Biostatistics (2011), **12**, 2, *pp.* 354–368
doi:10.1093/biostatistics/kxq061
Advance Access publication on September 21, 2010

Estimation of the 2-sample hazard ratio function using a semiparametric model

SONG YANG*

Office of Biostatistics Research, National Heart, Lung, and Blood Institute, 6701 Rockledge Drive, MSC 7913, Bethesda, MD 20892, USA
yangso@nhlbi.nih.gov

ROSS L. PRENTICE

Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, PO Box 19024 Seattle, WA 98109, USA

SUMMARY

The hazard ratio provides a natural target for assessing a treatment effect with survival data, with the Cox proportional hazards model providing a widely used special case. In general, the hazard ratio is a function of time and provides a visual display of the temporal pattern of the treatment effect. A variety of nonproportional hazards models have been proposed in the literature. However, available methods for flexibly estimating a possibly time-dependent hazard ratio are limited. Here, we investigate a semiparametric model that allows a wide range of time-varying hazard ratio shapes. Point estimates as well as pointwise confidence intervals and simultaneous confidence bands of the hazard ratio function are established under this model. The average hazard ratio function is also studied to assess the cumulative treatment effect. We illustrate corresponding inference procedures using coronary heart disease data from the Women's Health Initiative estrogen plus progestin clinical trial.

Keywords: Clinical trial; Empirical process; Gaussian process; Hazard ratio; Simultaneous inference; Survival analysis; Treatment–time interaction.

1. INTRODUCTION

Consider the comparison of failure times between a treated and control group under independent censorship. The hazard ratio provides a natural target of estimation in many applications since it permits a focus on relative failure rates across the study follow-up period, without the need to model absolute failure rates, which may be sensitive to study eligibility criteria and other factors. The proportional hazards special case of the Cox (1972) regression model is widely used for hazard ratio estimation. The maximum partial likelihood procedure (Cox, 1975) provides a convenient and robust means of estimating a constant hazard ratio and yields a log-rank procedure for testing equality of hazards between the 2 groups.

*To whom correspondence should be addressed.

In general, the hazard ratio may be a function of time, and estimation of the hazard ratio function may provide useful insights into temporal aspects of treatment effects. For example, Gilbert *and others* (2002) develop a nonparametric estimation procedure for the log-hazard ratio function with simultaneous confidence bands, for use as an exploratory data analytic tool. Naturally, confidence bands may be wide with such a nonparametric estimator, particularly at longer follow-up times where data may be sparse. See also Gray (1992), Kooperberg *and others* (1995), Cai and Sun (2003), Tian *and others* (2005), Abrahamowicz and Mackenzie (2007), and Peng and Huang (2007), and references therein, for additional related work.

Parametric or semiparametric hazard ratio models have potential to contribute valuably to treatment effect assessment. Hazard ratio models having parameters of useful interpretation, and that embrace a range of hazard ratio shapes, may be particularly valuable. The Cox model allows time-varying covariates to be defined that can, for example, allow separate hazard ratios for the elements of a partition of the time axis or allow the hazard ratio to be a parametric function of follow-up time more generally. Various other semiparametric regression models have been proposed for failure time data analyses, including accelerated failure time models, proportional odds models, and linear transformation models, many of which are embraced by the broad class of models for which Zeng and Lin (2007) develop maximum likelihood estimation procedures. Some more semiparametric models can be found in Vaupel *and others* (1979), Hsieh (1996), Chen and Wang (2000), Tsodikov (2002), Yang and Prentice (2005), and Chen and Cheng (2006). Many of these models induce a semiparametric class of models for the hazard ratio function that includes proportional hazards as a special case. Hazard ratio estimators under such semiparametric models can avoid the instability that may attend nonparametric hazard ratio function estimators.

One of these, proposed by Yang and Prentice (2005), involves short-term and long-term hazard ratio parameters, and a hazard ratio function that depends also on the control group survivor function. Assume absolutely continuous failure times and label the 2 groups control and treatment, with hazard functions $\lambda_C(t)$ and $\lambda_T(t)$, respectively. Let $h(t) = \lambda_T(t)/\lambda_C(t)$ be the hazard ratio function and $S_C(t)$ the survivor function of the control group. The model postulates that

$$h(t) = \frac{1}{e^{-\beta_2} + (e^{-\beta_1} - e^{-\beta_2})S_C(t)}, \quad t < \tau_0, \tag{1.1}$$

where β_1 and β_2 are scalar parameters and

$$\tau_0 = \sup\left\{x: \int_0^x \lambda_C(t)dt < \infty\right\}. \tag{1.2}$$

This model includes the proportional hazards model and the proportional odds model as special cases. It has a monotone $h(t)$ with a variety of patterns, including proportional hazards, no initial effect, disappearing effect, and crossing hazards, among others. Thus, the model presumably entails sufficient flexibility for many applications. It has also been studied for current status data in Tong *and others* (2007).

In comparison, for many commonly used special cases of the accelerated failure time model either $\lim_{t\downarrow 0} h(t) = 1$ or $\lim_{t\uparrow\tau_0} h(t) \in \{0, 1, \infty\}$ and the hazard ratio stays above or below one when λ_C is increasing. This is less flexible than desired. For the class of linear transformation models, with the logarithmic transformation, the hazard ratio also inherits some of these restrictions at many common baseline distributions. Similar properties hold as well for many other semiparametric models.

Under model (1.1), estimation procedures to date have focused on the finite-dimensional parameters, as has mostly been the case also for estimation under other semiparametric models. Here, we extend the estimation to pointwise and simultaneous inference on the hazard ratio function itself. First, consistency and asymptotic normality of the estimate at a fixed time point are established. Then procedures for constructing pointwise confidence intervals and simultaneous confidence bands for the hazard ratio function are developed, and some modifications are implemented to improve moderate sample size performance.

For additional display of the treatment effect, simultaneous confidence bands are also obtained for the average hazard ratio function over a time interval. The average hazard ratio gives a summary measure of treatment comparison and provides a picture of the cumulative treatment effect to augment display of the temporal pattern of the hazard ratio. These hazard ratio estimation procedures are applied to data from the Women's Health Initiative (WHI) estrogen plus progestin clinical trial (Writing Group For the Women's Health Initiative Investigators, 2002; Manson *and others*, 2003), which yielded a hazard ratio function for the primary coronary heart disease outcome that was decidedly nonproportional. Understanding the hazard ratio function shape in this setting was important to integrating the clinical trial data with a large body of preceding observational literature that had failed to identify an early hazard ratio increase (e.g. Manson *and others*, 2003; Prentice *and others*, 2005).

We organize the article as follows: In Section 2, the short-term and long-term hazard ratio model and the hazard ratio estimate are described. Pointwise confidence intervals of the hazard ratio are established. Simultaneous confidence bands for the hazard ratio and the average hazard ratio are provided in Section 3. Simulation results are presented in Section 4. Application to data from the WHI trial is given in Section 5. Some concluding remarks are given in Section 6. Proofs of the asymptotic results are contained in the Supplementary Material available at *Biostatistics* online.

2. HAZARD RATIO FUNCTION ESTIMATION

Let $T_1, \ldots, T_n$ be the pooled lifetimes of the 2 groups, with $T_1, \ldots, T_{n_1}$, $n_1 < n$, constituting the control group having the survivor function S_C. Let $C_1, \ldots, C_n$ be the censoring variables, and $Z_i = I(i > n_1), i = 1, \ldots, n$, where $I(\cdot)$ is the indicator function. The available data consist of the independent triplets (X_i, δ_i, Z_i), $i = 1, \ldots, n$, where $X_i = \min(T_i, C_i)$ and $\delta_i = I(T_i \leqslant C_i)$. We assume that T_i and C_i are independent given Z_i. The censoring variables (C_i's) need not be identically distributed, and in particular, the 2 groups may have different censoring patterns. For $t < \tau_0$ with τ_0 defined in (1.2), let $R(t)$ be the the odds function $1/S_C(t) - 1$ of the control group. The model of Yang and Prentice (2005) can be expressed as

$$\lambda_i(t) = \frac{1}{e^{-\beta_1 Z_i} + e^{-\beta_2 Z_i} R(t)} \frac{dR(t)}{dt}, \quad i = 1, \ldots, n, \quad t < \tau_0, \tag{2.1}$$

where $\lambda_i(t)$ is the hazard function for T_i given Z_i. Under the model, the hazard ratio is

$$h(t) = \frac{1 + R(t)}{e^{-\beta_1} + e^{-\beta_2} R(t)}.$$

To estimate $h(t)$, we need to estimate the parameter $\boldsymbol{\beta} = (\beta_1, \beta_2)^T$ and the baseline function $R(t)$, where "T" denotes transpose. Let us first introduce the estimators from Yang and Prentice (2005).

Define

$$K(t) = \sum_{i=1}^{n} I(X_i \geqslant t), \quad H_j(t; \mathbf{b}) = \sum_{i=1}^{n} \delta_i e^{-b_j Z_i} I(X_i \leqslant t), \quad j = 1, 2,$$

where $\mathbf{b} = (b_1, b_2)^T$. Let $\tau < \tau_0$ be such that

$$\lim_n K(\tau) > 0, \tag{2.2}$$

with probability 1. For $t \leqslant \tau$, let

$$\hat{P}(t; \mathbf{b}) = \prod_{s \leqslant t} \left\{ 1 - \frac{\Delta H_2(s; \mathbf{b})}{K(s)} \right\}, \quad \hat{R}(t; \mathbf{b}) = \frac{1}{\hat{P}(t; \mathbf{b})} \int_0^t \frac{\hat{P}_-(s; \mathbf{b})}{K(s)} H_1(ds; \mathbf{b}),$$

where $\Delta H_2(s;\mathbf{b})$ denotes the jump of $H_2(s;\mathbf{b})$ in s and $\hat{P}_-(s;\mathbf{b})$ denotes the left continuous (in s) version of $\hat{P}(s;\mathbf{b})$, Define the martingale residuals

$$\hat{M}_i(t;\mathbf{b}) = \delta_i I(X_i \leqslant t) - \int_0^t I(X_i \geqslant s)\frac{\hat{R}(\mathrm{d}s;\mathbf{b})}{\mathrm{e}^{-b_1 Z_i} + \mathrm{e}^{-b_2 Z_i}\hat{R}(s;\mathbf{b})}, \quad 1 \leqslant i \leqslant n.$$

Yang and Prentice (2005) proposed a pseudo maximum likelihood estimator $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \hat{\beta}_2)^T$ of $\boldsymbol{\beta}$, which is the zero of $Q(\mathbf{b})$, where

$$Q(\mathbf{b}) = \sum_{i=1}^{n}\int_0^\tau f_i(t;\mathbf{b})\hat{M}_i(\mathrm{d}t;\mathbf{b}), \tag{2.3}$$

with $f_i = (f_{1i}, f_{2i})^T$, where

$$f_{1i}(t;\mathbf{b}) = \frac{Z_i\,\mathrm{e}^{-b_1 Z_i}}{\mathrm{e}^{-b_1 Z_i} + \mathrm{e}^{-b_2 Z_i}\hat{R}(t;\mathbf{b})}, \quad f_{2i}(t;\mathbf{b}) = \frac{Z_i\,\mathrm{e}^{-b_2 Z_i}\hat{R}(t;\mathbf{b})}{\mathrm{e}^{-b_1 Z_i} + \mathrm{e}^{-b_2 Z_i}\hat{R}(t;\mathbf{b})}.$$

Once $\hat{\boldsymbol{\beta}}$ is obtained, $R(t)$ can be estimated by $\hat{R}(t;\hat{\boldsymbol{\beta}})$, and the hazard ratio $h(t)$ can be estimated by

$$\hat{h}(t) = \frac{1 + \hat{R}(t;\hat{\boldsymbol{\beta}})}{\mathrm{e}^{-\beta_1} + \mathrm{e}^{-\beta_2}\hat{R}(t;\hat{\boldsymbol{\beta}})}.$$

In Appendix A of the Supplementary Material available at *Biostatistics* online, we show that $\hat{h}(t)$ is strongly consistent for $h(t)$ under model (2.1).

To study the distributional properties of $\hat{h}(t)$, let

$$W_n(t) = \sqrt{n}(\hat{h}(t) - h(t)), \quad t \leqslant \tau.$$

For the asymptotic distribution of $\hat{\boldsymbol{\beta}}$, define

$$A(t) = \left(\frac{\mathrm{e}^{-\beta_1}}{\mathrm{e}^{-\beta_1} + \mathrm{e}^{-\beta_2}\hat{R}(t;\boldsymbol{\beta})}, \frac{\mathrm{e}^{-\beta_2}\hat{R}(t;\boldsymbol{\beta})}{\mathrm{e}^{-\beta_1} + \mathrm{e}^{-\beta_2}\hat{R}(t;\boldsymbol{\beta})}\right)^T,$$

$$K_1(t) = \sum_{i \leqslant n_1} I(X_i \geqslant t), \quad K_2(t) = \sum_{i > n_1} I(X_i \geqslant t),$$

$$\omega(t) = \int_t^\tau \frac{A(s)h(s)K_1(s)K_2(s)}{(1 + R(s))(1 + \hat{R}(s;\boldsymbol{\beta}))K(s)}(h(s)\mathrm{e}^{-\beta_2} - 1)\frac{\mathrm{d}R(s)}{\hat{P}(s;\boldsymbol{\beta})}.$$

From Theorem A2 of Yang and Prentice (2005) and some algebra,

$$Q(\boldsymbol{\beta}) = \left(\sum_{i \leqslant n_1}\int_0^\tau \mu_1\,\mathrm{d}M_i + \sum_{i > n_1}\int_0^\tau \mu_2\,\mathrm{d}M_i\right) + \mathrm{o}_p(1),$$

where

$$\mu_1(t) = -\frac{A(t)K_2(t)h(t)}{K(t)} + \frac{\hat{P}_-(t;\boldsymbol{\beta})(1+\hat{R}(t;\beta))}{K}\omega(t),$$

$$\mu_2(t) = A(t)\frac{K_1(t)}{K(t)} + \frac{\hat{P}_-(t;\boldsymbol{\beta})(\mathrm{e}^{-\beta_1} + \mathrm{e}^{-\beta_2}\hat{R}(t;\boldsymbol{\beta}))}{K(t)}\omega(t), \tag{2.4}$$

$$M_i(t) = \delta_i I(X_i \leqslant t) - \int_0^t I(X_i \geqslant s)\frac{\mathrm{d}R(s)}{\mathrm{e}^{-\beta_1 Z_i} + \mathrm{e}^{-\beta_2 Z_i}R(s)}, \quad i = 1, \ldots, n.$$

Now for $\hat{R}(t;\hat{\boldsymbol{\beta}})$, from Lemma A3 in Yang and Prentice (2005) and some algebra,

$$\sqrt{n}(\hat{R}(t;\boldsymbol{\beta}) - R(t)) = \frac{1}{\sqrt{n}\hat{P}(t;\boldsymbol{\beta})}\left(\sum_{i\leqslant n_1}\int_0^t \nu_1\,\mathrm{d}M_i + \sum_{i>n_1}\int_0^t \nu_2\,\mathrm{d}M_i\right), \tag{2.5}$$

where

$$\nu_1(t) = \frac{n\hat{P}_-(t;\boldsymbol{\beta})}{K(t)}(1+R(t)), \quad \nu_2(t) = \frac{n\hat{P}_-(t;\boldsymbol{\beta})}{K(t)}(\mathrm{e}^{-\beta_1} + \mathrm{e}^{-\beta_2}R(t)).$$

Let

$$D(t;\boldsymbol{\beta}) = \frac{\partial\hat{R}(t;\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}, \quad U = \left(-\frac{1}{n}\frac{\partial Q(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}\right)^{-1},$$

$$B(t) = h(t)A(t) + \frac{\mathrm{e}^{-\beta_1} - \mathrm{e}^{-\beta_2}}{(\mathrm{e}^{-\beta_1} + \mathrm{e}^{-\beta_2}R(t))^2}D(t;\boldsymbol{\beta}),$$

$$C(t) = \frac{\mathrm{e}^{-\beta_1} - \mathrm{e}^{-\beta_2}}{(\mathrm{e}^{-\beta_1} + \mathrm{e}^{-\beta_2}R(t))^2}\frac{1}{\hat{P}(t;\boldsymbol{\beta})}.$$

For $t \leqslant \tau$, define the process

$$\tilde{W}_n(t) = \frac{B^T(t)U}{\sqrt{n}}\left(\sum_{i\leqslant n_1}\int_0^\tau \mu_1\,\mathrm{d}M_i + \sum_{i>n_1}\int_0^\tau \mu_2\,\mathrm{d}M_i\right)$$
$$+\frac{C(t)}{\sqrt{n}}\left(\sum_{i\leqslant n_1}\int_0^t \nu_1\,\mathrm{d}M_i + \sum_{i>n_1}\int_0^t \nu_2\,\mathrm{d}M_i\right). \tag{2.6}$$

With the representations for $Q(\boldsymbol{\beta})$ and $\sqrt{n}(\hat{R}(t;\boldsymbol{\beta}) - R(t))$, in Appendix B of the Supplementary Material available at *Biostatistics* online, we show that W_n is asymptotically equivalent to $\tilde{W}_n$ which converges weakly to a zero-mean Gaussian process W^*. The weak convergence of W_n thus follows. The limiting covariance function $\sigma(s,t)$ of W^* involves the derivative $D(t;\boldsymbol{\beta})$ and the derivative matrix in U. Although analytic forms of these derivatives are available, they are quite complicated and cumbersome. Here, we approximate them by numerical derivatives. For the functions $B(t)$, $C(t)$, $\mu_1(t)$, $\mu_2(t)$, $\nu_1(t)$, and $\nu_2(t)$, define corresponding $\hat{B}(t), \hat{C}(t), \ldots$, by replacing $\boldsymbol{\beta}$ with $\hat{\boldsymbol{\beta}}$, $R(t)$ with $\hat{R}(t;\hat{\boldsymbol{\beta}})$ and $D(t;\boldsymbol{\beta})$ with the numerical derivatives. Similarly, let $\hat{U}$ be the numerical approximation of U. Simulation studies show

that the results are fairly stable with respect to the choice of the jump size in the numerical derivatives, and that the choice of $n^{-1/2}$ works well. With these approximations, we can estimate $\sigma(s,t)$, $s \leqslant t \leqslant \tau$, by

$$\begin{aligned}\hat{\sigma}(s,t) = \hat{B}^T(s)\hat{U}\Bigg\{\int_0^\tau \frac{1}{n(1+\hat{R}(w;\hat{\boldsymbol{\beta}}))}[\hat{\mu_1}(w)\hat{\mu_1}^T(w)K_1(w) \\ + \hat{\mu_2}(w)\hat{\mu_2}^T(w)K_2(w)\hat{h}(w)]\hat{R}(\mathrm{d}w,\hat{\boldsymbol{\beta}})\Bigg\}\hat{U}^T\hat{B}(t) \\ + \hat{C}(s)\hat{C}(t)\int_0^s \frac{1}{n(1+\hat{R}(w;\hat{\boldsymbol{\beta}}))}[\hat{\nu_1}^2(w)K_1(w) \\ + \hat{\nu_2}^2(w)K_2(w)\hat{h}(w)]\hat{R}(\mathrm{d}w,\hat{\boldsymbol{\beta}}) \\ + \hat{C}(t)\hat{B}^T(s)\hat{U}\int_0^t \frac{1}{n(1+\hat{R}(w;\hat{\boldsymbol{\beta}}))}[\hat{\mu_1}(w)\hat{\nu_1}(w)K_1(w) \\ + \hat{\mu_2}(w)\hat{\nu_2}(w)K_2(w)\hat{h}(w)]\hat{R}(\mathrm{d}w,\hat{\boldsymbol{\beta}}) \\ + \hat{C}(s)\hat{B}^T(t)\hat{U}\int_0^s \frac{1}{n(1+\hat{R}(w;\hat{\boldsymbol{\beta}}))}[\hat{\mu_1}(w)\hat{\nu_1}(w)K_1(w) \\ + \hat{\mu_2}(w)\hat{\nu_2}(w)K_2(w)\hat{h}(w)]\hat{R}(\mathrm{d}w,\hat{\boldsymbol{\beta}}). \qquad (2.7)\end{aligned}$$

Now for a fixed $t_0 \leqslant \tau$, from the above results, confidence intervals for $h(t_0)$ can be obtained from the asymptotic normality of $\hat{h}(t_0)$ and the estimated variance $\hat{\sigma}(t_0,t_0)$. The usual logarithm transformation results in the asymptotic $100(1-\alpha)\%$ confidence interval $\hat{h}(t_0)\exp\left(\mp z_{\alpha/2}\frac{\sqrt{\hat{\sigma}(t_0,t_0)}}{\sqrt{n}\hat{h}(t_0)}\right)$, where $z_{\alpha/2}$ is the $100(1-\alpha/2)\%$ percentile of the standard normal distribution.

3. Simultaneous confidence bands

To make simultaneous inference on $h(t)$ over a time interval $I=[a,b]\subset[0,\tau]$, consider

$$V_n(t) = \sqrt{n}\frac{\hat{h}(t)}{s(t)}(\ln(\hat{h}(t)) - \ln(h(t))),$$

where $s(t)$ converges in probability, uniformly in t over I, to a bounded function $s^*(t) > 0$. From the weak convergence of W_n to W^* and the functional delta method, we have the weak convergence of V_n to W^*/s^*. Thus, if c_α is the upper αth percentile of $\sup_{t\in I}|W^*/s^*|$, an asymptotic $100(1-\alpha)\%$ simultaneous confidence band for $h(t)$, $t\in I$, can be obtained as

$$\hat{h}(t)\exp\left(\mp\frac{c_\alpha s(t)}{\sqrt{n}\hat{h}(t)}\right).$$

It is difficult to obtain c_α analytically. One obvious alternative would be the bootstrapping method. However, it is very time-consuming and results in lower than nominal coverage probabilities in some simulation studies. Lin *and others* (1993) used a normal resampling approximation to simulate the asymptotic

distribution of sums of martingale residuals for checking the Cox regression model. The normal resampling approach reduces computing time significantly and has become a standard method. It has been used in many works, including Lin *and others* (1994), Cheng *and others* (1997), Gilbert *and others* (2002), Tian *and others* (2005), and Peng and Huang (2007). We will modify this approach for our problem here.

For $t \leqslant \tau$, define the process

$$\begin{aligned}\hat{W}_n(t) &= \frac{\hat{B}^T(t)\hat{U}}{\sqrt{n}}\left(\sum_{i\leqslant n_1}\int_0^\tau \hat{\mu_1}d(\epsilon_i N_i) + \sum_{i>n_1}\int_0^\tau \hat{\mu_2}d(\epsilon_i N_i)\right)\\ &\quad + \frac{\hat{C}(t)}{\sqrt{n}}\left(\sum_{i\leqslant n_1}\int_0^t \hat{\nu_1}d(\epsilon_i N_i) + \sum_{i>n_1}\int_0^t \hat{\nu_2}d(\epsilon_i N_i)\right)\\ &= \frac{\hat{B}^T(t)\hat{U}}{\sqrt{n}}\left(\sum_{i\leqslant n_1}\epsilon_i\delta_i\hat{\mu_1}(X_i)I(X_i\leqslant\tau) + \sum_{i>n_1}\epsilon_i\delta_i\hat{\mu_2}(X_i)I(X_i\leqslant\tau)\right)\\ &\quad + \frac{\hat{C}(t)}{\sqrt{n}}\left(\sum_{i\leqslant n_1}\epsilon_i\delta_i\hat{\nu_1}(X_i)I(X_i\leqslant t) + \sum_{i>n_1}\epsilon_i\delta_i\hat{\nu_2}(X_i)I(X_i\leqslant t)\right), \qquad (3.1)\end{aligned}$$

where $\epsilon_i, i = 1, \ldots, n$, are independent variables that are also independent from the data. Furthermore, these variables have mean zero and variance converging to one as $n \to \infty$. In the normal resampling approach mentioned above, the ϵ_i's are the standard normal variables. However, the standard normal variables often result in lower coverage probabilities in various simulation studies. Thus, with moderate sized samples, we need to make some adjustment.

Conditional on $(X_i, \delta_i, Z_i), i = 1, \ldots, n$, $\hat{W}_n$ is a sum of n independent variables at each time point. In Appendix B of the Supplementary Material available at *Biostatistics* online, we show that $\hat{W}_n$ given the data converges weakly to W^*. It follows that $\sup_{t\in I}|\hat{W}/s|$ given the data converges in distribution to $\sup_{t\in I}|W^*/s^*|$. Therefore, c_α can be estimated empirically from a large number of realizations of the conditional distribution of $\sup_{t\in I}|\hat{W}/s|$ given the data.

Several choices of the weight s arise from recommendations in the literature for confidence bands of the survivor function and the cumulative hazard function in the one sample case. The choice $s(t) = \sqrt{\hat{\sigma}(t,t)}$ results in equal precision bands (Nair, 1984), which differ from pointwise confidence intervals in that c_α replaces $z_{\alpha/2}$. The choice $s(t) = 1 + \hat{\sigma}(t,t)$ results in the Hall–Wellner type bands recommended by Bie *and others* (1987), which often have narrower widths in the middle of data range and wider widths near the extremes of data range (Lin *and others*, 1994). One could also choose $s(t) = \hat{h}(t)$. This choice does not involve $\hat{\sigma}(t,t)$ and hence is easier to implement. It may be adequate when $\hat{\sigma}(t,t)$ only varies mildly over time.

Let $a \in (0, \tau)$ and define the average hazard ratio, over $[a, t]$,

$$\bar{h}(t) = \frac{1}{t-a}\int_a^t h(s)\mathrm{d}s, \quad a < t < \tau.$$

Note that the average hazard ratio involves an integral of the hazard ratio rather than a ratio of integrated hazards. It provides a measure for the cumulative treatment effect over a time interval to augment the temporal effect display from the hazard ratio estimates. It can be estimated by

$$\hat{\bar{h}}(t) = \frac{1}{t-a}\int_a^t \hat{h}(s)\mathrm{d}s, \quad a < t < \tau.$$

To obtain simultaneous confidence bands for the average hazard ratio, let

$$\bar{W}_n(t) = \sqrt{n}(\widehat{\bar{h}}(t) - \bar{h}(t)), \quad a < t < \tau.$$

In Appendix B of the Supplementary Material available at *Biostatistics* online, we show that $\bar{W}_n(t)$ converges weakly to the zero-mean Gaussian process $\int_a^t W^*(s)\mathrm{d}s/(t-a)$. Also, $\widehat{\bar{h}}(t)$ behaves more stably than $\hat{h}(t)$ and its covariance function is comparatively insensitive to the choice of weight function. Hence, for simplicity, we consider only the process

$$\bar{V}_n(t) = \sqrt{n}(\ln(\widehat{\bar{h}}(t)) - \ln(\bar{h}(t))).$$

From the functional delta method, it follows that $\bar{V}_n(t)$ converges weakly to $\int_a^t W^*(s)\mathrm{d}s/((t-a)\bar{h}(t))$. Thus, if $\bar{c}_\alpha$ is the upper αth percentile of $\sup_{t\in[a,b]} \left| \int_a^t W^*(s)\mathrm{d}s/((t-a)\bar{h}(t)) \right|$, an asymptotic $100(1-\alpha)\%$ simultaneous confidence band for $\bar{h}(t)$, $t \in I$, can be obtained as

$$\hat{h}(t) \exp\left(\mp \frac{\bar{c}_\alpha}{\sqrt{n}}\right).$$

To approximate the critical value $\bar{c}_\alpha$, again we use a resampling approximation. In Appendix B of the Supplementary Material available at *Biostatistics* online, the process $\int_a^t \hat{W}_n(s)\mathrm{d}s/(t-a)$ given the data is shown to converge weakly to $\int_a^t W^*(s)\mathrm{d}s/(t-a)$. From this and strong consistency of $\widehat{\bar{h}}(t)$, $\bar{c}_\alpha$ can be approximated empirically from a large number of realizations of the conditional distribution of $\sup_{t\in[a,b]} \left| \int_a^t \hat{W}_n(s)\mathrm{d}s/((t-a)\widehat{\bar{h}}(t)) \right|$ given the data.

4. Simulation studies

Without any finite-sample modifications, it was found that the empirical coverage probabilities of the proposed confidence bands for the hazard ratio were often lower than the nominal levels for small samples, especially with substantial censoring. In a series of simulation studies, we have gone through an extensive trial and error process to evaluate various modifications. In the end, we recommend that the left continuous versions of the integrands in (2.3) be used. Also, instead of $\hat{P}(t;\mathbf{b})$, we will use the asymptotically equivalent form $\exp\left(-\int_0^t \frac{H_2(\mathrm{d}s;\mathbf{b})}{K(s)}\right)$. In addition, it is best to restrict to the time range $[\inf\kappa, \sup\kappa]$, where κ is the set of observations at which the weight function $s(t)$ is less than or equal to the 90%th percentile of $s(t_i)$, $i = 1, \ldots, n$, with t_i's being the uncensored observations. This restriction is similar in spirit to the recommendations of Nair (1984) and Bie *and others* (1987), except we measure the extremeness of data by $s(t_i)$. For the hazard ratio and small to moderate n, we choose the ϵ_i's in (3.1) to be a multiple of the standard normal variables. We will use an *ad hoc* multiplier of $1 + 1/(2\sqrt{n})$ based on various simulations. For n equal to 400 or larger, the standard normal variables can be used. For the average hazard ratio, no such multiplier adjustment is necessary.

Next, we report the results from some representative simulation studies. Here and for the real data application in Section 5 later, τ was set to exclude the last-order statistic. All numerical computations were done in "Matlab." First, under the model of Yang and Prentice (2005), lifetime variables were generated with $R(t)$ chosen to yield the standard exponential distribution for the control group. The values of β were $(\log(0.9), \log(1.2))$ and $(\log(1.2), \log(0.8))$, representing $1/3$ increase or decrease over time from the initial hazard ratio, respectively. The censoring variables were independent and identically distributed with the log-normal distribution, where the normal distribution had mean c and standard deviation 0.5, with c chosen to achieve various censoring rates. The empirical coverage probabilities were obtained

from 1000 repetitions, and for each repetition, the critical values c_α and $\bar{c}_\alpha$ were calculated empirically from 1000 realizations of relevant conditional distributions. The results of these simulations are summarized in Table 1, where the equal precision bands, Hall–Wellner type bands and unweighted bands for the hazard ratio are denoted by EP, HW, and UW, respectively. Results for simultaneous confidence bands of the average hazard ratio are also included with the column header $\bar{h}$. From Table 1, the empirical coverage probabilities for the hazard ratio were mostly close to the nominal level. The empirical coverage probabilities for the average hazard ratio were mostly conservative. The conservative results were partially due to the finite-sample modifications intended for the hazard ratio. Those modifications improved the performance of the hazard ratio estimation procedure under some scenarios, while yielding conservatism in others, particularly for the more stable average hazard ratio estimator. The coverage probabilities for the equal precision bands overall were closer to the nominal level than other types of bands.

To check the robustness of the proposed procedures, we carried out various simulation studies with monotone hazard ratio not satisfying the model of Yang and Prentice (2005). For Table 2, the control group lifetime variables were standard exponential. The hazard ratio was linear from 0 to the 99th percentile of the standard exponential and continuous and constant afterward. The initial and end hazard ratios again were $(0.9, 1.2)$ and $(1.2, 0.8)$, respectively, and the censoring variables were the same as before. From Table 2, the results are similar to those in Table 1, with slight undercoverage under some scenarios.

Table 1. *Empirical coverage probabilities of the simultaneous confidence bands, for the hazard ratio (EP, HW, and UW) and the average hazard ratio ($\bar{h}$), under the model of Yang and Prentice* (2005), *based on* 1000 *repetitions*

Hazard ratio	Censoring rate (%)	$n_1 = n_2$	EP	HW	UW	$\bar{h}$
$0.9 \uparrow 1.2$	10	40	0.954	0.946	0.963	0.973
	30		0.952	0.946	0.961	0.970
	50		0.971	0.960	0.976	0.977
	75		0.967	0.966	0.977	0.964
	10	80	0.955	0.957	0.959	0.963
	30		0.947	0.940	0.955	0.962
	50		0.955	0.943	0.956	0.965
	75		0.967	0.979	0.979	0.976
	10	160	0.960	0.966	0.966	0.977
	30		0.954	0.950	0.951	0.969
	50		0.941	0.937	0.940	0.964
	75		0.960	0.970	0.971	0.967
$1.2 \downarrow 0.8$	10	40	0.966	0.980	0.983	0.976
	30		0.936	0.948	0.967	0.980
	50		0.943	0.948	0.954	0.967
	75		0.956	0.959	0.964	0.966
	10	80	0.959	0.974	0.975	0.971
	30		0.926	0.946	0.945	0.964
	50		0.930	0.946	0.939	0.953
	75		0.959	0.966	0.968	0.965
	10	160	0.957	0.973	0.963	0.967
	30		0.949	0.965	0.945	0.968
	50		0.944	0.962	0.947	0.970
	75		0.951	0.957	0.954	0.961

Table 2. *Empirical coverage probabilities of the simultaneous confidence bands, for the hazard ratio (EP, HW, and UW) and the average hazard ratio ($\bar{h}$), under a monotone hazard ratio model not satisfying the model of Yang and Prentice* (2005)*, based on* 1000 *repetitions*

Hazard ratio	Censoring rate (%)	$n_1 = n_2$	EP	HW	UW	$\bar{h}$
0.9 ↑ 1.2	10	40	0.955	0.957	0.954	0.973
	30		0.965	0.952	0.964	0.976
	50		0.945	0.941	0.960	0.962
	75		0.971	0.972	0.9754	0.970
	10	80	0.959	0.963	0.958	0.983
	30		0.935	0.943	0.940	0.968
	50		0.938	0.943	0.937	0.956
	75		0.956	0.958	0.965	0.955
	10	160	0.963	0.964	0.950	0.974
	30		0.952	0.949	0.937	0.966
	50		0.940	0.935	0.920	0.960
	75		0.957	0.969	0.976	0.971
1.2 ↓ 0.8	10	40	0.976	0.969	0.975	0.982
	30		0.952	0.956	0.967	0.970
	50		0.955	0.963	0.966	0.961
	75		0.966	0.970	0.975	0.967
	10	80	0.965	0.967	0.969	0.975
	30		0.954	0.967	0.969	0.972
	50		0.941	0.948	0.960	0.968
	75		0.965	0.965	0.971	0.973
	10	160	0.969	0.977	0.960	0.976
	30		0.963	0.967	0.969	0.967
	50		0.937	0.943	0.941	0.963
	75		0.955	0.963	0.970	0.970

5. Application

Let us illustrate the proposed methods with data from the WHI randomized controlled trial of combined (estrogen plus progestin) postmenopausal hormone therapy, which reported an elevated coronary heart disease risk and overall unfavorable health benefits versus risks over a 5.6-year study period (Writing Group For the Women's Health Initiative Investigators, 2002; Manson *and others*, 2003). Few research reports have stimulated as much public response, since preceding observational research literature suggested a 40–50% reduction in coronary heart disease incidence among women taking postmenopausal hormone therapy. Analysis of the WHI observational study shows a similar discrepancy with the WHI clinical trial for each of coronary heart disease, stroke, and venous thromboembolism. The discrepancy is partially explained by confounding in the observational study. A remaining source of discrepancy between the clinical trial and the observational study is elucidated by recognizing a dependence of the hazard ratio on the therapy duration (e.g. Prentice *and others*, 2005). Here, we look at the time to coronary heart disease in the WHI clinical trial, which included 16 608 postmenopausal women initially in the age range of 50–79 with uterus ($n_1 = 8102$). There were 188 and 147 events observed in the treatment and control group, respectively, implying about 98% censoring, primarily by the trial stopping time. Fitting model (2.1) to this data set, we get $\hat{\boldsymbol{\beta}} = (0.65, -3.63)^T$. Due to heavy censoring, the value 0.03 of $\exp(\hat{\boldsymbol{\beta}}_2)$ cannot be interpreted as the estimated long-term hazard ratio in the range of study follow-up times. The estimated hazard ratio function is needed for a more complete and accurate assessment of the treatment effect.

To examine model adequacy, we can use a residual plot that is similar to the method for the Cox regression model (Cox and Snell, 1968). Let Λ_C and Λ_T be the cumulative hazard functions of the 2 groups, respectively. Then $\Lambda_C(T_i),\ i \leqslant n_1,\ \Lambda_T(T_i),\ i > n_1$ are i.i.d. from the standard exponential distribution. Let $\hat{\Lambda}_C$ and $\hat{\Lambda}_T$ be the model-based estimator of Λ_C and Λ_T, respectively, and define the residuals $\hat{\Lambda}_C(X_i),\ i \leqslant n_1,\ \hat{\Lambda}_T(X_i),\ i > n_1$. If model (2.1) is correct, the residuals should behave like a censored sample from the standard exponential distribution. Thus, the Aalen–Nelson cumulative hazard estimator based on them should be close to the identity function. If there is noticeable deviation, then model (2.1) is questionable. Similarly, the residual plot can be obtained for the piecewise constant hazards ratio model used in Prentice *and others* (2005). Both residual plots, not shown here, suggest that the 2 models fit the data adequately, with similar residual behaviors.

The 95% pointwise confidence intervals and simultaneous confidence bands for the hazard ratio function are given in Figure 1. For comparison, the 95% confidence intervals for 0–2, 2–5, and >5 years from Prentice *and others* (2005) are included, over the median of uncensored data in each time interval. Compared with the piecewise constant hazards ratio model, the confidence bands do not depend on partitioning of the data range and provide more continuously changing display of the treatment effect. The confidence bands are generally in agreement with the results from Prentice *and others* (2005). The UW band is wider than the other 2 bands most of the time. The HW band is the narrowest in the middle section but is quite wide at the beginning. Both the EP band and the HW band give narrower intervals for the middle portion of the data range than the piecewise Cox model. Near the end of the data range, all 3 bands have about the same width as the confidence interval from Prentice *and others* (2005). Overall the EP band matches most closely with the results for the piecewise constant hazards ratio model. The width of the EP band is

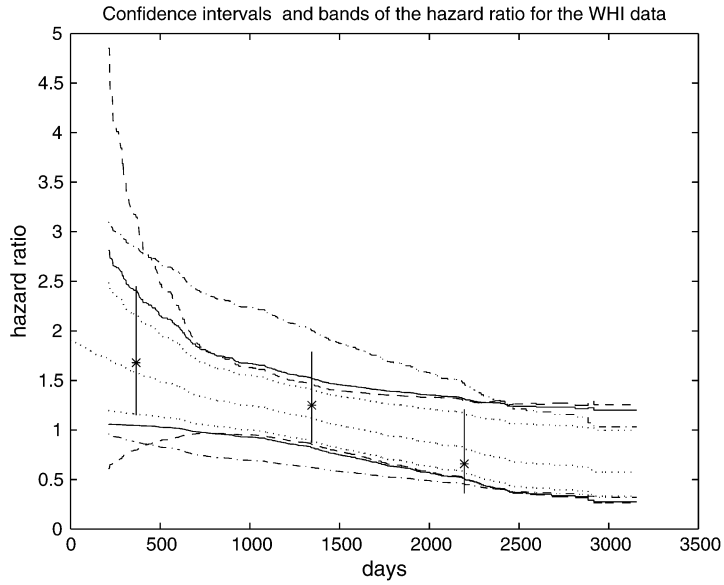

Fig. 1. The 95% pointwise confidence intervals and simultaneous confidence bands of the hazard ratio function for the WHI data: Solid lines—equal precision confidence band; dashed lines— Hall–Wellner type confidence band; dash dotted lines— unweighted confidence band; outside dotted lines—pointwise confidence limits; and central dotted line—the estimated hazard ratio function; vertical segments—95% confidence intervals for the hazard ratios in 0–2, 2–5, and >5 years intervals from Prentice *and others* (2005), over the median of uncensored data in each time interval, with "$*$" indicating the point estimates.

less than or equal to the piecewise model–based confidence intervals for most of the data range, except at the beginning. Note that the constant function 1 is not excluded in the HW and UW bands. In comparison, the EP band stays above 1 for about the first 600 days. From Prentice *and others* (2005), the confidence interval for $0 - 2$yr excludes 1, indicating an elevation in coronary heart disease risk for the treatment early on. For this data set, the standard error of the estimated hazard ratio begins at 0.43, quickly comes down to below 0.20 at 600 days and stays below 0.20 for the rest of data range. Since the UW band does not take the variance into account and the HW band emphasizes the middle range, the elevated standard error at early follow-up times likely explains the discrepancy among the results. Compared with the original analysis that showed an overall difference between the 2 groups, the results here and those from Prentice *and others* (2005) give more detailed analysis on the dependence of the hazard ratio on time and help explaining the discrepancy between the results of the WHI clinical trial and preceding observational research, much of which involved cohorts where women could be enrolled some years after initiating hormone therapy.

For the average hazard ratio function, the estimator and the 95% simultaneous confidence band are given in Figure 2. The standard error of the estimated average hazard ratio varies more mildly over time, and both the estimated average hazard ratio and the confidence band are changing much more smoothly compared with the results for the hazard ratio in Figure 1. Note that the confidence band stays above 1 for $t < 700$ days. This is in agreement with the results of Prentice *and others* (2005).

To compare with the nonparametric approach, Figure 3 gives the estimated hazard ratio, the 95% pointwise confidence intervals and simultaneous confidence band of Gilbert *and others* (2002), based on the R programs from the author's site. The same scale as that in Figure 1 is used for comparison and results in truncation of some portion of the plot. The estimated hazard ratio suggests that the hazard ratio is reasonably monotonic. The nonparametric hazard ratio estimate is somewhat lower than the hazard ratio estimates in Figure 1 under either model (2.1) or the piecewise constant hazards ratio model. The confidence band is wider than those in Figure 1 for the beginning and later parts of the data range, reflecting

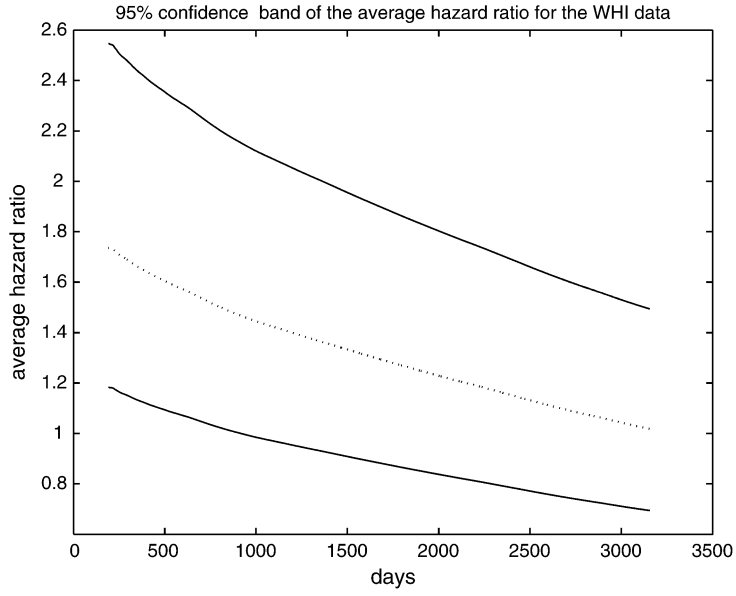

Fig. 2. The 95% simultaneous confidence band of the average hazard ratio function for the WHI data: dotted line—estimated average hazard ratio; and solid lines—95% simultaneous confidence band.

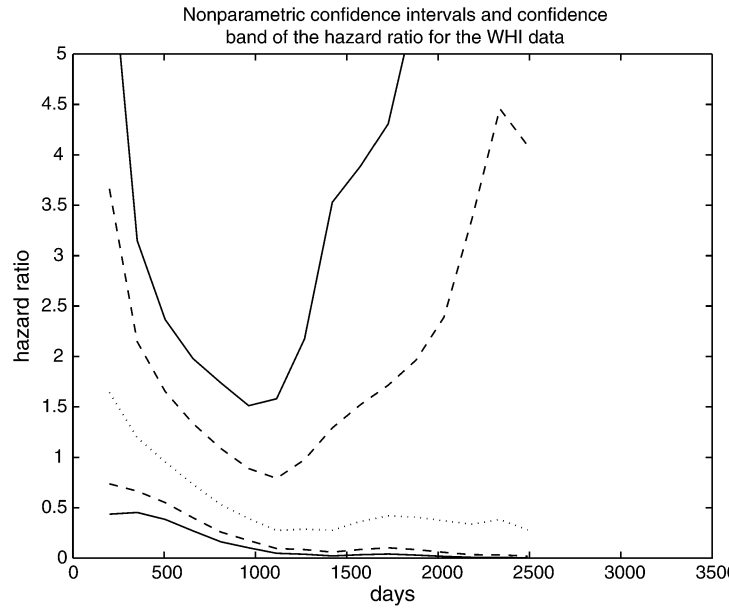

Fig. 3. Nonparametric 95% pointwise confidence intervals and simultaneous confidence band of the hazard ratio function for the WHI data: dotted line—estimated hazard ratio; solid lines—simultaneous confidence band; and dashed lines—pointwise confidence intervals.

the difficulty in making nonparametric inference on the hazard functions, especially with heavy censoring and in the tail region.

From the results here and additional numerical studies and real data applications, we find that for the hazard ratio, the EP bands are preferable if the interest is in the largest possible data range; if the interest is in part of the middle portion, then the HW bands are usually better. For the average hazard ratio, the simple confidence band proposed here works adequately, although could possibly be improved if more elaborate weights are used.

6. DISCUSSION

We have focused on the model of Yang and Prentice (2005) in deriving inference procedures for the hazard ratio function. Under this model, the hazard ratio involves the baseline survivor function, but not the baseline density function, a property shared by some other semiparametric models. Thus, inference on the hazard ratio may be easier and more reliable than approaches involving densities, such as those under the accelerated failure time model or the nonparametric approaches.

To assess the cumulative treatment effect, we have worked with the average hazard ratio function here, partly due to its close connection with the hazard ratio and its corresponding ready interpretation. Alternatively, the ratios $S_T(t)/S_C(t)$ and $(1 - S_T(t))/(1 - S_C(t))$ or the difference $S_T(t) - S_C(t)$, could be considered.

We expect that the model of Yang and Prentice (2005) can provide an adequate approximation for a wide range of applications. More rigorous model checking procedures would be useful to address model fit and robustness issues. Note that the form of this model is not closed under a relabeling of treatment and control groups, so its use may be more natural if there is a “no treatment ” or “standard treatment” control

group. It would be possible to study hazard ratio function estimation for larger classes of semiparametric models to incorporate an even wider range of time dependence of the hazard ratio, though there is a trade off between the model fit and increasing variance, as well as analysis cost. Also, while we have focused on the 2-sample comparison here, adjustment for covariates may be considered. These and other problems are worthy of further exploration.

SUPPLEMENTARY MATERIAL

Supplementary material is available at http://biostatistics.oxfordjournals.org.

ACKNOWLEDGMENTS

We thank Co-Editor Anastasios Tsiatis, a referee, and an associate editor for helpful comments and suggestions. *Conflict of Interest:* None declared.

FUNDING

National Institutes of Health (CA 53996 to Ross L. Prentice).

REFERENCES

ABRAHAMOWICZ, M. AND MACKENZIE, T. A. (2007). Joint estimation of time-dependent and non-linear effects of continuous covariates on survival. *Statistics in Medicine* **26**, 392–408.

BIE, O., BORGAN, Ø. AND LIESTØL, K. (1987). Confidence intervals and confidence bands for the cumulative hazard rate function and their small-sample properties. *Scandinavian Journal of Statistics* **14**, 221–233.

CAI, Z. AND SUN, Y. (2003). Local linear estimation for time-dependent coefficients in Coxs regression models. *Scandinavian Journal of Statistics* **30**, 93–111.

CHEN, Y. Q. AND CHENG, S. (2006). Linear life expectancy regression with censored data. *Biometrika* **93**, 303–313.

CHEN, Y. Q. AND WANG, M. C. (2000). Analysis of accelerated hazards models. *Journal of the American Statistical Association* **95**, 608–618.

CHENG. S. C., WEI, L. J. AND YING, Z. (1997). Predicting survival probabilities with semiparametric transformation models. *Journal of the American Statistical Association* **92**, 227–235.

COX, D. R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society, Series B* **34**, 187–220.

COX, D. R. (1975). Partial likelihood. *Biometrika* **62**, 269–276.

COX, D. R. AND SNELL, E. J. (1968). A general definition of residuals (with discussion). *Journal of the Royal Statistical Society, Series B* **30**, 248–275.

GILBERT, P. B., WEI, L. J., KOSOROK, M. R. AND CLEMENS, J. D. (2002). Simultaneous inferences on the contrast of two hazard functions with censored observations. *Biometrics* **58**, 773–780.

GRAY, R. J. (1992). Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis. *Journal of the American Statistical Association* **87**, 942–951.

HSIEH, F. (1996). A transformation model for two survival curves: an empirical process approach. *Biometrika* **83**, 519–528.

KOOPERBERG, C., STONE, C. J. AND TRUONG, Y. K. (1995). Hazard regression. *Journal of the American Statistical Association* **90**, 78–94.

LIN, D. Y., FLEMING, T. R. AND WEI, L. J. (1994). Confidence bands for survival curves under the proportional hazards model. *Biometrika* **81**, 73–81.

LIN, D. Y., WEI, L. J. AND YING, Z. (1993). Checking the Cox model with cumulative sums of martingale-based residuals. *Biometrika* **80**, 557–572.

MANSON, J. E., HSIA, J., JOHNSON, K. C., ROSSOUW, J. E., ASSAF, A. R., LASSER, N. L., TREVISAN, M., BLACK, H. R., HECKBERT, S. R., DETRANO, R. *and others* (2003). Estrogen plus progestin and the risk of coronary heart disease. *The New England Journal of Medicine* **349**, 523–534.

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

PENG, L. AND HUANG, Y. (2007). Survival analysis with temporal covariate effects. *Biometrika* **94**, 719–733.

PRENTICE, R. L., LANGER, R., STEFANICK, M. L., HOWARD, B. V., PETTINGER, M., ANDERSON, G., BARAD, D., CURB, J. D., KOTCHEN, J., KULLER, L. *and others* (2005). Combined postmenopausal hormone therapy and cardiovascular disease: toward resolving the discrepancy between observational studies and the womens health initiative clinical trial. *American Journal of Epidemiology* **162**, 404–414.

TIAN, L., ZUCKER, D. AND WEI, L. J. (2005). On the Cox model with time-varying regression coefficients. *Journal of the American statistical Association* **100**, 172–183.

TONG, X., ZHU, C. AND SUN, J. (2007). Semiparametric regression analysis of two-sample current status data, with applications to tumorigenicity experiments. *The Canadian Journal of Statistics* **35**, 575–584.

TSODIKOV, A. (2002). Semi-parametric models of long- and short-term survival: an application to the analysis of breast cancer survival in Utah by age and state. *Statistics in Medicine* **21**, 895–920.

VAUPEL J. W., MANTON, K. G. AND STALLARD, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography* **16**, 439–454.

WRITING GROUP FOR THE WOMEN'S HEALTH INITIATIVE INVESTIGATORS (2002). Risks and benefits of estrogen plus progestin in healthy postmenopausal women: principal results from the women's health initiative randomized controlled trial. *Journal of the American Medical Association* **288**, 321–333.

YANG, S. AND PRENTICE, R. L. (2005). Semiparametric analysis of short-term and long-term hazard ratios with two-sample survival data. *Biometrika* **92**, 1–17.

ZENG, D. AND LIN, D. Y. (2007). Maximum likelihood estimation in semiparametric regression models with censored data. *Journal of the Royal Statistical Society, Series B* **69**, 507–564.

[*Received October 21, 2009; revised August 23, 2010; accepted for publication August 24, 2010*]