UW Biostatistics Working Paper Series

10-23-2013

Hypothesis Testing for an Extended Cox Model with Time-Varying Coefficients

Takumi Saegusa
University of Washington - Seattle Campus, tsaegusa@uw.edu

Chongzhi Di
Fred Hutchinson Cancer Research Center, cdi@fredhutch.org

Ying Qing Chen
Fred Hutchinson Cancer Research Center, yqchen@fhcrc.org

1. Introduction

The HIVNET 012 Study is a randomized clinical trial conducted by the HIV Prevention Trial Network (HPTN) between 1997 and 2001 (The HIVNET/HPTN Group, 2003). It showed an astounding prevention benefit for a short course nevirapine (NVP) versus zidovudine (AZT) among HIV infected pregnant women in Uganda: NVP was associated with a 41% reduction in relative risk of mother-to-child transmission (MTCT) of HIV-1 through to the age of 18 months (Jackson et al., 2003).

Besides the prevention benefit, it was also important for this study to assess whether or not the NVP would eventually improve the newborn's 18-month survival. As shown in Figure 1(a), Kaplan-Meier curves indicate that the NVP group appears to have better infant survival. However, a log-rank test does not show a desired statistical significance (p value: 0.147). Although such a lack of significance may be due to insufficient sample size, if the hazard ratio is not proportional, the power of the log-rank test may be compromised as well. We hence plot the log-log-transformed estimated cumulative hazard functions in Figure 1(b). As shown in the figure, the two Kaplan-Meier curves indeed appear to be closer as time progresses. This may suggest that the hazards ratio is not constant at any time.

[Figure 1 about here.]

In fact, for the HIVNET 012 Study, the NVP was only given once to the mothers at labor onset and once to the babies within 72 hours of birth, while AZT was given to the mothers from labor onset to delivery and to the babies twice daily for 7 days after birth. As a result, it was not expected that the NVP effect would necessarily sustain throughout the entire 18-month of follow-up. In addition, even if the babies were born uninfected, they could be infected after birth via breastmilk feeding. For example, the concentration of cells in breastmilk decreases over time while the ratio of HIV susceptible cells to total breastmilk cells increases over time. Therefore, the risk reduction of 18-month infant mortality may be unlikely to stay constant over time. If so, it is important to

develop appropriate statistics to test whether or not there is any infant survival benefit for NVP under the alternative of time-varying hazards ratio.

As alluded, the power of log-rank test, or the score test based on the partial likelihood for the Cox proportional hazards model (Cox, 1972), relies on the alternative assumption of proportional hazards. They may lose power if the assumption does not hold, as in the HIVNET 012 Study example. This paper aims to develop some flexible testing procedures when a treatment effect measured by the hazards ratio is potentially time-varying.

To be specific, we consider an extended Cox proportional hazards model as follows,

$$\lambda(t|X,Z) = \lambda_0(t)\exp\{X^{\intercal}\beta + Z\theta(t)\}. \tag{1}$$

Here, $\lambda(\cdot \mid X, Z)$ is the hazard function for the p-dimensional covariate vector $X \in \mathcal{R}^p$ and the treatment indicator $Z \in \mathcal{R}$, and $\lambda_0(\cdot)$ is the baseline hazard function. The superscript $^{\intercal}$ is for vector (matrix) transpose. Moreover, β is the time-independent regression parameter of the same p-dimension for X, and $\theta(\cdot)$ is the time-varying coefficient for Z that is assumed to be a smooth function of time. Apparently, when $\theta(t)$ is constant, model (1) reduces to the usual Cox model. Under model (1), we are interested in testing the null hypothesis if Z has effect at any time while adjusting for X, i.e., $\mathrm{H}_0 : \theta(t) = 0$ for any $t \geqslant 0$.

There have been several approaches in the statistical literature for hypothesis testing involving $\theta(\cdot)$ in (1). For example, **?** and **?** considered adaptively weighted log-rank tests, assuming parametric forms of $\theta(\cdot)$ to be polynomial functions. For nonparametric or semiparametric approaches, Gray (1994) applied B-spline bases to approximate $\theta(\cdot)$ with a careful manual choice of tuning parameters, such as the degrees of freedom and the number and location of knots. Nevertheless, the choice of tuning parameters would depend on the functional shape of true $\theta(\cdot)$ and was generally unknown; and different choices of tuning parameters could affect power considerably and lead to different p-values. Other approaches may resort to direct nonparametric estimation of $\theta(\cdot)$, as seen in O'Sullivan (1988), Hastie and Tibshirani (1990), Zucker and Karr (1990), Sleeper and Harrington (1990), Kooperberg et al. (1995), Brown et al. (2007), and references therein. Generally speaking, these works tend to be useful in understanding the overall shape of $\theta(\cdot)$. Although

asymptotic properties are developed for an estimated $\widehat{\theta}(t)$ at a specific time t, they are usually not intended for testing the global null $\mathrm{H}_0 : \theta(t) = 0$ for any $t \geqslant 0$.

The null hypothesis of interest is related to but fundamentally different from $\mathrm{H}_{0,PH} : \theta(t) = c$ for some constant $c \in \mathcal{R}$. In fact, $\mathrm{H}_{0,PH}$ is exactly equivalent to the proportional hazards assumption, with testing procedures including Pettitt and Bin Daud (1990), Gray (1994) and Lin et al. (2006). Specifically in Lin et al. (2006), $\theta(\cdot)$ was approximated by smoothing spline bases, and the authors proposed a score test for $\mathrm{H}_{0,PH}$ that does not involve tuning parameters. Nevertheless, we would like to emphasize that these approaches are proposed to test a different null hypothesis from the proposed test, and are thus not comparable with our work. The differences between our work and other previous literature are further clarified in Section 3.5.

In this article, we aim to develop proper testing methods specifically for the null hypothesis $\mathrm{H}_0 : \theta(t) = 0$ for any $t \geqslant 0$ under model (1), based on spline representation of the hazard ratio $\theta(t)$. The rest of the paper is organized as follows. In Section 3, we study the extended Cox model and derive the proposed statistics. Extensive Monte-Carlo simulations studies are presented in Section 4 with various choices of $\theta(t)$ to evaluate a finite sample properties of our score statistics. Their performances are compared with the log-rank statistic and Gray's statistics (Gray, 1994). We also apply our proposed statistics to compare the 18-month infant survival of the HIVNET 012 Study. We conclude in Section 5 by summarizing our results and discussing relevant issues and future directions.

2. The extended Cox model

Throughout the rest of this paper, we assume the extended Cox model as in (1). Our aim is to develop powerful and omnibus hypothesis testing procedures for $\mathrm{H}_0 : \theta(t) = 0$ for any $t \geqslant 0$. Apparently when $\theta(\cdot)$ is constant, model (1) reduces to the usual Cox proportional hazards model. Nevertheless, $\theta(t)$ can be quite flexible. It allows the hazard ratio between two treatment groups to change over time for any given X. Some properties regarding this model are summarized as follows:

Property. Denote the cumulative hazard function by $\Lambda(t) = \int_0^t \lambda(s)ds$. Let $\theta(0) = \theta_0$ and assume that $\lim_{t\to\infty} \theta(t) = \theta_1 < \infty$. Then

(1) $\lim_{t\to 0} \log\{\Lambda(t \mid X, Z)/\Lambda_0(t)\} = X^{\intercal}\beta + Z\theta_0$, and

(2) $\lim_{t\to\infty} \log\{\Lambda(t \mid X, Z)/\Lambda_0(t)\} = X^{\intercal}\beta + Z\theta_1$.

Proof of this property is straightforward. Based on this, when $\theta(\cdot)$ is monotone, θ_0 and θ_1 define the boundaries of the relative risk in cumulative hazard functions adjusted for X. For example, when $\theta(t) = 0.7\exp(-t)$, the hazard ratio for Z is 0.7 at $t = 0$, but gradually reduces to zero as time progresses. More examples of $\theta(\cdot)$ are shown in Figure 2. **?** and **?** considered special parametric submodels, where the shape $\theta(\cdot)$ is represented by a family of polynomial functions indexed by a few parameters, to account for early or late effects.

2.1 *Spline representations of* $\theta(t)$

We assume that the collected data consist of n independent and identically distributed (iid) copies of (Y, Δ, X, Z), where Y is the minimum of time to event T and censoring time C, i.e., $Y = \min(T, C)$, $\Delta = I(T \leqslant C)$ is the event indicator, X is the vector of covariates other than the treatment indicator, and Z is the treatment indicator, namely $\{(Y_i, \Delta_i, X_i, Z_i),\ i = 1, 2, \ldots, n\}$. Let $t_1^o \leqslant t_2^o \leqslant \cdots \leqslant t_r^o$ denote the ordered observed failure times, i.e., ordered statistics of $\{Y_i : i = 1, 2, \cdots, n, \text{ and } \Delta_i = 1\}$, where $r = \sum_{i=1}^n \Delta_i$ is the number of observed failure time points.

To model the time-varying treatment effect flexibly, we consider representing $\theta(t)$ by fixed knots B-splines or smoothing splines, i.e.,

$$\theta(t) = \theta_0 + \sum_{k=1}^{K} \theta_k B_k(t), \tag{2}$$

where $B_k(t)$'s form a set of basis functions. Note that the methods development below works for both B-spline or smoothing spline approaches, and our experience is that the performance of the two approaches are comparable as long as the number of knots are reasonably dense. If one is using fixed knots B-splines, the number of basis function K is fixed and depends on the number of knots

and the order of polynomials. For the smoothing spline approach, on the other hand, the number of basis functions depends on the sample size and the order of polynomials, i.e., $K = r + m - 1$.

Since the partial likelihood involves $\theta(t)$ evaluated at the observed failure time $t_1^o, \ldots, t_r^o$ only, we define $\gamma = \{\, \theta(t_1^o), \ldots, \theta(t_r^o) \,\}^T = \mathbf{1}\theta_0 + B\underline{\theta}$, where $\mathbf{1} \in \mathbb{R}^r$ is a vector whose elements are all 1, $\underline{\theta} = (\theta_1, \ldots, \theta_K)^T \in \Theta \subset \mathbb{R}^r$ and B is a $r \times K$ matrix whose (i, j) element is $B_i(t_j^o)$.

2.2 *Penalized likelihood*

The spline representation introduces K parameters θ_k's for the hazard ratio function to allow flexiblity. However, with a large degree of freedom, the model could overfit and the power to detect deviations from the null may be low. One strategy to avoid overfitting is to introduce a penalized partial likelihood function by penalizing the roughness of $\theta(t)$, e.g., in the form of $\int \{\theta'(t)\}^2 dt$. It can be shown that the penalty term is a quadratic function of $\underline{\theta}$,

$$\int \{\theta'(t)\}^2 dt = \int \left\{ \textstyle\sum_k \theta_k B_k'(t) \right\}^2 dt = \underline{\theta}^T \Sigma \underline{\theta},$$

where Σ is a $K \times K$ matrix whose (i, j) element is $\int B_i'(t) B_j'(t) dt$, which is fully determined by the choice of splines. Thus, the penalized partial log likelihood function is defined by

$$\begin{aligned} \ell\{\beta, \theta(\cdot), \tau\} &\equiv \ell_P\{\beta, \theta(\cdot)\} - \frac{1}{2\tau} \int \{\theta'(t)\}^2 dt \\ &= \ell_P(\beta, \theta_0, \underline{\theta}) - \frac{1}{2\tau} \underline{\theta}^T \Sigma \underline{\theta}, \end{aligned} \tag{3}$$

where τ is a tuning parameter that controls smoothness of $\theta(t)$, and ℓ_P is the partial likelihood corresponding to hazard ratio function $\theta(t)$,

$$\ell_P(\beta, \theta(\cdot)) = \sum_{i=1}^n \Delta_i \left[X_i^T \beta + S_i \theta(Y_i) - \log \left\{ \sum_{j=1}^n \exp\left\{ X_j^T \beta + S_j \theta(Y_i) \right\} I(Y_j \geqslant Y_i) \right\} \right].$$

Note that τ controls the level of smoothness of $\theta(t)$ and thus effective degree of freedom. When τ is small, the penalized partial likelihood encourages solutions that are close to the Cox proportional model with $df = 1$ for treatment effect. When τ is large, the effect of penalty is negligible and the model involves $K + 1$ parameters for the treatment effect, e.g., $df = K + 1$. Under this model, the null hypothesis can be represented as $H_0 : \theta_k = 0, \ , k \in \{0, 1, \cdots, K\}$. Gray (1994) used B-splines and studied asymptotic properties of Wald, score and likelihood ratio tests for fixed tuning

parameter τ. For the choice of tuning parameter, they suggested choosing a suitable degree of freedom ("df") and find τ to achieve the desired df. However, in practice, the choice of suitable df is subjective, and the power performance depends on the tuning parameter.

To construct tests that does not depend on tuning parameters, one can exploit the connection between the penalized splines and random effects models (**?**). Note that the second term of (3) is proportional to the logarithm of a multivariate normal density with mean zero and the covariance matrix $\tau\Sigma^{-1}$. Hence, one can treat $\underline{\theta}$ as "random effects," and integrate $\underline{\theta}$ out with respect to the multivariate normal density to obtain the marginal partial log likelihood

$$\ell_m(\beta, \theta_0, \tau) = \int_{\theta\in\Theta} \exp\left\{\ell_P(\beta, \theta_0, \underline{\theta})\right\} \exp\left(-\frac{1}{2\tau}\underline{\theta}^T\Sigma\underline{\theta}\right) d\underline{\theta}. \tag{4}$$

We use this marginal likelihood as our objective function to derive the score statistics. In fact, since τ is the variance component for the random effects, setting $\tau = 0$ shall lead to $\underline{\theta} = 0$. As a result, testing H_0 is equivalent to $H_0 : \theta_0 = 0, \tau = 0$. Note that this mixed effect model representation holds for both B-splines and smoothing splines.

Remark 1. Lin et al. (2006) considered a smoothing splines approach, where the log hazard ratio $\theta(t)$ can be represented by

$$\theta(t) = \theta_0 + \sum_{j=1}^{m-1} \delta_j\, t^j + \sum_{k=1}^{r} \theta_k R(t, t_k^o),$$

where δ_j's and θ_k's are spline coefficients, $\{1, t, \cdots, t^{m-1}\}$ is a set of basis functions for $(m-1)^{th}$ order polynomials, and $R(t,s) = \int_0^1 (t-u)_+^{m-1}(s-u)_+^{m-1}/\{(m-1)!\}^2 du$, where $x_+ = x \vee 0 = \max\{x, 0\}$. The choice of m in (2) depends on a pre-specified level of smoothness. If one choose $m = 1$, the time-varying coefficient (2) is simplified as $\theta(t) = \theta_0 + \sum_{k=1}^{r} \theta_k R(t, t_k^o)$, where $R(t,s) = \min\{t, s\}$. It can be shown that the matrix B is exactly the same as Σ in this case.

Remark 2. There are several key differences between Lin et al. (2006) and our work. First, they are interested in testing the proportional hazards assumption, $H_0^{PH} : \tau = 0$, which is a different null hypothesis. Thus, our approach and theirs are not directly comparable. Second, their null hypothesis involves the variance component parameter only, where ours involve in an additional fixed effects parameter θ_0. Statistically, as will be seen later, our work involves address challenges

to combine score tests for both θ_0 and τ, due to some non-standard properties for the variance component τ. Third, our approach here is not limited to the smoothing splines approach, and works for other choices of basis functions such as B-splines and others.

3. *Proposed test statistics*

The aim of this paper is to propose "omnibus" testing procedures for possily time-varying treatment effects, without making parametric assumptions on the shape of the hazard ratio. The idea is to combine evidence from both the average magnitude and shape of the hazard ratio function and construct test statistics that are powerful under both PH and various non-PH alternatives.

3.1 *A two-stage test*

In the literature, there were developments on hypothesis testing procedures for the proportional hazards assumption, e.g., Lin et al. (2006), denoted by T_{LZD}. If one aims to test treatment effect H_0 while accounting for potential non-proportionality, a natural strategy is to construct a two-stage procedure as follows (henceforth denoted T^{2stg}),

S1: apply the test T_{LZD} for the PH assumption. Reject H_0, if the p value is less than a pre-determined significance level α_1; otherwise, go to the second stage;

S2: apply the standard log-rank test for treatment effect. Reject H_0 if the p value is less than another pre-determined significance level α_2.

This procedure is a straightforward extension of the log-rank test and T_{LZD}. Note that S1 tests the proportional hazards assumption, while S2 tests treatment effect given the proportional hazards assumption is plausible. The overall null hypothesis is rejected, if either stage rejects the null. The overall type I error rate of this two-stage procedure depends on the correlation of two tests, but is bounded by $\alpha_1 + \alpha_2$ from above according to the Bonferroni inequality. The parameters α_1 and α_2 controls how much type I error was assigned to the two tests. In practice, one can choose $\alpha_1 = \alpha_2 = \alpha/2$ without prior information on the plausibility of proportional hazards, where α is

the targeted overall significance level. The performance of this simple two-stage procedure will be compared with the standard log-rank test as well as other proposed methods to be discussed below.

3.2 *Score statistics*

Next, we construct a few test statistics based by combining score statistics for θ_0 and τ. Taking the derivatives of (4) with respect to θ_0 and τ, one obtains

$$U_{\theta_0} = \left.\frac{\partial \ell_m(\beta, \theta_0, \tau)}{\partial \theta_0}\right|_{\beta=\widehat{\beta},\, \theta_0=0,\, \tau=0} = \mathbf{1}^T \frac{\partial \ell_P(\widehat{\beta}, \theta_0 = 0, \underline{\theta} = \mathbf{0})}{\partial \gamma},$$

and

$$\begin{aligned} U_\tau = \left.\frac{\partial \ell_m(\beta, \theta_0, \tau)}{\partial \tau}\right|_{\beta=\widehat{\beta},\, \theta_0=0,\, \tau=0} &= \frac{1}{2} \frac{\partial \ell_P(\widehat{\beta}, \theta_0 = 0, \underline{\theta} = \mathbf{0})}{\partial \gamma^T} B\Sigma^{-1}B^T \frac{\partial \ell_P(\widehat{\beta}, \theta_0 = 0, \underline{\theta} = \mathbf{0})}{\partial \gamma} \\ &\quad + \frac{1}{2}\mathrm{tr}\left[\frac{\partial^2 \ell_P(\widehat{\beta}, \theta_0 = 0, \underline{\theta} = \mathbf{0})}{\partial \gamma \partial \gamma^T} B\Sigma^{-1}B^T\right], \end{aligned} \tag{5}$$

where $\widehat{\beta}$ is the maximum partial likelihood estimate of the Cox model without treatment effect, i.e., the maximizer of the partial likelihood $\ell_P(\beta, \theta_0 = 0, \underline{\theta} = \mathbf{0})$. First, we look at the derivative with respect to θ_0. Denote $S(\beta, \theta_0, \underline{\theta}) = (\partial/\partial\gamma)\ell_P(\beta, \theta_0, \underline{\theta})$, the k^{th} element of $S(\widehat{\beta}, 0, 0)$ is then $\sum_i I(Y_i = t_k^o)\Delta_i\{S_i - \sum_j S_j \exp(\widehat{\beta}^T X_j) I(Y_j \geqslant t_k^o)/\sum_j \exp(\widehat{\beta}^T X_j) I(Y_j \geqslant t_k^o)\}$. It can be shown that the covariance matrix of $S(\beta, \theta_0, \underline{\theta})$ is given by $V = I_{\gamma\gamma} - I_{\gamma\beta} I_{\beta\beta}^{-1} I_{\beta\gamma}$, where the Fisher information matrices are evalued under the true parameter values. Thus, $\mathbf{1}^T S(\widehat{\beta}, 0, 0)$ is in fact the usual partial likelihood score function evaluated under the null. The standard score test statistic for θ_0 can be written as

$$T_{LR} = \{\, \mathbf{1}^T S(\widehat{\beta}, 0, 0)\,\}^2 / \mathrm{var}\{\mathbf{1}^T S(\widehat{\beta}, 0, 0)\} = {I_0}^{-1} S(\widehat{\beta}, 0, 0)^T\, \mathbf{1}\mathbf{1}^T\, S(\widehat{\beta}, 0, 0),$$

which is a quadratic form of $S(\widehat{\beta}, 0, 0)$ and converges to χ_1^2 due to $rank(\mathbf{1}\mathbf{1}^T) = 1$. Note here $I_0 = \mathrm{var}\{\mathbf{1}^T S(\widehat{\beta}, 0, 0)\} = \mathbf{1}^T V \mathbf{1}$ is exactly the efficient Fisher information for θ_0 under the Cox model. Thus, the standardized score test statistic T_{LR} converges to a χ_1^2 distribution under the null hypothesis. Next, we look at the derivative with respect to τ. It has been shown that the variation of second term of (5) is negligible relative to the first term (Lin et al., 2006). The first term of (5) is a quadratic form of $S(\widehat{\beta}, 0, 0)$. According to quadratic form theory, its limiting distribution is

weighted sum of χ^2's, with weights determined by the eigenvalues of Σ. This suggests that the asymptotic behavior of the score function for τ is non-standard.

Remark 3. For the smoothing splines approach with first order polynomials (Lin et al., 2006), the term $B\Sigma^{-1}B^T$ simplifies to Σ as $B = \Sigma = B^T$ in this case. They evaluated the score for τ at $(\beta = \widetilde{\beta}, \theta_0 = \widetilde{\theta}_0, \underline{\theta} = 0)$ instead, where $\widetilde{\beta}$ and $\widetilde{\theta}_0$ are their maximum likelihood estimates under the Cox model, because they are testing a different null hypothesis H_0^{PH}.

3.3 *Combine scores for θ_0 and τ*

To test $H_0 : \theta_0 = 0, \tau = 0$, since the null hypothesis involves both θ_0 and τ, one needs to combine score functions with respect to θ_0 and τ (denoted as U_{θ_0} and U_τ, respectively), which reflect information from the average magnitude and shape of the hazard ratio function, respectively. Under regularity conditions, score functions for multiple parameters follows multivariate normal distribution asymptotically, and the standard approach is contructing linear combination of U_{θ_0} and U_τ weighted by the square root of the joint Fisher information matrix. However, this procedure does not work here, due to non-standard properties in testing variance components.

As discussed above, the parameter $\tau = 0$ is on the boundary of its parameter space under the null. The dominating term of the score function with respect to τ converges to a weighted sum of χ_1^2's, instead of a Gaussian distribution. To overcome the challenges, we propose a few methods of combing two score functions to construct test statistics. U_{θ_0} and U_τ are linear and quadratic forms of $S(\widehat{\beta}, 0, 0)$, respectively. We propose a few combinations that are quadratic forms of $S(\widehat{\beta}, 0, 0)$, whose asymptotic distribution can be derived conveniently.

The first test statistics T_1 is constructed by taking the sum of U_τ and $U_{\theta_0}^2$,

$$\begin{aligned} T_1 &\equiv S^T(\widehat{\beta}, 0, 0)B\Sigma^{-1}B^T S(\widehat{\beta}, 0, 0) + (\widehat{I}_0)^{-1} S^T(\widehat{\beta}, 0, 0)\mathbf{1}\mathbf{1}^T S(\widehat{\beta}, 0, 0) \\ &= S^T(\widehat{\beta}, 0, 0)\left\{B\Sigma^{-1}B^T + (\widehat{I}_0)^{-1}\mathbf{1}\mathbf{1}^T\right\} S(\widehat{\beta}, 0, 0), \end{aligned} \tag{6}$$

where $\widehat{I}_0$ is an estimate of the efficient information for θ_0 in the usual Cox model. Note that T_1 is also a quadratic form of $S(\widehat{\beta}, 0, 0)$, and thus its limiting distribution can be easily calculated. It converges to a weighted sum of χ_1^2's, with weights determined by eigenvalues of the matrix

$M_1 = B\Sigma^{-1}B^T + (\widehat{I}_0)^{-1}\mathbf{1}\mathbf{1}^T$. Using the Satterthwaite approximation, the limiting distribution can be further simplified. Our score statistic rejects the null hypothesis H_0 at the nominal level α if $T_1 \geqslant k_1\chi^2_{\alpha,v_1}$ where $\chi^2_{\alpha,v}$ is a $100(1-\alpha)$ percentile of the χ^2_v random variable with degree of freedom v. Here, $k_1 = \text{tr}(M_1VM_1V)/\text{tr}(M_1V)$, and $v_1 = \{\text{tr}(M_1V)\}^2/\text{tr}(M_1VM_1V)$, where $V = \partial^2\ell_P\{\widehat{\beta}, (\theta_0, \underline{\theta}^T) = 0\}/\partial\gamma\partial\gamma^T$. The details on deriving the limiting distribution and its approximations are discussed in Appendix A.

The test statistic T_1 is a sum of two quadratic forms. However, these two parts are not independent, and taking the sum directly may not be optimal in terms of power, with potential power loss depending their correlation. We propose to remove the projection of S_τ on S_{θ_0}, so that the modified score statistics for τ (denoted by subscript "mPH", i.e., modified test for PH) is asymptotically independent to the score statistics for θ_0. Let

$$T_{mPH} \equiv S^T(\widehat{\beta}, 0, 0)W^TB\Sigma^{-1}B^TWS(\widehat{\beta}, 0, 0), \tag{7}$$

where $W = I_{r\times r} - V\mathbf{1}\mathbf{1}^T/\{\mathbf{1}^TV\mathbf{1}\}$. The matrix W is constructed so that $WS(\widehat{\beta}, 0, 0)$ and $\mathbf{1}^TS(\widehat{\beta}, 0, 0)$ are asymptotically independent (see Appendix C). Note that T_{mPH} is also a quadratic form of $S(\widehat{\beta}, 0, 0)$ that reflects evidence on proportionality, on the direction that is orghogonal to the average magnitude of hazard ratio. The degree of freedom of T_{mPH} depends on realizations of the data, and can be calculated by $tr(V^{-1}W^TB\Sigma^{-1}B^TW)$.

We now construct the second test statistic by taking the sum of T_{LR} and T_{mPH},

$$T_2 \equiv T_{LR} + T_{mPH} = S^T(\widehat{\beta}, 0, 0)\left\{W^TB\Sigma^{-1}B^TW + (\widehat{I}_0)^{-1}\mathbf{1}\mathbf{1}^T\right\}S(\widehat{\beta}, 0, 0), \tag{8}$$

The test statistic T_2 is also a quadratic form, and we can obtain the approximate asymptotic distribution according to Appendix A. Our score statistic T_2 rejects the null hypothesis H_0 at the nominal level α if $T_2 \geqslant k_2\chi^2_{\alpha,v_2}$ where $k_2 = \text{tr}(M_2VM_2V)/\text{tr}(M_2V)$, $v_2 = \{\text{tr}(M_2V)\}^2/\text{tr}(M_2VM_2V)$, and $M_2 = W^TB\Sigma^{-1}B^TW + (\widehat{I}_0)^{-1}\mathbf{1}\mathbf{1}^T$.

Remark 4. More generally, one can consider the family of linear combinations of T_{LR} and T_{mPH}, i.e., using test statistics

$$T(\rho) = \rho T_{LR} + (1-\rho)T_{mPH}, \text{ where } 0 \leqslant \rho \leqslant 1, \tag{9}$$

where ρ determines the weights from two parts. The test statistics $T(\rho)$ becomes the standard log-rank test when $\rho = 0$, and modified test for proportionality when $\rho = 1$. The proposed test T_2 corresponds to $\rho = 0.5$. The optimal choice of tuning parameter ρ is not clear and we plan to investigate this family of test statistics in the future. Our simulation studies suggest that T_2 performs well in terms of power in finite samples.

3.4 *Combination procedures based on p values*

We also explored a few other methods to combine information, e.g., **?**, taking advantage of the fact that T_{LR} and T_{mPH} are asymptotically independent. For example, two commonly used procedures for combining independent tests are based on p values, Fisher's and Tippett's procedures. Specifically, let P_{LR} and P_{mPH} denote p values from T_{LR} and T_{mPH} respectively. The test statistics for Fisher's procedure is

$$T_3 = -2\log P_{LR} - 2\log P_{mPH}.$$

Under H_0, it can be shown that T_3 follows χ^2_4, and thus p values can be calculated by $P_3 = 1 - F_{\chi^2_4}(T_3)$, where $F_{\chi^2_4}$ is the cumulative distribution function of χ^2_4. We reject the null when $T_3 > \chi^2_{4,1-\alpha}$ at significance level α. On the other hand, Tippett's procedure rejects the null when either P_{LR} or P_{mPH} is small, i.e., the test statistics is the minimum of two p values

$$T_4 = min(P_{LR}, P_{mPH}).$$

Under (asymptotic) independence of two tests, one can show that the formula for p values for the combined test is $P_4 = 1 - (1 - T_4)^2$. At significance level α, the rejection regaion is given by $T_4 < 1 - (1 - \alpha)^{1/2}$.

Note that the two-stage test T^{2stg} described in Section can also be viewed as a combination procedure based on p values. However, the proposed T_3 and T_4 combines two test statistics that are asymptotically independent, while the two statistics of T_{2stg} are possibly correlated. The finite sample performance of these procedures will be evaluated in simulation studies.

3.5 *Connection with previous literature*

As pointed above, there are two existing work, Gray (1994) and Lin et al. (2006), mostly relevant to our proposed test statistics. However, there are substantial differences between their work and the proposed tests.

First, Gray (1994) used B-splines to model $\theta(t)$ and proposed Wald, score and likelihood ratio statistics based on penalized partial likelihood for fixed values of tuning parameter. Their approach is applicable to several hypothesis testing problems, including testing time-varying treatment effect H_0 (see Section 4 of their paper). However, it depends on a tuning parameter that controls the effective degree of freedom of splines. The tuning parameter affects the power of these tests substantially and is often difficult to choose in practice. If we use the same B-spline approach, our work can be viewed as extensions of Gray (1994) through mixed effects model framework. In contrast, our proposed tests are automatic procedures that do not depend on tuning parameters, and are shown to be as or more powerful in finite samples (see the next section).

Second, Lin et al. (2006) proposed smoothing spline based score tests in extended Cox models. They discussed several hypothesis testing problems, including testing the proportional hazards null $H_{0,PH} : \theta(t) = \theta_0$ versus an alternative model with a time-varying hazard function. From a random effects model perspective, their null hypothesis can be represented via the variance component, i.e., $H_{0,PH} : \tau = 0$. Although our development adapts some technical arguments from Lin et al. (2006), we target at a different null hypothesis $H_0 : \theta(t) = 0$, or equivalently $H_0 : \theta_0 = 0, \tau = 0$. Their approach is applicable only when the null hypothesis is represented via the variance component τ only, thus excluding the problem of testing H_0. Other than the fundamental differences in the null hypotheses of interest, the primary challenge of extending their approach to our problem is how to combine information from θ_0 and τ, and this is not straightforward given the non-standard nature of variance components. We proposed a few approaches to combine score statistics for θ_0 and τ, and this is one of the main contributions of this paper from a methological perspective.

4. Numerical Studies

4.1 *Simulations*

To study the finite sample performance of the proposed tests, survival data of sample sizes $n = 100$ and $n = 500$ were generated from the extended Cox model (1) with a binary treatment indicator and no additional covariates. The baseline hazard function was a constant function, i.e., $\lambda_0(t) = 1$. The censoring distribution was the uniform distribution on $[0, c]$ where c was chosen to yield censoring probability 30% for each scenario. We conducted extensive simulation studies with various choices of $\theta(t)$, corresponding to different shapes of hazard ratio functions. The shapes of $\theta(t)$ are shown in Figure 2, most of which were considered by Gray (1994) and Lin et al. (2006) up to scale changes.

[Figure 2 about here.]

For each simulated dataset, we compared our methods with the log-rank test (denoted T_{LR}), the two stage procedure with $\alpha_1 = \alpha_2 = 0.025\%$ as described in Section 3.4 (denoted T^{2stg}), and the score statistic of Gray (1994) with pre-selected degree of freedom (denoted $G^K_{S,df}$ with K knots and degree of freedom df). For the score statistic of Gray (1994), we considered four choices, with the number of knots $K = 10$ and 20, and degrees of freedom $df = 1.5$ and 5. The performances of the likelihood ratio and Wald statistics of Gray (1994) are similar to the score statistic, and thus we only reported results for their score statistics. We also reported power from the modified test for proportionality, T_{mPH}, since it is a valid test for H_0 with a correct type I error. However, we do not mean to compare its performance with other tests directly, as T_{mPH} summarizes evidence from non-constant shape of the hazard ratio only; rather our intention is to gain insights of how each method combine information from the magnitude and shape of the hazard ratio.

Table 1 summarizes results from our simulations. In terms of type I error, all test statistics maintained the nominal level of size ($\alpha = 0.05$) approximately under the null hypothesis H_0. In terms of power, none of the test statistics is optimal in detecting all types of alternatives. It is well known that the log-rank test is most powerful for treatment effect under the PH assumption, but may lose power otherwise. The aim of our work is to propose new "omnibus" tests that do not

lose much power compared to log-rank under the PH assumption and also have decent power to detect non-PH alternatives.

We now compare power of various test statistics under both PH alternatives and a wide variety of non-PH alternatives. When the PH assumption holds and the true model is the Cox model ($H_{0,PH}$), the log-rank test was the most powerful as expected. Gray's score tests performed well, particularly with low degree of freedom ($df = 1.5$), but lost power substantially with higher degree of freedom ($df = 5$). The proposed linear combination statistics T_1 and T_2 have slightly lower power compared to the log-rank test, but the differences are small. The proposed p-value based statistics T_3 and T_4 are less powerful than the linear combination tests from simulations.

[Table 1 about here.]

Under non-PH alternatives, there is no universally best test according to Table 1. First, we compare power performance between two proposed statistics T_1 and T_2. Under several scenarios (*L, Q, E1, Expit, Log2*), T_1 has slightly higher power than T_2 by around 2% and 6%. On the other hand, T_2 outperforms T_1 in other settings (*E2, Log1, S, C*), but generally with a more substantial power gain that varies between 7% to as high as 38% (Log1). Thus, the statistic T_2 is considered "omnibus" in the sense that it has decent power against all types of alternatives and is our preferred choice, while T_1 is prone to very low power to detect certain types of alternatives. The substantial power gain of T_2 is likely due to the fact that the latter exploits orthogonality between information on the magnitude and shape of the hazard ratio function.

Next, we compare the proposed score statistic T_2 versus the log-rank test (T_{LR}), the modified PH test (T_{mPH}) and the two-stage procedure (T^{2stg}) under non-PH alternatives. Under many settings, T_2 outperforms both T_{LR} and $T_{two-stg}$. For example, under alternative curve *Log2* and sample size of 500, the power for T_2 is 61.4%, much higher than both T_{LR} (44.9%) and T^{2stg} (55.7%). Under some alternatives (*E1, E2, Expit*), the power of T_2 is lower than the log-rank test T_{LR} but only by very slight margins. Thus, the proposed test T_2 is generally comparable or more powerful than T_{LR}, since it combines information from the shape of hazard functions. The proposed test can potentially pick up evidence of treatment effects even if both T_{LR} and $TwoStg$ fail to suggest so.

We also compared T_2 versus Gray's score tests. The power of Gray's score tests varies with tuning parameters, especially with the degree of freedom, which needs to be pre-specified and can affect power performance considerably. Their score tests with low *df* ($G^{10}_{S,1.5}$ and $G^{20}_{S,1.5}$) perform well when the hazard function is close to a constant or linear function but poorly otherwise, while the opposite holds for the tests with high *df* ($G^{10}_{S,5}$ and $G^{20}_{S,5}$). The proposed test T_2 is often comparable or close to Gray's statistics with the "better" choice of *df* in terms of power, and sometimes outperform all of them under certain alternatives (*Log1* and *C*).

To summarize, the proposed test statistics, especially the preferred T_2, demonstrate decent power performances under various alternative hypotheses in simulations. Other testing procedures, such as the log-rank test, the two-stage procedure and Gray (1994)'s score tests, are often powerful against certain alternatives but may lose power substantially against others. In addition, T_2 does not depend on tuning parameters, making it an omnibus and desirable testing procedure to use in practice.

4.2 *Application to HIVNET 012 Study*

We demonstrate our proposed methods by analyzing our motivating example of infant survival in the HIVNET 012 Study mentioned in Section 1. In our data set, 310 women were assigned to the NVP group and 306 women to the AZT group. We exclude the second twins or more and babies with still birth from the analysis. Mean CD4+ counts at the baseline for mothers in both groups were 482 and 465 ($p = .41$); mean log RNA viral copies with base 10 at visit 101 were 4.35 and 4.39 ($p = .59$); mean birth weights were 3080 kg and 3197 kg ($p = 0.001$), respectively. The total follow-up time is 18 months. The HIV-1 transmission risk and risk of death at the end of the study is 14.9 $\%$ and 23.5$\%$ ($p = 0.007$), and 9.3$\%$ and 12.7$\%$ ($p = .18$).

As we discussed in Section 1, the log-rank test does not show significance ($p = 0.145$), with estimated hazard ratio 0.701 (95 $\%$ CI: 0.43-1.13). However, the lack of statistical significance may be due to the fact that the log-rank test is not powerful when the PH assumption does not hold, which seemed to be the case in this application (Figure 5).

Next, we applied alternative tests, including weighted log-rank test, Gray's test and the proposed tests. For weighted log-rank tests, we used weights corresponding to Peto-Peto modification of the Wilcoxon statistic (Peto and Peto, 1972) considering the drug effect gradually disappeared over time, but the result is not significant ($P = 0.125$). We also tried weights from the G^ρ family (Harrington and Fleming, 1982) with different ρ's, but the usual choice of ρ between 0 and 1 did not yield a significant result. Gray's score tests yielded different results depending on tuning parameters such as degrees of freedom and numbers of knots. Specifically, p values corresponding to $df = 1.5, 3$ and 5 are $0.073, 0.032$ and 0.057, respectively, with $K = 10$ knots, and are $0.069, 0.032$ and 0.053, respectively, with $K = 20$ knots. These results may be confusing to practitioners since different df led to different p values, and it is not clear which df one should choose for this specific application. The proposed test T_2 suggested significant treatment effects ($p = 0.015$). To confirm the proportionality of the hazard ratio, we also applied the test and obtained p value ($p = 0.013$), which suggested that the hazard ratio is time-varying and that the log-rank test is not optimal in this setting.

5. Discussion

We developed spline based score tests for time-varying treatment effects in an extended Cox model. The proposed approach is designed to test treatment effects when the proportional hazards assumption may not hold. These test statistics do not depend on tuning parameters and are easy to compute since they only requires fitting the null model (Cox model). Simulation studies suggested that our methods gained power substantially compared to the log-rank test when the proportional hazards assumption do not hold.

There are some connections between the proposed tests and the widely used weighted log-rank test. As shown in Appendix B, the family of quadratic form tests $Q = S^T U S$, which include the weighted log-rank test, Lin et al. (2006)'s proportionality test and the proposed tests, are equivalent to a linear combination of several weighted log-rank tests with weights determined by the eigenvectors of the matrix U. The weighted log-rank test (when U has rank 1) is powerful when

the chosen weights are close to the true hazard ratio functions, but may lose power substantially otherwise. In contrast, the proposed tests combine several plausible weighted log-rank tests, some of which reflect information on the smoothness of the hazard ratio function, and thus provide omnibus testing procedures that are powerful to detect a variety of alternatives.

There are a few areas for future research. As discussed in Section 2.2, one main challenge is how to combine scores for parameters θ_0 and τ, given the non-standard nature of variance components testing. While the two proposed statistics are natural choices, they are not necessarily optimal in terms of power. There may be other ways to combine the score statistics. For example, one may explore the family of linear combinations and find an optimal weight within this family, or even identify the optimal statistics among nonlinear combinations if possible. The proposed methods can also be applied to other models and setting to detect a nonlinear trend. Although we focused on an extended Cox model, our method can be extended to other models such as an additive hazards models.

A reviewer raised an interesting question of deriving a method to differentiate three hypotheses: the null, constant or truly time-varying effect of treatment. Two step procedures will be needed to accomplish this, as standard hypothesis testing is designed to distinguish two hypotheses only. For example, one can first apply the test of Lin et al. (2006), which differentiate "null or constant effect" vs. time-varying effect. If the null was not rejected in the first step, then apply the log-rank test to further differentiate null vs. constant effect. Alternatively, one can first apply the proposed test statistics, which differentiates null vs. "constant or time-varying effect." If the null was rejected, then apply the test of Lin et al. (2006) to further differentiate constant vs. time-varying effect.

REFERENCES

Brown, D., Kauermann, G., and Ford, I. (2007). A partial likelihood approach to smooth estimation of dynamic covariate effects using penalised splines. *Biom. J.* **49,** 441–452.

Cox, D. R. (1972). Regression models and life-tables. *J. Roy. Statist. Soc. Ser. B* **34,** 187–220. With discussion.

Gray, R. J. (1994). Spline-based tests in survival analysis. *Biometrics* **50,** 640–652.

Group, T. H. (2003). The hivnet 012 protocol: Phase iib trial to evaluate the efficacy of oral nevirapine and the efficacy of oral azt in infants born to hiv-infected mothers in uganda for prevention of vertical hiv transmission. *The HIV Prevention Trials Network: http://www.hptn.org.* .

Harrington, D. P. and Fleming, T. R. (1982). A class of rank test procedures for censored survival data. *Biometrika* **69,** 553–566.

Hastie, T. and Tibshirani, R. (1990). Exploring the nature of covariate effects in the proportional hazards model. *Biometrics* **46,** pp. 1005–1016.

Jackson, J. B., Musoke, P., Fleming, T., Guay, L. A., Bagenda, D., Allen, M., Nakabiito, C., Sherman, J., Bakaki, P., Owor, M., Ducar, C., Deseyve, M., Mwatha, A., Emel, L., Duefield, C., Mirochnick, M., Fowler, M. G., Mofenson, L., Miotti, P., Gigliotti, M., Bray, D., and Mmiro, F. (2003). Intrapartum and neonatal single-dose nevirapine compared with zidovudine for prevention of mother-to-child transmission of HIV-1 in Kampala, Uganda: 18-month follow-up of the HIVNET 012 randomised trial. *Lancet* **362,** 859–868.

Kooperberg, C., Stone, C. J., and Truong, Y. K. (1995). Hazard regression. *J. Amer. Statist. Assoc.* **90,** 78–94.

Lin, J., Zhang, D., and Davidian, M. (2006). Smoothing spline-based score tests for proportional hazards models. *Biometrics* **62,** 803–812.

Murphy, S. A. and van der Vaart, A. W. (2000). On profile likelihood. *J. Amer. Statist. Assoc.* **95,** 449–485. With comments and a rejoinder by the authors.

O'Sullivan, F. (1988). Nonparametric estimation of relative risk using splines and cross-validation. *SIAM J. Sci. Statist. Comput.* **9,** 531–542.

Peto, R. and Peto, J. (1972). Asymptotically efficient rank invariant test procedures. *Journal of the Royal Statistical Society. Series A (General)* **135,** pp. 185–207.

Pettitt, A. N. and Bin Daud, I. (1990). Investigating time dependence in Cox's proportional hazards model. *J. Roy. Statist. Soc. Ser. C* **39,** 313–329.

Sleeper, L. A. and Harrington, D. P. (1990). Regression splines in the cox model with application

to covariate effects in liver disease. *Journal of the American Statistical Association* **85,** pp. 941–949.

Zucker, D. M. and Karr, A. F. (1990). Nonparametric survival analysis with time-dependent covariate effects: a penalized partial likelihood approach. *Ann. Statist.* **18,** 329–353.

APPENDIX

A. *Asymptotic distributions of test statistics*

The two proposed test statistics (T_1 and T_2), as well as the log-rank (T_{LR}) and modified PH (T_{mPH}) test statistics, are all asymptotically equivalent to quadratic forms of $S(\widehat{\beta}, 0, 0)$. In the following, we describe asymptotic distributions of such quadratic forms, as well as approximation methods to calculate p values. We follow similar arguments.

We consider a general quadratic form $Q = S^T U S$, where $S := S(\widehat{\beta}, 0, 0)$ is a vector of length r and U is a positive semi-definite matrix of size $r \times r$. Since each element of S is a realization of the score function, S has mean 0 and its variance-variance matrix is the Fisher information V. One can rewrite $Q = S^T U S = (V^{-1/2}S)^T (V^{1/2} U V^{1/2})(V^{-1/2}S)$, where $V^{-1/2}S$ are standardized S with identity matrix as its covariance matrix. Using quadratic form theory and the central limit theorem, one obtains the following result.

Proposition. Asymptotically, the distribution of the quadratic form $Q = S^T U S$ is approximately a weighted average of χ^2_1, more specifically,

$$Q \to \sum_k \lambda_k \chi^2_1,$$

where λ_k's are eigenvalues of the matrix UV. The mean and variance of the limiting distribution are $tr(UV)$ and $tr(UVUV)$, respectively.

In practice, it is often the case that the first few eigenvalues capture the most variations and the remaining ones are negligible. To calculate p values, one can use further approximation $c\chi^2_v$, i.e., a scaled χ^2 distribution with degree of freedom v. By matching the mean and variance of the two distributions, one can obtain the choice of parameters $c = tr(UVUV)/tr(UV)$ and $v =$

$\{tr(UV)\}^2/tr(UVUV)$. In simulations, we found that both approximations work reasonably well in finite samples.

B. *Connection with weighted log-rank tests via spectral decomposition*

In this section, we will apply the spectral decomposition to understand the connection between the proposed tests and the weighted log-rank test. Consider the general quadratic form $Q = S^T U S$, where U is a non-negative semi-definite matrix. One has spectral decomposition $U = \sum_k \lambda_k P_k P_k^T$, where λ_k's and P_k's are eigenvalues and eigenvectors of U, respectively. Using such decomposition, the quadratic form can be written as

$$Q = S^T U S = \sum_k \lambda_k \, S^T P_k P_k^T S = \sum_k \lambda_k (P_k^T S)^T (P_k^T S). \tag{A.1}$$

Note that the k^{th} term is equivalent to the weighted log-rank statistic with weight P_k. Thus, the test statistic Q is equivalent to a linear combination of several weighted log-rank statistics, with weights determined by the eigenvectors of the matrix U. The relative importance of each weighted log-rank statistics in the linear combination is determined by the eigenvalues λ_k's.

If the matrix U has rank 1 and thus only one eigenvector P_1, the test statistic Q is actually weighted log-rank test with weight P_1 (unweighted log-rank test if and only if $U \propto \mathbf{1}\mathbf{1}^T$, or equivalently, $P_1 \propto \mathbf{1}$). If $rank(U) > 1$, the quadratic form Q is equivalent to a linear combination of several weighted log-rank statistics, different from any weighted log-rank tests. The resulting test statistics incorporate information from deviation from the null in several different directions, and thus are expected to be omnibus when the shape of true hazard ratio function is unknown. In Lin et al. (2006), they chose $U = \Sigma$, which was derived from the differential operator, and their test statistic would summarize information from possible non-proportionality. For the proposed tests T_1 and T_2, we choose the matrix U to be a linear combination of $\mathbf{1}\mathbf{1}^T$ and Σ, and thus our test statistics combine information from both the magnitude and shape of the hazard ratio function.

C. *A sketch of proof of the properties of* T_2

We provide a sketch of proof to show that $\mathbf{1}^T S(\widehat{\beta}, 0, 0)$ and $W(\widehat{\beta}) S(\widehat{\beta}, 0, 0)$ are approximately uncorrelated under the null. Therefore, T_2 is expected to combine information from $\mathbf{1}^T S(\widehat{\beta}, 0, 0)$

and $S(\widehat{\beta}, 0, 0)$ effectively. Because the profile likelihood is an approximately least favorable submodel of the Cox model (Murphy and van der Vaart, 2000), we treat the partial likelihood as a legitimate likelihood from a parametric model without a nuisance parameter. Denote the partial likelihood as $\ell(\beta, \theta_0, \underline{\theta})$. Note that the projection of a random vector X onto a random vector Y is $Z = \{cov(X, Y)/var(Y)\}Y$. We consider $X = S(\beta_0, 0, 0)$ and $Y = \mathbf{1}^T S(\beta_0, 0, 0)$ where β_0 is a true parameter. Then the projection of X onto Y is given by

$$\begin{aligned} Z &= \frac{cov(S(\beta_0, 0, 0), \mathbf{1}^T S(\beta_0, 0, 0))}{var(\mathbf{1}^T S(\beta_0, 0, 0))} \mathbf{1}^T S(\beta_0, 0, 0) \\ &= \frac{E\{S(\beta_0, 0, 0) S(\beta_0, 0, 0)^T\})\mathbf{1}}{\mathbf{1}^T var\{S(\beta_0, 0, 0)\}\mathbf{1}} \mathbf{1}^T S(\beta_0, 0, 0), \end{aligned}$$

Since

$$ES(\beta_0, 0, 0)^{\otimes 2} = E\left[\{(\partial/\partial\underline{\theta})\ell(\beta_0, 0, 0)\}^{\otimes 2}\right] = -E\left[(\partial^2/\partial\underline{\theta}\partial\underline{\theta}^T)\ell(\beta_0, 0, 0)\right] = -E\dot{S}(\beta_0, 0, 0),$$

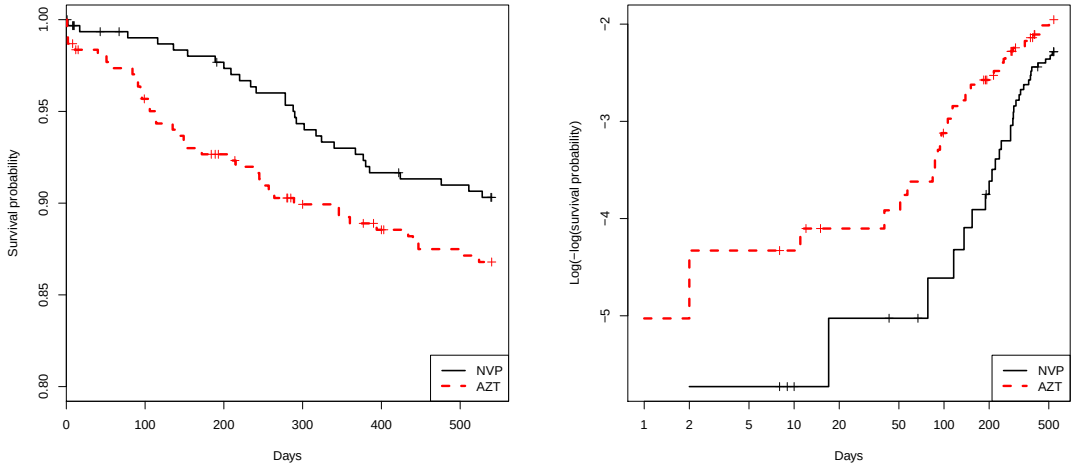
we obtain

$$Z = \frac{E\{\dot{S}(\beta_0, 0, 0)\}}{\mathbf{1}^T E\{\dot{S}(\beta_0, 0, 0)\}\mathbf{1}} \mathbf{1}\mathbf{1}^T S(\beta_0, 0, 0).$$

Since $E\{\dot{S}(\beta_0, 0, 0)\}$ and β_0 are unknown, we replace these by a empirical version, $n^{-1}\dot{S}(\widehat{\beta}, 0, 0)$, and an estimate, $\widehat{\beta}$, to obtain

$$\widetilde{Z} = \frac{\dot{S}(\widehat{\beta}, 0, 0)}{\mathbf{1}^T \dot{S}(\widehat{\beta}, 0, 0)\mathbf{1}} \mathbf{1}\mathbf{1}^T S(\widehat{\beta}, 0, 0) = (W(\widehat{\theta}) - I)S(\widehat{\beta}, 0, 0).$$

Thus, we expect $\mathbf{1}^T S(\widehat{\beta}, 0, 0)$ and $W(\widehat{\theta})S(\widehat{\beta}, 0, 0)$ are approximately uncorrelated.



(a) Kaplan-Meier Curves

(b) Log negative log of KM curves

Figure 1. Kaplan-Meier curves for two treatment arms and their transformation by log negative log.

Figure 2. Functional shapes of hazard ratio $\theta(t)$ that were used in simulations. We considered nine hazard functions, including linear (L), quadratic (Q), exponential (E1, E2), inverse-logistic (Expit), logarithm (Log1, Log2), step (S) and cosine (C) functions. Their specific functional forms are specified in Table 1.

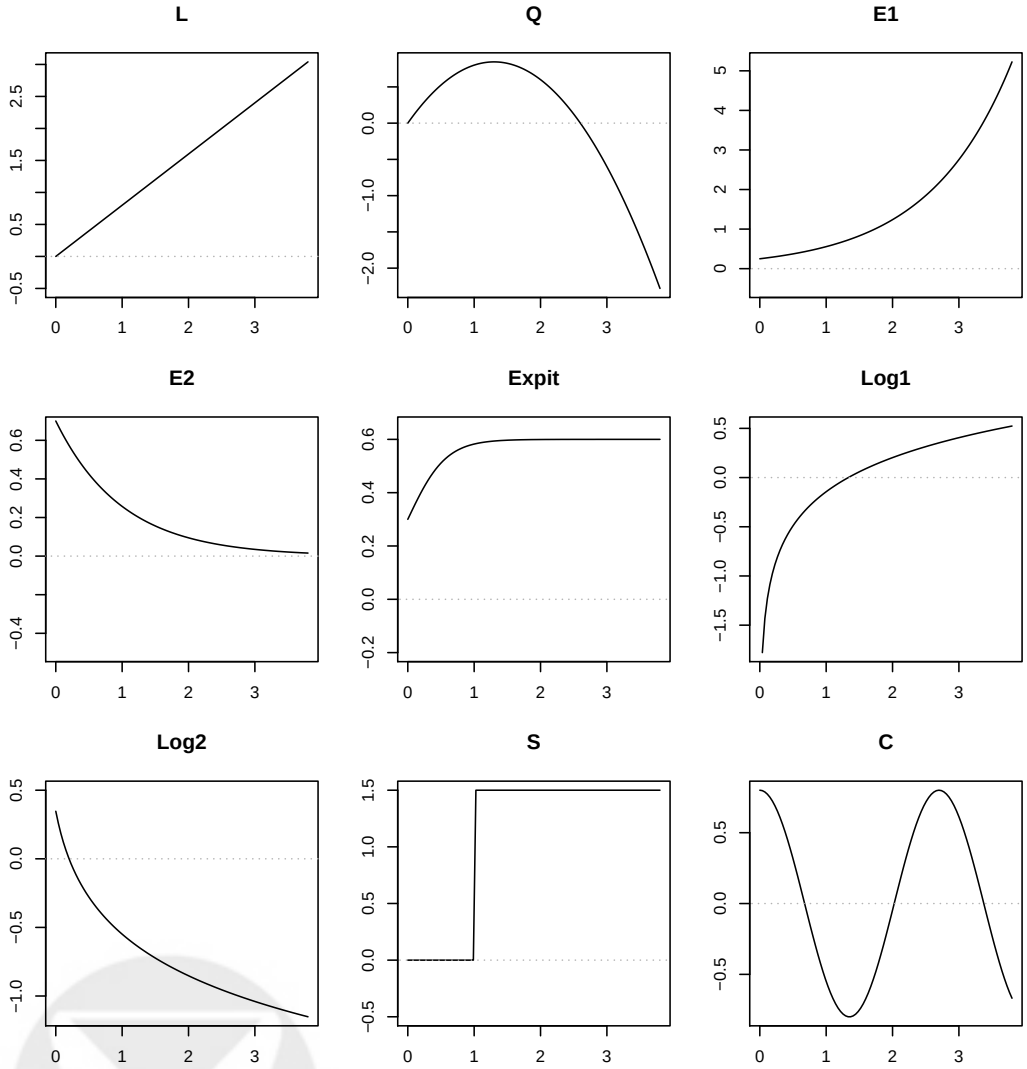


Table 1

Type I error rates and power for the proposed tests versus alternatives in simulations. The testing procedures include standard log-rank test (T_{LR}), modified test for proportionality (T_{mPH}), the two-stage test (T^{2stg}), Gray's score tests ($G^{K}_{S,df}$ with K knots and degree of freedom df) and the proposed test statistics (T_1, T_2, T_3 and T_4).

$\theta(t)$	n	T_{LR}	T_{mPH}	T^{2stg}	$G^{10}_{S,1.5}$	$G^{10}_{S,5}$	$G^{20}_{S,1.5}$	$G^{20}_{S,5}$	T_1	T_2	T_3	T_4
H_0: 0	100	0.058	0.049	0.059	0.056	0.051	0.057	0.061	0.056	0.052	0.055	0.059
	500	0.042	0.063	0.060	0.056	0.058	0.059	0.053	0.054	0.062	0.059	0.060
$H_{0,PH}$: $\log 1.5$	100	0.363	0.042	0.292	0.369	0.258	0.418	0.271	0.329	0.323	0.278	0.298
	500	0.971	0.053	0.933	0.966	0.905	0.964	0.912	0.943	0.946	0.932	0.934
L: $0.8t$	100	0.350	0.221	0.369	0.421	0.373	0.444	0.424	0.491	0.428	0.407	0.373
	500	0.968	0.843	0.986	0.994	0.988	0.990	0.994	0.998	0.995	0.992	0.986
Q: $-0.5t(t-2.6)$	100	0.423	0.163	0.409	0.499	0.390	0.494	0.446	0.534	0.466	0.429	0.414
	500	0.990	0.636	0.987	0.995	0.991	0.993	0.991	0.996	0.993	0.992	0.987
E1: $0.25\exp(0.8t)$	100	0.352	0.082	0.296	0.382	0.279	0.401	0.306	0.393	0.334	0.293	0.305
	500	0.968	0.316	0.954	0.980	0.950	0.971	0.954	0.981	0.971	0.969	0.956
E2: $0.7\exp(-t)$	100	0.493	0.095	0.414	0.508	0.368	0.530	0.380	0.368	0.468	0.444	0.420
	500	0.990	0.285	0.991	0.997	0.988	0.994	0.981	0.975	0.995	0.993	0.991
Expit: $\frac{0.6\exp(3.5t)}{1+\exp(3.5t)}$	100	0.503	0.064	0.413	0.520	0.398	0.528	0.432	0.505	0.457	0.406	0.422
	500	0.995	0.151	0.993	1.000	0.983	0.991	0.987	0.995	0.993	0.991	0.993
Log1: $0.5\log(.75t)$	100	0.449	0.393	0.557	0.525	0.549	0.518	0.550	0.231	0.614	0.619	0.561
	500	0.978	0.976	0.999	0.997	0.999	0.997	1.000	0.835	1.000	1.000	0.999
Log2: $0.5\log\frac{2}{1+5t}$	100	0.165	0.263	0.275	0.234	0.255	0.223	0.279	0.348	0.314	0.300	0.283
	500	0.652	0.862	0.911	0.882	0.922	0.871	0.946	0.952	0.941	0.940	0.912
S: $1.5I(t<1.0)$	100	0.187	0.439	0.413	0.297	0.494	0.307	0.514	0.372	0.441	0.449	0.417
	500	0.760	0.996	0.997	0.996	0.998	0.987	0.999	0.990	0.998	0.998	0.997
C: $0.8\cos(2\pi t/2.7)$	100	0.354	0.440	0.545	0.475	0.501	0.484	0.499	0.245	0.576	0.606	0.549
	500	0.949	0.981	0.999	0.998	0.999	0.994	0.997	0.933	0.999	0.999	0.999