pr(Y \geq \tau) > 0$.

5. $\Gamma(\boldsymbol{\beta})$ is positive definite.

6. $0 < w_C(\tau) < \infty$ and $\int_0^\tau \left[\{\int_t^\tau S_C(u)\mathrm{d}u\}^2 / \{S_C^2(t) S_V(t)\} \right] \mathrm{d}S_C(t) < \infty$ where $S_V(\cdot)$ is the survival function of the residual survival time.

Appendix B: Proof of Theorem 1

By applying the Taylor series expansions, we have

$$
\begin{aligned}
& n^{-1/2} \sum_{i=1}^{n} \int_0^t \hat{E}_{\mathbf{Z}}(\boldsymbol{\beta}, u, z) \mathrm{d}N_i(u) \\
& = n^{-1/2} \sum_{i=1}^{n} \int_0^t \hat{E}_{\mathbf{Z}}(\boldsymbol{\beta}_0, u, z) \mathrm{d}N_i(u) + n^{-1} \sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\beta}} \int_0^t \hat{E}_{\mathbf{Z}}(\boldsymbol{\beta}_0, u, z) \mathrm{d}N_i(u) \sqrt{n}(\boldsymbol{\beta} - \boldsymbol{\beta}_0) + o_p(1), \\
& = n^{-1/2} \sum_{i=1}^{n} \int_0^t \hat{E}_{\mathbf{Z}}(\boldsymbol{\beta}_0, u, z) \mathrm{d}N_i(u) \\
& + n^{-1} \sum_{i=1}^{n} \int_0^t \left[\frac{\hat{S}_{\mathbf{Z}}^{(1)}(\boldsymbol{\beta}_0, u, z)}{\hat{S}^{(0)}(\boldsymbol{\beta}_0, u)} - \frac{\hat{S}_{\mathbf{Z}}^{(0)}(\boldsymbol{\beta}_0, u, z) \hat{S}^{(1)}(\boldsymbol{\beta}_0, u)}{\{\hat{S}^{(0)}(\boldsymbol{\beta}_0, u)\}^2} \right] \mathrm{d}N_i(u) \sqrt{n}(\boldsymbol{\beta} - \boldsymbol{\beta}_0) + o_p(1),
\end{aligned}
\tag{5}
$$

and

$$n^{-1/2}\sum_{i=1}^{n}\int_0^{\tau}\left\{\boldsymbol{Z}_i-\hat{E}(\boldsymbol{\beta},u)\right\}\mathrm{d}N_i(u) \tag{6}$$

$$=n^{-1/2}\sum_{i=1}^{n}\int_0^{\tau}\left\{\boldsymbol{Z}_i-\hat{E}(\boldsymbol{\beta}_0,u)\right\}\mathrm{d}N_i(u)+n^{-1}\sum_{i=1}^{n}\frac{\partial}{\partial\boldsymbol{\beta}}\int_0^{\tau}\left\{\boldsymbol{Z}_i-\hat{E}(\boldsymbol{\beta}_0,u)\right\}\mathrm{d}N_i(u)\times\sqrt{n}(\boldsymbol{\beta}$$

$$-\boldsymbol{\beta}_0)+o_p(1)$$

$$=n^{-1/2}\sum_{i=1}^{n}\int_0^{\tau}\left\{\boldsymbol{Z}_i-\hat{E}(\boldsymbol{\beta}_0,u)\right\}\mathrm{d}N_i(u)$$

$$=n^{-1}\sum_{i=1}^{n}\int_0^{\tau}\left[\frac{\hat{S}^{(2)}(\boldsymbol{\beta}_0,u)}{\hat{S}^{(0)}(\boldsymbol{\beta}_0,u)}-\left\{\frac{\hat{S}^{(1)}(\boldsymbol{\beta}_0,u)^2}{\hat{S}^{(0)}(\boldsymbol{\beta}_0,u)}\right\}^2\right]\mathrm{d}N_i(u)\sqrt{n}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)+o_p(1).$$

The stochastic processes $\boldsymbol{G}(t,\boldsymbol{z})$ can be expressed in two terms, $\boldsymbol{G}_1(t,\boldsymbol{z})$ and $\boldsymbol{G}_2(t,\boldsymbol{z})$, as follows.

$$\boldsymbol{G}(t,z)$$

$$=\sum_{i=1}^{n}f(\boldsymbol{Z}_i)I(\boldsymbol{Z}_i\le z)\widehat{M}_i(t)$$

$$=\sum_{i=1}^{n}f(\boldsymbol{Z}_i)I(\boldsymbol{Z}_i\le z)N_i(t)-f(\boldsymbol{Z}_i)I(\boldsymbol{Z}_i\le z)\int_0^t\widehat{w}_C(u)R_i(u)\{\widehat{w}_C(Y_i)\}^{-1}\exp(\hat{\boldsymbol{\beta}}^{\top}\boldsymbol{Z}_i)\mathrm{d}\hat{\Lambda}_0(u)$$

$$=\sum_{i=1}^{n}f(\boldsymbol{Z}_i)I(\boldsymbol{Z}_i\le z)N_i(t)-\int_0^t\frac{\hat{S}_Z^{(0)}(\hat{\boldsymbol{\beta}},u,z)}{\hat{S}^{(0)}(\hat{\boldsymbol{\beta}},u)}\mathrm{d}N_i(u)$$

$$=\sum_{i=1}^{n}\int_0^t\{f(\boldsymbol{Z}_i)I(\boldsymbol{Z}_i\le z)-E_Z(\boldsymbol{\beta}_0,u,z)\}\,\mathrm{d}N_i(u)+\sum_{i=1}^{n}\int_0^t\left\{E_Z(\boldsymbol{\beta}_0,u,z)-\hat{E}_Z(\hat{\boldsymbol{\beta}},u,z)\right\}\mathrm{d}N_i(u)$$

$$=\boldsymbol{G}_1(t,z)+\boldsymbol{G}_2(t,z)$$

We exploit the Taylor expansion and empirical process approximation techniques. It is straight-forward that the first term can be approximated by

$$n^{-1/2}\boldsymbol{G}_1(t,z)=n^{-1/2}\sum_{i=1}^{n}\int_0^t\{f(\boldsymbol{Z}_i)I(\boldsymbol{Z}_i\le z)-E_Z(\boldsymbol{\beta}_0,u,z)\}\,\mathrm{d}N_i(u)$$

$$=n^{-1/2}\sum_{i=1}^{n}\int_0^t\{f(\boldsymbol{Z}_i)I(\boldsymbol{Z}_i\le z)-e_Z(\boldsymbol{\beta}_0,u,z)\}\,\mathrm{d}M_i(u)+o_p(1).$$

Then, we re-express the second term based on equations (5) and (6).

$$n^{-1/2}\boldsymbol{G}_2(t,z) = n^{-1/2}\sum_{i=1}^{n}\int_0^t E_Z(\boldsymbol{\beta}_0,u,z)\mathrm{d}N_i(u) - n^{-1/2}\sum_{i=1}^{n}\int_0^t \hat{E}_Z(\hat{\boldsymbol{\beta}},u,z)\mathrm{d}N_i(u)$$
$$= n^{-1/2}\sum_{i=1}^{n}\int_0^t \left\{E_Z(\boldsymbol{\beta}_0,u,z) - \hat{E}_Z(\boldsymbol{\beta}_0,u,z)\right\}\mathrm{d}N_i(u) - \Gamma_Z(\boldsymbol{\beta}_0,t,z)\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0) + o_p(1)$$
$$= n^{-1/2}\sum_{i=1}^{n}\int_0^t \left\{E_Z(\boldsymbol{\beta}_0,u,z) - \hat{E}_Z(\boldsymbol{\beta}_0,u,z)\right\}\mathrm{d}N_i(u) + \Gamma_Z(\boldsymbol{\beta}_0,t,z)\{\Gamma(\boldsymbol{\beta}_0)\}^{-1}n^{-1/2}\hat{U}(\boldsymbol{\beta}_0) + o_p(1)$$

where

$$\Gamma_Z(\boldsymbol{\beta},t,z) = \mathrm{E}\left\{\int_0^t\left[\frac{S_{\boldsymbol{Z}}^{(1)}(\boldsymbol{\beta},u,z)}{S^{(0)}(\boldsymbol{\beta},u)} - \frac{S_{\boldsymbol{Z}}^{(0)}(\boldsymbol{\beta},u,z)S^{(1)}(\boldsymbol{\beta},u)}{\{S^{(0)}(\boldsymbol{\beta},u)\}^2}\right]\mathrm{d}N_i(u)\right\}$$

and

$$\Gamma(\boldsymbol{\beta}) = -\mathrm{E}\left\{\int_0^\tau\left[\frac{S^{(2)}(\boldsymbol{\beta},u)}{S^{(0)}(\boldsymbol{\beta},u)} - \left\{\frac{S^{(1)}(\boldsymbol{\beta},u)}{S^{(0)}(\boldsymbol{\beta},u)}\right\}^2\right]\mathrm{d}N_i(u)\right\}.$$

The second equation can be derived by plugging in equation (5). The third equation naturally follows by replacing $\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)$ with equation (6) after some algebra (Qin and Shen, 2010). Note that the leading term in the last equation can be rewritten as

$$n^{-1/2}\sum_{i=1}^{n}\int_0^t \left\{E_Z(\boldsymbol{\beta}_0,u,z) - \hat{E}_Z(\boldsymbol{\beta}_0,u,z)\right\}\mathrm{d}N_i(u)$$
$$= n^{-1/2}\sum_{i=1}^{n}\int_0^t \left\{\frac{S_Z^{(0)}(\boldsymbol{\beta}_0,u) - \hat{S}_Z^{(0)}(\boldsymbol{\beta}_0,u)}{S^{(0)}(\boldsymbol{\beta}_0,u)}\right\}\mathrm{d}N_i(u) + o_p(1)$$
$$= n^{-1/2}\sum_{i=1}^{n}\int_0^t \frac{\sum_{k=1}^n f(\boldsymbol{Z}_k)I(\boldsymbol{Z}_k\le z)R_k(u)\exp(\boldsymbol{\beta}_0^\top\boldsymbol{Z}_k)\left\{\frac{1}{w_C(Y_k)} - \frac{1}{\hat{w}_C(Y_k)}\right\}}{\sum_{k=1}^n R_k(u)\{w_C(Y_k)\}^{-1}\exp(\boldsymbol{\beta}_0^\top\boldsymbol{Z}_k)}\mathrm{d}N_i(u) + o_p(1)$$
$$= n^{-1/2}\sum_{i=1}^{n}\int_0^t \frac{\sum_{k=1}^n f(\boldsymbol{Z}_k)I(\boldsymbol{Z}_k\le z)w_C(u)R_k(u)\exp(\boldsymbol{\beta}_0^\top\boldsymbol{Z}_k)}{nS^{(0)}(\boldsymbol{\beta}_0,u)}\frac{\{\hat{w}_C(Y_k) - w_C(Y_k)\}}{\{\hat{w}_C(Y_k)\}^2}\mathrm{d}N_i(u) + o_p(1)$$
$$= n^{-1/2}\sum_{i=1}^{n}\int_0^t H(\boldsymbol{\beta}_0,u)\frac{\mathrm{d}M_{C_i}(u)}{\pi(u)} + o_p(1),$$

where

$$H(\boldsymbol{\beta}, t) = \lim_{n \to \infty} \frac{1}{n^2} \sum_{i=1}^{n} \sum_{k=1}^{n} \frac{f(\mathbf{Z}_k) I(\mathbf{Z}_k \le z) w_C(Y_i) R_k(Y_i) \exp(\boldsymbol{\beta}^\top \mathbf{Z}_k) \{w_C(Y_k)\}^{-2} h_k(t)}{S^{(0)}(\boldsymbol{\beta}, Y_i)}$$

$$M_{C_i}(t) = I(V_i \le t, \delta_i = 0) - \int_0^t I(V_i \ge u) \mathrm{d}\Lambda_C(u)$$

$$h_k(t) = I(Y_k \ge t) \int_t^{Y_k} S_C(u) \mathrm{d}u$$

$$\pi(t) = S_C(t) S_V(t),$$

in which $\Lambda_C(t)$ is the cumulative hazard function of the residual censoring time and $S_V(t)$ is the survival function of the residual survival time. The last equation can be obtained by expressing $\{\hat{w}_C(y) - w_C(y)\}$ as an i.i.d. sum of martingales (Pepe and Fleming, 1991). Finally, the general class of stochastic processes $G(t, \mathbf{z})$ can be asymptotically represented by

$$\begin{aligned} n^{-1/2} \boldsymbol{G}(t, z) &= n^{-1/2} \sum_{i=1}^{n} \int_0^t \{f(\mathbf{Z}_i) I(\mathbf{Z}_i \le z) - e_Z(\boldsymbol{\beta}_0, u)\} \, \mathrm{d}M_i(u) \\ &+ n^{-1/2} \sum_{i=1}^{n} \int_0^t H(\boldsymbol{\beta}_0, u) \frac{\mathrm{d}M_{C_i}(u)}{\pi(u)} \\ &+ \Gamma_Z(\boldsymbol{\beta}_0, t, z) \{\Gamma(\boldsymbol{\beta}_0)\}^{-1} n^{-1/2} \sum_{i=1}^{n} \int_0^\infty \{\mathbf{Z}_i - e(\boldsymbol{\beta}_0, u)\} \, \mathrm{d}M_i(u) + o_p(1) \end{aligned} \tag{7}$$

$$= n^{-1/2} \sum_{i=1}^{n} \boldsymbol{G}_i^*(t, z) + o_p(1). \tag{8}$$

Under the regularity conditions, for any given $\mathbf{z}$, $\boldsymbol{G}_i^*(t, z)$ is a mean zero process bounded on $[0, \tau]$. This process can be classified as a Donsker class (Kosorok, 2008). Thus, as $n \to \infty$, the summation of $\boldsymbol{G}_i^*(t, z)$ in (8) converges weakly to a mean zero Gaussian process, for which the asymptotic covariance function is $\mathrm{E}\{\boldsymbol{G}_i^*(t_1, z_1) \boldsymbol{G}_i^*(t_2, z_2)^\top\}$.

Appendix C: Proof of Theorem 2

We note that $\Gamma_Z(\boldsymbol{\beta}_0, t, \mathbf{z})\{\Gamma(\boldsymbol{\beta}_0)\}^{-1}$ in the third term of (7) converges in probability to a non-random function. Conditional on the observed data, the process (7) is a linear combination of independent normally distributed processes with mean zero. Thus, given that $\hat{\boldsymbol{\beta}}$ is a consistent estimator for $\boldsymbol{\beta}_0$, we can show that $n^{-1} \sum_{i=1}^{n} \hat{\boldsymbol{G}}_i^*(t_1, z_1) \hat{\boldsymbol{G}}_i^*(t_2, z_2)^\top$ converges in

probability to the asymptotic covariance function $\mathrm{E}\{\boldsymbol{G}_i^*(t_1, z_1)\boldsymbol{G}_i^*(t_2, z_2)^\top\}$ as $n \to \infty$. By applying the multiplier central limit theorem (Kosorok, 2008), it follows that conditional on the observed data and $\boldsymbol{z}$, $n^{-1/2}\tilde{\boldsymbol{G}}_m(t, \boldsymbol{z})$ and $n^{-1/2}\sum_{i=1}^n \boldsymbol{G}_i^*(t, z)$ converge to the same mean zero Gaussian process.

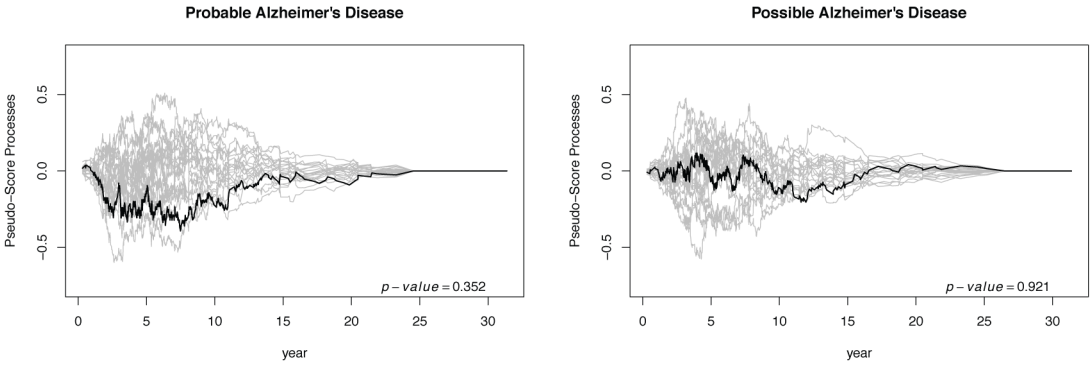

Fig. 1.

Graphical results of testing the proportional hazards assumption with the dementia data from the Canadian study

Table 1

Summary of type I error rate under the null hypothesis and the powers under the alternative hypotheses for testing the functional form of the first covariate with sample sizes (n) of 200 and 400 and censoring rates (cr) of 30% and 45% at significance levels of 5% and 10%.

		Null hypothesis		**(a) Indicator function**		**(b) Quadratic function**	
n	**cr**	**5%**	**10%**	**5%**	**10%**	**5%**	**10%**
200	30%	0.051	0.124	0.484	0.665	0.815	0.904
	45%	0.052	0.137	0.356	0.540	0.671	0.820
400	30%	0.044	0.106	0.885	0.952	0.981	0.993
	45%	0.048	0.100	0.766	0.878	0.947	0.975
		(c) Square root function		(d) Log function		(e) Exponential function	
n	cr	5%	10%	5%	10%	5%	10%
200	30%	0.338	0.445	0.397	0.554	0.536	0.678
	45%	0.235	0.367	0.334	0.496	0.439	0.615
400	30%	0.602	0.726	0.763	0.884	0.822	0.926
	45%	0.423	0.561	0.687	0.820	0.758	0.864

Table 2

Summary of the powers under alternative hypotheses (b)–(d) for testing the functional form of the first covariate with sample sizes (n) of 200 and 400, censoring rates (cr) of 30% and 45%, and varying βs at significance levels of 5% and 10%.

			(b)		(c)		(d)		(e)	
n	**cr**	$\boldsymbol{\beta}$	**5%**	**10%**	**5%**	**10%**	**5%**	**10%**	**5%**	**10%**
200	30%	1.0	0.819	0.904	0.178	0.291	0.358	0.517	0.613	0.772
		1.5	0.815	0.904	0.338	0.445	0.397	0.554	0.536	0.678
		2.0	0.828	0.911	0.536	0.684	0.429	0.579	0.452	0.609
	45%	1.0	0.692	0.811	0.135	0.247	0.278	0.454	0.515	0.683
		1.5	0.671	0.820	0.235	0.367	0.334	0.496	0.439	0.615
		2.0	0.696	0.827	0.389	0.497	0.353	0.504	0.370	0.515
400	30%	1.0	0.981	0.993	0.282	0.400	0.694	0.833	0.919	0.962
		1.5	0.981	0.993	0.602	0.726	0.763	0.884	0.822	0.926
		2.0	0.993	1.000	0.852	0.913	0.781	0.868	0.796	0.875
	45%	1.0	0.952	0.987	0.201	0.314	0.611	0.772	0.856	0.936
		1.5	0.947	0.975	0.423	0.561	0.687	0.820	0.758	0.864
		2.0	0.969	0.992	0.652	0.772	0.707	0.834	0.672	0.814

Table 3

Summary of type I error rate under the null hypothesis and the powers under the alternative hypotheses for testing the proportional hazards assumption with sample sizes (n) of 200 and 400 and censoring rates (cr) of 30% and 45% at significance levels of 5% and 10%.

		Null hypothesis		(a) Time-dependent I		(b) Time-dependent II	
n	cr	5%	10%	5%	10%	5%	10%
200	30%	0.066	0.138	0.395	0.556	0.684	0.791
	45%	0.063	0.130	0.312	0.481	0.628	0.752
400	30%	0.051	0.107	0.746	0.864	0.924	0.963
	45%	0.060	0.124	0.624	0.760	0.905	0.944