On graphical tests for proportionality of hazards in two samples

Technical Report No. ASU/2014/5

Dated: 19 June 2014

Shyamsundar Sahoo, Haldia Government College, Haldia

and

Debasis Sengupta, Indian Statistical Institute, Kolkata

Indian Statistical Institute

Applied Statistics Unit

Kolkata 700 108

On graphical tests for proportionality of hazards in two samples

Shyamsundar Sahoo, Haldia Government College, Haldia, India
Debasis Sengupta, Indian Statistical Institute, Kolkata, India

June 19, 2014

Abstract

In this paper, we present a class of graphical tests for the proportional hazards hypothesis in two-sample censored survival data. The proposed tests are improvements over some existing tests based on asymptotic confidence bands of certain functions of the estimated cumulative hazard functions. These methods are based on comparison of parametric and nonparametric estimates of the said functions, and they attempt to combine the rigour of analytical tests with the descriptive value of plots. We show theoretically as well as through simulations, how the existing tests can be improved upon. The simulations suggest that the proposed asymptotic procedures have reasonable small sample properties. The methods are then illustrated through the analysis of a data set on malignant melanoma and also on a data set on bone marrow transplantation for Leukemia patients.

*Key words and phrases.*Proportional hazards model; Monotone hazard ratio; Confidence band; Acceptance band; Two-sample problem.

1 Introduction

A widely used model for comparing two survival distributions from censored samples is the *proportional hazards* (PH) model. The problem of testing for the proportionality of hazards in two samples has been visited often by various researchers. Apart from being an important problem by its own right, it is also an important diagnostic issue for Cox's regression model (Cox, 1972), which has application in several fields such as clinical trial, reliability analysis, demography, actuarial science, environmental science, microeconomics and so on.

There are several analytical tests of the PH assumption. Some examples can be found in Wei (1984), Breslow et al. (1984), Gill and Schumacher (1987), Deshpande and Sengupta (1995), Sengupta et al. (1998) and the references therein.

A well-known drawback of analytical tests is that, in the case of rejection, they do not provide a clear pointer towards possible alternative modeling. In view of this limitation, *plots* for checking the proportionality of hazards in two samples have become increasingly popular in recent years (see Lee and Pirie (1981), Gill and Schumacher (1987), Kay (1977), Dabrowska et al. (1989) and Dabrowska et al. (1992)). These plots permit the human eye to pick up patterns that may suggest specific forms of departure from the ideal shape corresponding to proportional hazards.

Some researchers (see, e.g., Andersen (1982), Gill and Schumacher (1987)) viewed analytical tests and diagnostic plots as mutually complementary tools of data analysis. Dabrowska et al. (1989) and Dabrowska et al. (1992), on the other hand, sought to merge them by augmenting the plots with probabilistic threshold limits computed under the PH hypothesis. The modified plots, which may be regarded as graphical tests, combine the insight provided by plots with the formalism of analytical tests. Therneau et al. (1990) and Lin et al. (1993) proposed graphical tests that make use of martingale based residuals, for which critical values of the limiting process can serve as formal reference.

The graphical tests developed by Dabrowska et al. (1989) and Dabrowska et al. (1992) are based on asymptotic confidence bands for the plots. The population version of these plots reduce to a straight line under the PH hypothesis. Therefore, rejection of the null hypothesis is recommended if one cannot draw an appropriate straight line through the gap between the upper and the lower confidence bands, designed so that the probability of false rejection does not exceed a specified threshold.

In this paper, we argue that this procedure is too conservative. We recommend a set of alternative graphical tests based on what we call 'acceptance bands' for the same plots.

The aim of the paper is to demonstrate how a class of existing graphical methods can be improved, rather than to propose a completely new method. The main idea may be extended beyond survival analysis for improving other graphical methods, as pointed out in Section 6.

In Section 2, we demonstrate theoretically the limitations of the existing graphical tests for checking the PH assumption in two populations. We propose in Section 3 a class of modified graphical tests based on the deviation of the sample plot from the ideal population plot under the null hypothesis, present some weak convergence results for the deviation processes, and show how asymptotic tests based on acceptance bands may be constructed. In Section 4, we study the graphical tests through Monte Carlo simulations, and report their size and power properties for finite sample size. In Section 5, we illustrate the graphical methods through the analysis of some real data sets. We provide some concluding remarks in Section 6.

2 Existing graphical tests

2.1 Diagnostic plots

Let $X_{11}, X_{12}, \ldots, X_{1n_1}$ and $X_{21}, X_{22}, \ldots, X_{2n_2}$ be independent samples from life distributions F_1 and F_2, respectively. Observations consist of randomly right censored versions of these, namely the pairs (X'_{ji}, δ_{ji}) for $i = 1, 2, \ldots, n_j$, $j = 1, 2$, where $X'_{ji} = \min(X_{ji}, C_{ji})$, δ_{ji} is the binary indicator of the event $X_{ji} \leq C_{ji}$, and $C_{11}, C_{12}, \ldots, C_{1n_1}$ and $C_{21}, C_{22}, \ldots, C_{2n_2}$ are independent samples from the respective censoring distributions F_1^c and F_2^c that are independent of the X_{ji}'s. Let $N_1(t)$ and $N_2(t)$, $t > 0$, be independent counting processes representing the total number of observed failures till time t for the two groups, and $Y_1(t)$ and $Y_2(t)$ be the 'number at risk of failure' immediately prior to time t for the two groups. We assume that all the distributions are absolutely continuous with respect to the Lebesgue measure. It follows that the cumulative hazard functions, $\Lambda_i(t) = -\log(1 - F_i(t))$, $i = 1, 2$, are also continuous. The Nelson-Aalen estimators of the cumulative hazard functions are

$$\widehat{\Lambda}_i(t) = \int_0^t \{Y_i(s)\}^{-1} dN_i(s), \qquad i = 1, 2.$$

The Nelson-Aalen estimator has been used to construct many diagnostic plots for checking the appropriateness of the PH model. The proportionality of hazards in two samples means that there is a real positive value θ such that $\Lambda_2(t) = \theta\Lambda_1(t)$. To detect the proportionality of hazards, Kay (1977) suggested simultaneous plotting of $\log \widehat{\Lambda}_2(t)$ and $\log \widehat{\Lambda}_1(t)$ versus t. Now, since $\log \Lambda_2(t) = \log \theta + \log \Lambda_1(t)$ under the PH model, the separation between the two estimated curves should be approximately constant if the PH model holds. Further, when the hazard rates are proportional (i.e., the cumulative hazard functions are proportional), the graph of $\Lambda_2(t)$ vs. $\Lambda_1(t)$ for $t > 0$ is a straight line through origin, with slope equal to the constant of proportionality between the two hazards. Therefore, one may check the plotting of $\widehat{\Lambda}_2(t)$ vs. $\widehat{\Lambda}_1(t)$ and look for the approximate shape of a straight line through the origin – as an indicator of proportionality of the hazards. Lee and Pirie (1981) proposed, in a some what different context, the plot of the function $\Lambda_2 \circ {\Lambda_1}^{-1}(u)$ vs. u which is a straight line through origin with slope θ when the PH model holds. If one defines the function ${\widehat{\Lambda}_1}^{-1}$ by the relation ${\widehat{\Lambda}_1}^{-1}(u) = \inf\{t : t \geq 0, \widehat{\Lambda}_1(t) \geq u\}$, then it is easy to see that the graph of the function $\widehat{\Lambda}_2 \circ {\widehat{\Lambda}_1}^{-1}(u)$ has the shape of a staircase, passing through pairs $(\widehat{\Lambda}_2(t), \widehat{\Lambda}_1(t))$ for different values of t. For this reason, the plot of Lee and Pirie is often referred to as the relative trend function (RTF) plot. Building further on Kay's (1977) idea, Dabrowska et al. (1992) considered a plot of the log cumulative hazard difference (LCHD) $\log \widehat{\Lambda}_2(t) - \log \widehat{\Lambda}_1(t)$ versus t, which makes it easier to see whether or not the difference curve is *constant*. This plot is related to a plot of the relative cumulative hazard difference (RCHD) $(\widehat{\Lambda}_2(t) - \widehat{\Lambda}_1(t))/\widehat{\Lambda}_1(t)$ versus t, which should also resemble a horizontal straight line if the hazards are proportional. This plot had been proposed by Dabrowska et al. (1989).

An interesting aspect of these plots is their behavior when the hazard ratio is strictly increasing or decreasing. This scenario of 'monotone hazard ratio', which

is considered to be an important form of departure from the PH situation, implies that the ratio of the cumulative hazard functions is also strictly monotone. It turns out that in this case, the theoretical counterparts of all these plots have a special shape. In particular, the graph of $\Lambda_2(t)$ vs. $\Lambda_1(t)$ becomes strictly convex or concave (rather than being a straight line), and the graphs of $\log \Lambda_2(t) - \log \Lambda_1(t)$ vs. t and $(\Lambda_2(t) - \Lambda_1(t))/\Lambda_1(t)$ vs. t have a strictly increasing or decreasing trend (rather than being a horizontal straight line). Therefore, the empirical versions of the plots are expected to be sensitive to the scenario of 'monotone hazard ratio'. Other shapes of the graph of $\Lambda_2(t)$ vs. $\Lambda_1(t)$ also have interesting interpretation; see Sengupta and Deshpande (1994).

A drawback of these plots is that these are prone to high variability near the right tail, especially for moderate and small sample sizes. This variability occurs because of the impact of the smaller risk sets (near the right tail) on the Nelson-Aalen estimator. This problem may be rectified following the suggestion of Gill and Schumacher (1987), who proposed a generalization of the plot of $\widehat{\Lambda}_2(t)$ vs. $\widehat{\Lambda}_1(t)$ by replacing $\widehat{\Lambda}_i(t)$ by $\widehat{\Lambda}_i^K(t)$, i=1, 2, where

$$\widehat{\Lambda}_i^K(t) = \int_0^t K(u) d\widehat{\Lambda}_i(u), \qquad i = 1, 2,$$

and $K(u)$ is a predictable weight function, in the sense that it depends only on observations up to (but excluding) time u. A monotone decreasing weight function can bring more stability to this plot near the right tail. One can similarly modify the plots of $\log \widehat{\Lambda}_2(t) - \log \widehat{\Lambda}_1(t)$ vs. t and $(\widehat{\Lambda}_2(t) - \widehat{\Lambda}_1(t))/\widehat{\Lambda}_1(t)$ vs. t with suitable weight functions. Gill and Schumacher (1987) suggested a number of weight functions satisfying the requisite conditions.

2.2 Graphical test procedures

There have been attempts to build graphical tests based on diagnostic plots, with the objective of combining the exploratory value of plots and the formalism of analytical tests. A typical route to obtaining goodness of fit tests (for the PH assumption) from diagnostic plots is through asymptotic confidence bands of the plots.

Dabrowska et al. (1989) provided asymptotic confidence bands of the RTF plot and proposed a conservative graphical test for the hypothesis of proportional hazards by checking whether the band contains a straight line through the origin. This procedure can be adapted in a straightforward manner to the trend function $\widehat{\Lambda}_2^K \circ \widehat{\Lambda}_1^{K^{-1}}(u)$ based on the weighted cumulative hazard functions.

Dabrowska et al. (1989) proposed another test, in which the null hypothesis of PH model is rejected when no horizontal line fits in the asymptotic confidence bands of the RCHD plot. The asymptotic confidence bands of the LCHD plot, obtained by Dabrowska et al. (1992) can also be used to obtain a formal test, by checking whether a horizontal line can be passed through them. Both the tests would continue to work, with minor modification of the asymptotic results (indicated in Theorem 3.1 below), if the estimated cumulative hazard functions are replaced by their suitably weighted versions.

2.3 Weakness of the graphical tests

The three plots mentioned in the foregoing section are graphs of the RTF, the LCHD and the RCHD functions. If we represent by h the function used in these plots, the null hypothesis corresponds to a particular (parametric) form of this function. We can express the null and the alternative hypotheses as

$$\begin{aligned} H_0 &: h(t) = h_0(t;\theta) \text{ for some positive } \theta \in \Theta_0, \\ \text{vs. } H_1 &: h(t) \neq h_0(t;\theta) \text{ for any positive } \theta \in \Theta_0, \end{aligned}$$

where $h_0(t;\theta)$ is a completely known function of t except for the parameter θ. The different forms of $h(t)$ and $h_0(t;\theta)$ corresponding to the three diagnostic plots are given in Table 1.

Table 1
$h(t)$ and $h_0(t;\theta)$ for different plots

Plot	Form of $h(t)$	Form of $h_0(t;\theta)$	Interpretation of θ
RTF	$\Lambda_2 \circ \Lambda_1^{-1}(t)$	θt	slope parameter
LCHD	$\log \Lambda_2(t) - \log \Lambda_1(t)$	$\log \theta$	parameter controlling intercept
RCHD	$(\Lambda_2(t) - \Lambda_1(t))/\Lambda_1(t)$	$\theta - 1$	parameter controlling intercept

If θ had been completely specified under the null hypothesis, i.e., if $\Theta_0 = \{\theta_0\}$, then a graphical test would amount to finding a pair of curves $\widehat{h}_l(t)$ and $\widehat{h}_u(t)$ such that $h_0(t;\theta_0)$ is sandwiched between this pair with specified probability, say $1-\alpha$, under the null hypothesis. In other words, the requirement is that

$$P_{\theta_0}\left(\widehat{h}_l(t) \leq h_0(t;\theta_0) \leq \widehat{h}_u(t), \forall t\right) = 1 - \alpha,$$

where $P_\theta(A)$ indicates the probability of the event A for parameter value θ. The decision rule of this test is to reject H_0 if $h_0(t;\theta_0)$ ever strays beyond the gap between $\widehat{h}_l(t)$ and $\widehat{h}_u(t)$, which may presumably be obtained on the basis of a nonparametric estimator of h.

In the case of a composite null hypothesis, one may obtain a $(1-\alpha)$ level confidence band of $h(t)$ in the form of $\widehat{h}_l(t)$ and $\widehat{h}_u(t)$, such that

$$\inf_{\theta \in \Theta_0} P_\theta\left\{\widehat{h}_l(t) \leq h(t) \leq \widehat{h}_u(t), \forall t\right\} = 1 - \alpha.$$

One may reject H_0 at level α if *no* $h_0(t;\theta)$ with $\theta \in \Theta_0$ lies completely within the band, that is, if the event

$$\bigcup_{\theta \in \Theta_0} \left\{\widehat{h}_l(t) \leq h_0(t;\theta) \leq \widehat{h}_u(t), \forall t\right\}$$

does not occur. Indeed, this has been the test procedure recommended for the three plots mentioned in Table 1 (see Dabrowska et al. (1989), Dabrowska et al. (1992), Andersen et al. (1992)).

The effective size of each of these graphical tests happens to be

$$\begin{aligned} & 1 - \inf_{\theta \in \Theta_0} P_\theta \left[\bigcup_{\phi \in \Theta_0} \left\{ \widehat{h}_l(t) \leq h_0(t;\phi) \leq \widehat{h}_u(t), \forall t \right\} \right] \\ \leq \; & 1 - \inf_{\theta \in \Theta_0} P_\theta \left\{ \widehat{h}_l(t) \leq h_0(t;\theta) \leq \widehat{h}_u(t), \forall t \right\} \\ = \; & \alpha. \end{aligned}$$

Thus, the test procedures are conservative. There may even be a large gap between the two sides of the above inequality, in which case the test cannot have good power. There appears to be some room for improvement.

3 Modified graphical tests

Since the null hypothesis $h(t) = h_0(t;\theta)$ corresponds to a parametric model for the function h, we can obtain a parametric estimator of h under this hypothesis and compare it with an unconstrained, nonparametric estimator. If $\widehat{\theta}$ is a consistent estimator of θ under H_0 and $\widehat{h}$ is a consistent estimator of h under H_1, then the difference $\widehat{h}(t) - h_0(t;\widehat{\theta})$ should not be large when H_0 is true. The supremum of the absolute value of the difference process $\widehat{h}(t) - h_0(t;\widehat{\theta})$ can be used as a test statistic. In fact, this is a commonly used approach for obtaining a lack-of-fit test for a parametric model, based on a nonparametric estimator (Hart (1997)).

For the weighted versions of the three plots mentioned in Table 1, the function $h(\cdot)$ takes the following special forms

$$\begin{aligned} r^k(u) &= \Lambda_2^k \circ {\Lambda_1^k}^{-1}(u), \quad 0 \leq u \leq \Lambda^k(\tau), \\ \rho^k(t) &= \log \Lambda_2^k(t) - \log \Lambda_1^k(t), \quad 0 \leq t \leq \tau, \\ \Delta^k(t) &= (\Lambda_2^k(t) - \Lambda_1^k(t))/\Lambda_1^k(t), \quad 0 \leq t \leq \tau, \end{aligned}$$

where $\Lambda_i^k(t) = \int_0^t k(u) d\Lambda_i(u)$ for $i = 1, 2$, $k(\cdot)$ is the probability limit of the predictable weight function $K(\cdot)$ and τ is a pre-specified and suitably large time point such that $F_i(\tau) < 1$ for $i = 1, 2$. Nonparametric estimators of these functions can be obtained by replacing Λ_1^k and Λ_2^k by $\widehat{\Lambda}_1^K$ and $\widehat{\Lambda}_2^K$, respectively. On the other hand, the corresponding parametric forms under the null hypothesis (see Table 1) can be based on the estimator of θ given by $\widehat{\theta}_K = \widehat{\Lambda}_2^K(\tau)/\widehat{\Lambda}_1^K(\tau)$ (see Andersen (1983)). For the purpose of computation, τ may be chosen as a time point that is larger than the largest observed failure time in the combined sample. All such choices lead to the same value of $\widehat{\theta}_K$.

3.1 Convergence of plotted functions

We use the nonparametric estimators of $r^k(\cdot)$, $\rho^k(\cdot)$ and $\Delta^k(\cdot)$, defined by

$$\begin{aligned}
\widehat{r}^K(u) &= \widehat{\Lambda}_2^K \circ \widehat{\Lambda}_1^{K\,-1}(u), \quad 0 \le u \le \widehat{\Lambda}^K(\tau), \\
\widehat{\rho}^K(t) &= \left[\log \widehat{\Lambda}_2^K(t) - \log \widehat{\Lambda}_1^K(t)\right] I[\widehat{\Lambda}_1^K(t) > 0] I[\widehat{\Lambda}_2^K(t) > 0], \quad 0 \le t \le \tau, \\
\widehat{\Delta}^K(t) &= \left[(\widehat{\Lambda}_2^K(t) - \widehat{\Lambda}_1^K(t))/\widehat{\Lambda}_1^K(t)\right] I[\widehat{\Lambda}_1^K(t) > 0], \quad 0 \le t \le \tau,
\end{aligned}$$

where $\widehat{\Lambda}_1^{K\,-1}(u) = \inf\{t : t \ge 0, \widehat{\Lambda}_1^K(t) \ge u\}$. The three graphical tests, as suggested above, would be based on estimators of the difference function $\widehat{h}(t) - h_0(t;\widehat{\theta})$ in the three cases, namely, $\widehat{r}^K(u) - \widehat{\theta}_K u$, $\widehat{\rho}^K(t) - \widehat{\rho}^K(\tau)$ and $\widehat{\Delta}^K(t) - \widehat{\Delta}^K(\tau)$.

Let $u_\tau = \Lambda_1^k(\tau)$, $M = n_1 n_2/(n_1+n_2)$, $\widehat{\eta}_1 = n_1/(n_1+n_2)$ and $\widehat{\eta}_2 = n_2/(n_1+n_2)$. The following theorem is an adaptation of some results obtained in Dabrowska et al. (1989) and Dabrowska et al. (1992) to the case of weighted hazard functions. The convergence mentioned in this theorem takes place in the space $\mathrm{D}[0,u_\tau]$ ($\mathrm{D}[0,\tau]$) of right continuous functions with finite left hand limits over $[0,u_\tau]$ ($[0,\tau]$), equipped with the Skorohod topology. The limiting processes are time-transformed versions of the standard Brownian motion (i.e., Gaussian processes with zero mean, independent increments and variances equal to the time parameters).

THEOREM 3.1. *Let $M \to \infty$ in such a way that $\widehat{\eta}_1 \to \eta$ and $\widehat{\eta}_2 \to 1-\eta$ for some $\eta \in (0,1)$. Then we have*

(i) $\sqrt{M}\left(\widehat{r}^K(u) - r^k(u)\right)$ *converges weakly in the space $D[0,u_\tau]$ to a time-transformed Brownian motion with variance function $g_r(\cdot)$ defined over $[0,u_\tau]$ as*

$$\begin{aligned}
g_r(u) &= \eta \int_0^{\Lambda_1^{k\,-1}(u)} \frac{k(t)}{\overline{F}_2(t)\overline{F}_2^c(t)} d\Lambda_2^k(t) \\
&\quad + (1-\eta)\left(\frac{\lambda_2\left(\Lambda_1^{k\,-1}(u)\right)}{\lambda_1\left(\Lambda_1^{k\,-1}(u)\right)}\right)^2 \int_0^{\Lambda_1^{k\,-1}(u)} \frac{k(t)}{\overline{F}_1(t)\overline{F}_1^c(t)} d\Lambda_1^k(t),
\end{aligned}$$

provided $\Lambda_1(\cdot)$ and $\Lambda_2(\cdot)$ have derivatives $\lambda_1(\cdot)$ and $\lambda_2(\cdot)$, respectively, and $\lambda_1(\cdot)$ is continuous and strictly positive over $[0,\tau]$.

(ii) $\sqrt{M}\,\widehat{\Lambda}_2^K(t)\left(\widehat{\rho}^K(t) - \rho^k(t)\right)$ *converges weakly in the space $D[0,\tau]$ to a time-transformed Brownian motion with variance function $g_\rho(\cdot)$ defined over $[0,\tau]$ as*

$$g_\rho(t) = \left[\eta \int_0^t \frac{k(s)}{\overline{F}_2(s)\overline{F}_2^c(s)} d\Lambda_2^k(s) + (1-\eta)\frac{\Lambda_2^{k\,2}(t)}{\Lambda_1^{k\,2}(t)} \int_0^t \frac{k(s)}{\overline{F}_1(s)\overline{F}_1^c(s)} d\Lambda_1^k(s)\right],$$

(iii) $\sqrt{M}\,\widehat{\Lambda}_1^K(t)\left(\widehat{\Delta}^K(t) - \Delta^k(t)\right)$ *converges weakly in the space $D[0,\tau]$ to a time-transformed Brownian motion with variance function $g_\Delta(\cdot)$ defined over $[0,\tau]$*

as

$$g_\Delta(t) = \left[\eta \int_0^t \frac{k(s)}{\overline{F}_2(s)\overline{F}_2^c(s)} d\Lambda_2^k(s) + (1-\eta)\frac{{\Lambda_2^k}^2(t)}{{\Lambda_1^k}^2(t)} \int_0^t \frac{k(s)}{\overline{F}_1(s)\overline{F}_1^c(s)} d\Lambda_1^k(s)\right].$$

Under the hypothesis of proportional hazards, the variance functions simplify as follows:

$$\begin{aligned} g_r(u) &= g\left({\Lambda_1^k}^{-1}(u)\right), \\ g_\rho(t) &= g_\Delta(t) = g(t), \end{aligned}$$

where

$$g(t) = \eta \int_0^t \frac{k(s)}{\overline{F}_2(s)\overline{F}_2^c(s)} d\Lambda_2^k(s) + (1-\eta)\theta^2 \int_0^t \frac{k(s)}{\overline{F}_1(s)\overline{F}_1^c(s)} d\Lambda_1^k(s), \tag{1}$$

and a consistent estimator of $g(t)$ is

$$\widehat{g}(t) = M \int_0^t \frac{K(s)}{Y_2(s)} d\widehat{\Lambda}_2^K(s) + M\widehat{\theta}_K^2 \int_0^t \frac{K(s)}{Y_1(s)} d\widehat{\Lambda}_1^K(s). \tag{2}$$

Empirical versions of the variance functions given in Theorem 3.1 can be used to produce a transformation of the three processes to the Brownian bridge over $[0, 1/2]$, and the supremum of the Brownian bridge can be used (as argued in Anderson and Borgan (1985)) as a pivot for generating asymptotic confidence bands for the RTF, the LCHD and the RCHD. These, in turn, lead to the weighted versions of the conventional tests, mentioned in Section 2.

3.2 Convergence of difference functions

We now turn to the difference functions $G^k(u) = r^k(u) - \theta u$, $Q^k(t) = \rho^k(t) - \log\theta$ and $R^k(t) = \Delta^k(t) - (\theta - 1)$, estimated by $\widehat{G}^K(u) = \widehat{r}^K(u) - \widehat{\theta}_K u$, $\widehat{Q}^K(t) = \widehat{\rho}^K(t) - \widehat{\rho}^K(\tau)$ and $\widehat{R}^K(t) = \widehat{\Delta}^K(t) - \widehat{\Delta}^K(\tau)$, respectively. The difference functions simplify to 0 under H_0.

THEOREM 3.2. *Let $M \to \infty$ in such a way that $\widehat{\eta}_1 \to \eta$ and $\widehat{\eta}_2 \to 1 - \eta$ for some $\eta \in (0, 1)$, and B be the standard Brownian motion. Then the following convergence results hold under the hypothesis of proportional hazards.*

(i) If $\Lambda_1(\cdot)$ is strictly increasing over $[0, \tau]$ and has a continuous derivative, then the process $\sqrt{M}\widehat{G}^K(u)$ converges weakly in the space $D[0, u_\tau]$ to a mean zero Gaussian process

$$B\left(g({\Lambda_1^k}^{-1}(u))\right) - \frac{u}{u_\tau} B\left(g({\Lambda_1^k}^{-1}(u_\tau))\right),$$

whose variance function

$$\sigma_G^2(u) = g\left({\Lambda_1^k}^{-1}(u)\right) + \left(\frac{u}{u_\tau}\right)^2 g(\tau) - 2\left(\frac{u}{u_\tau}\right) g\left({\Lambda_1^k}^{-1}(u)\right)$$

is consistently estimated by

$$\widehat{\sigma}_G^2(u) = \widehat{g}(\widehat{\Lambda}_1^{K\,-1}(u)) + \left(\frac{u}{\widehat{\Lambda}_1^K(\tau)}\right)^2 \widehat{g}(\tau) - 2\left(\frac{u}{\widehat{\Lambda}_1^K(\tau)}\right)\widehat{g}(\widehat{\Lambda}_1^{K\,-1}(u)),$$

the functions $g(\cdot)$ and $\widehat{g}(\cdot)$ being as defined in (1) and (2), respectively.

(ii) The process $\sqrt{M}\,\widehat{\Lambda}_2^K(t)\,\widehat{Q}^K(t)$ converges weakly in the space $D[0,\tau]$ to a mean zero Gaussian process

$$B(g(t)) - \left(\frac{\Lambda_1^k(t)}{\Lambda_1^k(\tau)}\right) B(g(\tau)),$$

whose variance function

$$\sigma_Q^2(t) = g(t) + \left(\frac{\Lambda_1^k(t)}{\Lambda_1^k(\tau)}\right)^2 g(\tau) - 2\left(\frac{\Lambda_1^k(t)}{\Lambda_1^k(\tau)}\right) g(t),$$

is consistently estimated by

$$\widehat{\sigma}_Q^2(t) = \widehat{g}(t) + \left(\frac{\widehat{\Lambda}_1^K(t)}{\widehat{\Lambda}_1^K(\tau)}\right)^2 \widehat{g}(\tau) - 2\left(\frac{\widehat{\Lambda}_1^K(t)}{\widehat{\Lambda}_1^K(\tau)}\right)\widehat{g}(t).$$

(iii) The process $\sqrt{M}\,\widehat{\Lambda}_1^K(t)\widehat{R}^K(t)$ converges weakly in the space $D[0,\tau]$ to a mean zero Gaussian process

$$B(g(t)) - \left(\frac{\Lambda_1^k(t)}{\Lambda_1^k(\tau)}\right) B(g(\tau)),$$

whose variance function

$$\sigma_R^2(t) = g(t) + \left(\frac{\Lambda_1^k(t)}{\Lambda_1^k(\tau)}\right)^2 g(\tau) - 2\left(\frac{\Lambda_1^k(t)}{\Lambda_1^k(\tau)}\right) g(t),$$

is consistently estimated by

$$\widehat{\sigma}_R^2(t) = \widehat{g}(t) + \left(\frac{\widehat{\Lambda}_1^K(t)}{\widehat{\Lambda}_1^K(\tau)}\right)^2 \widehat{g}(\tau) - 2\left(\frac{\widehat{\Lambda}_1^K(t)}{\widehat{\Lambda}_1^K(\tau)}\right)\widehat{g}(t).$$

3.3 A pivot for asymptotic inference

Since the function $g(\cdot)$ is continuous and monotone increasing over $[0,\tau]$, the scaled variance functions $\sigma_G^2(u)/g(\tau)$, $\sigma_Q^2(t)/g(\tau)$ and $\sigma_R^2(t)/g(\tau)$ have the common form

$$v^2 + f(v) - 2vf(v), \tag{3}$$

for a strictly increasing and bijective map $f(v) : [0,1] \to [0,1]$. The function given in (3) is not monotone; rather it assumes the value 0 at the end-points 0 and 1. Therefore, the usual construction of asymptotic confidence band based on a suitable transformation of the Brownian motion to the standard Brownian bridge over $[0, 1/2]$ (see Dabrowska et al. (1989), Andersen et al. (1992)) would not work. On the other hand, the variance function resembles that of a Brownian bridge over $[0, 1]$ in the sense that the variance is 0 at both end-points and is positive in between. However, it does not have the form $\psi(v)(1 - \psi(v))$, which would have enabled us to model it as a Brownian bridge with time transformation ψ.

Note that the limiting processes in Theorem 3.2, scaled by the factor $1/\sqrt{g(\tau)}$, can be written in the common form as

$$B(f(v)) - vB(1) \tag{4}$$

This could have been a Brownian bridge if f had been the identity function.

The above discussion leads us to the following.

THEOREM 3.3. *Let $M \to \infty$ in such a way that $\widehat{\eta}_1 \to \eta$ and $\widehat{\eta}_2 \to 1 - \eta$ for some $\eta \in (0, 1)$, and F_B denote the distribution of $\sup_{0 \le v \le 1} |B(f(v)) - vB(1)|$. The following convergence results hold under the hypothesis of proportional hazards.*

(i) If $\Lambda_1(\cdot)$ is strictly increasing over $[0, \tau]$ and has a continuous derivative, then the distribution of the supremum

$$\sup_{0 \le u \le \widehat{\Lambda}_1^K(\tau)} \sqrt{M}\, \frac{|\widehat{G}^K(u)|}{\sqrt{\widehat{g}(\tau)}}$$

converges to F_B.

(ii) The distribution of the supremum

$$\sup_{0 \le t \le \tau} \sqrt{M}\, \widehat{\Lambda}_2^K(t) \frac{|\widehat{Q}^K(t)|}{\sqrt{\widehat{g}(\tau)}}$$

converges to F_B.

(iii) The distribution of the supremum

$$\sup_{0 \le t \le \tau} \sqrt{M}\, \widehat{\Lambda}_1^K(t) \frac{|\widehat{R}^K(t)|}{\sqrt{\widehat{g}(\tau)}}$$

converges to F_B.

In the present circumstances, we can use an estimate of f from the available data and simulate the distribution of F_B.

3.4 Construction of graphical tests

Suppose that b_α is such that $\overline{F}_B(b_\alpha) = \alpha$ for a suitable level α, where F_B is as defined in Theorem 3.3. It follows from Theorem 3.3 that, under the PH hypothesis, the line $\widehat{\theta}_K u$ through origin would be sandwiched between the graphs of $\widehat{r}^K(u) \pm \frac{b_\alpha\sqrt{\widehat{g}(\tau)}}{\sqrt{M}}$ with asymptotic coverage probability $1-\alpha$, and the horizontal lines at levels $\widehat{\rho}^K(\tau)$ and $\widehat{\Delta}^K(\tau)$ are sandwiched between the pairs of curves $\widehat{\rho}^K(t) \pm \frac{b_\alpha\sqrt{\widehat{g}(\tau)}}{\sqrt{M}\widehat{\Lambda}_2^K(t)}$ and $\widehat{\Delta}^K(t) \pm \frac{b_\alpha\sqrt{\widehat{g}(\tau)}}{\sqrt{M}\widehat{\Lambda}_1^K(t)}$, respectively, with asymptotic coverage probability $1-\alpha$. We refer to the ranges $\widehat{r}^K(u) \pm \frac{b_\alpha\sqrt{\widehat{g}(\tau)}}{\sqrt{M}}$, $\widehat{\rho}^K(t) \pm \frac{b_\alpha\sqrt{\widehat{g}(\tau)}}{\sqrt{M}\widehat{\Lambda}_2^K(t)}$ and $\widehat{\Delta}^K(t) \pm \frac{b_\alpha\sqrt{\widehat{g}(\tau)}}{\sqrt{M}\widehat{\Lambda}_1^K(t)}$ as level-α asymptotic *acceptance bands* for the graphs of $\widehat{\theta}_K u$, $\widehat{\rho}^K(\tau)$ and $\widehat{\Delta}^K(\tau)$, respectively.

According to our modified version of the three graphical tests, the PH hypothesis is rejected if the above parametrically estimated curve does not lie entirely within the corresponding acceptance band. The p-value of the test is the smallest level that does not permit the acceptance band to completely contain the parametric estimate.

It may be noted that the acceptance band for the RTF plot has constant width. For all the plots, the width shrinks to zero in probability as $M \to \infty$ at all points (except for the left end-points of the LCHD and the RCHD plots).

4 Simulation Study of Graphical Tests

We evaluate the graphical tests through censored survival data sets simulated from the Weibull distribution with survival function $\exp(-\gamma t^\delta)$ (denoted in this section by Weibull(γ, δ)), having scale parameter γ and shape parameter δ. The censoring distribution is chosen as exponential with mean $1/\lambda$ (denoted in this section by $\exp(\lambda)$).

4.1 Illustration through a single simulation run

We generate data separately from the PH and a model with bathtub shaped ratio of hazard rates. Note that when the hazard ratio is increasing, the RTF is convex, and when the hazard ratio is decreasing, the RTF is concave. If the hazard ratio has the bathtub shape (i.e., decreasing initially and increasing afterwards), then the RTF has an inverse S-shape (i.e., concave initially and convex afterwards).

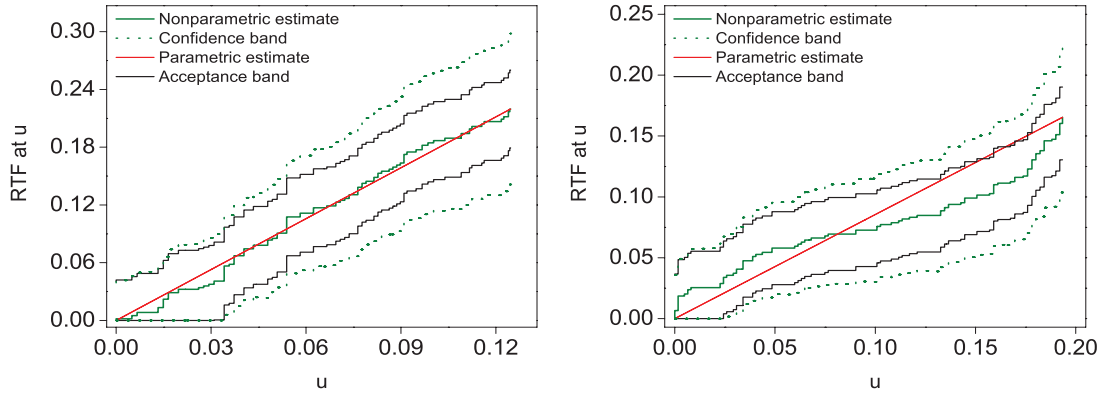
For the PH model, we choose F_1, F_2, F_1^c and F_2^c as Weibull$(1,2)$, Weibull$(2,2)$, $\exp(0.335)$ and $\exp(0.46)$, respectively. For the bathtub hazard ratio (BHR) model, we choose F_1, F_1^c and F_2^c as $\exp(3.78)$, $\exp(1.25)$ and $\exp(0.85)$, respectively, and $F_2 = \left[1 - \exp\left(1 - e^{t^2}\right)\right]^{0.3}$. In all the cases, the parameters are chosen to achieve about 25% censored observations in each sample. The sample size for each population (n_1 as well as n_2) is chosen as 150. We choose the weight function as $K(t) = Y_1(t)Y_2(t)[(n_1+n_2)(Y_1(t)+Y_2(t))]^{-1}$ (see Gill and Schumacher (1987)).

For computing b_α, it is necessary to simulate a number of sample paths from the process $B(f(v)) - vB(1)$, where $B(\cdot)$ is a standard Brownian motion over $[0, 1]$. We use the following approximation over the grid i/n_b, $i = 0, 1, \ldots, n_b$:

$$\sum_{j=1}^{i} \sqrt{\widehat{f}(v_j) - \widehat{f}(v_{j-1})}\, z_j - v_i \sum_{j=1}^{n_b} \sqrt{\widehat{f}(v_j) - \widehat{f}(v_{j-1})}\, z_j,$$

where $z_1, z_2, \ldots, z_{n_b}$ are independent standard normal variates. We use $n_b = 500$ and 2000 sample paths. Since the same $\widehat{f}(\cdot)$ is used in every run, a small adjustment of b_α is necessary to achieve the nominal size. Adjustment factor determined by simulation of the RTF statistic happens to be 1.012589.

Figure 1 exhibits plots of the parametric and nonparametric estimators ($\widehat{\theta}_K u$ and $\widehat{\Lambda}_2^K \circ \widehat{\Lambda}_1^{K\,-1}(u)$, respectively) of the RTF for simulated data sets from the PH and the BHR models, together with the 95% confidence band for $\Lambda_2^k \circ \Lambda_1^{k\,-1}(u)$ and the 95% acceptance band for $\widehat{\theta}_K u$. The plot on the right panel reveals an approximately inverse S-shape of the estimated RTF, indicating the bathtub shape of the hazard ratio. The conventional test based on confidence bands (described in Section 2.2) is unable to reject the PH hypothesis for either of the two data sets, since a straight line through origin can easily be drawn through the 95% confidence bands in both cases. In contrast, the parametric estimate of the RTF fits into its acceptance band in the PH case, but does not fit in the BHR case. Thus, at the 5% level, the proposed test (described in Section 3.4) correctly accepts the PH hypothesis for the data set generated from the PH model and correctly rejects it for the data set generated from the BHR model. The p-values of the test are 0.756 and 0.023 for the PH and BHR cases, respectively.

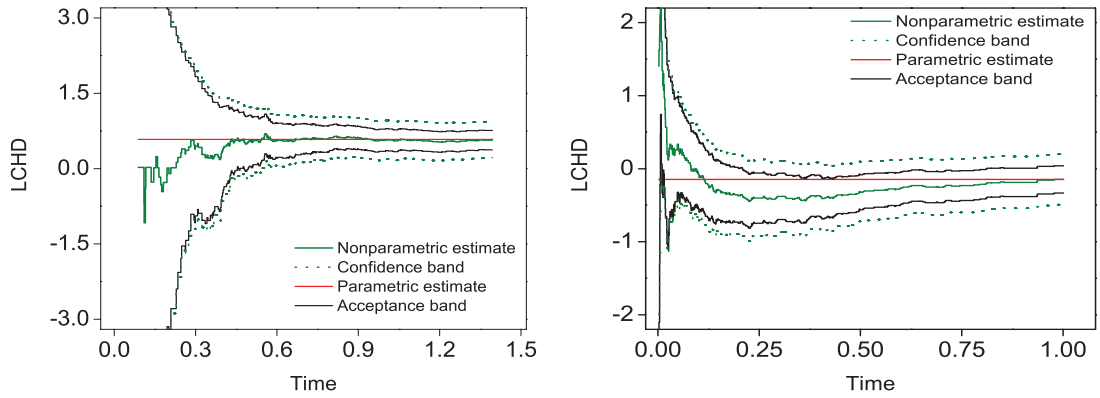


(a) PH model (b) BHR model

Figure 1: Plots of parametric and nonparametric estimates of RTF, together with 95% asymptotic acceptance bands and 95% asymptotic confidence bands for data simulated from PH and BHR models.

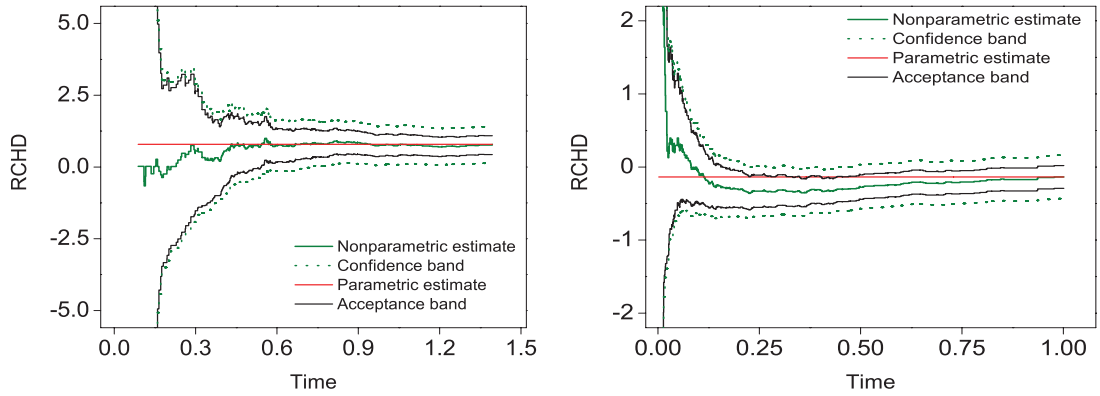
The LCHD and the RCHD plots for the two data sets are shown in Figures 2 and 3, respectively. The horizontal lines representing parametric estimates of LCHD and RCHD fit into the respective 95% acceptance bands in the PH case, but do not

fit there in the BHR case. In contrast, horizontal lines can easily be passed through the nonparametric confidence bands of LCHD and RCHD in the PH case and also in the BHR case. Therefore, the proposed graphical test correctly accepts the null hypothesis for the PH data set and rejects it for the BHR data set, while the existing graphical tests are unable to reject the null hypothesis for BHR data set. The p-values of the proposed test based on the LCHD and the RCHD plots are, respectively, 0.914 and 0.724 for the PH data set, and 0.001 and 0.017 for the BHR data set.



(a) PH model (b) BHR model

Figure 2: Plots of parametric and nonparametric estimates of LCHD, together with 95% asymptotic acceptance bands and 95% asymptotic confidence bands for data simulated from PH and BHR models.



(a) PH model (b) BHR model

Figure 3: Plots of parametric and nonparametric estimates of RCHD, together with 95% asymptotic acceptance bands and 95% asymptotic confidence bands for data simulated from PH and BHR models.

4.2 Systematic study through multiple simulation runs

We run multiple simulations by generating uncensored survival data separately from PH, IHR and decreasing hazard ratio (DHR) models. For the PH case, we choose F_1 and F_2 as Weibull$(1, 2)$ and Weibull$(2, 2)$, respectively. For the IHR case, we choose F_1 and F_2 as Weibull$(2.73, 1.25)$ and Weibull$(2, 2)$, respectively. For the DHR case, we choose F_1 and F_2 as Weibull$(2, 0.5)$ and Weibull$(3.03, 0.25)$, respectively. For the alternative models, the parameters are so chosen that the two groups have the same 0.99th quantile. The weight function continues to be $K(t) = Y_1(t)Y_2(t)[(n_1 + n_2)(Y_1(t) + Y_2(t))]^{-1}$.

Results reported in this section are based on 20,000 simulation runs with level of significance 0.05. This corresponds to a standard error of 0.0015 (approximately) on the achieved size $\widehat{\alpha}$, so that a 95% confidence interval for the effective α is approximately $\widehat{\alpha} \pm 0.003$. As in the case of the single simulation run reported in section 4.1, we use uniform grid of size 500 and simulate 2000 sample paths in order to generate the distribution of F_B (described in Theorem 3.3) for computing b_α, separately for every simulation run. We use the same adjustment factor of b_α for the empirical results, and consider equal sample sizes of 25, 50, 100, 200 and 500 in the two groups.

Table 2 shows the empirical sizes of the proposed and the conventional graphical tests (described in Sections 3.4 and 2.2, respectively) based on RTF, LCHD and RCHD plots for varying sample sizes at the level of significance 0.05. It is seen that the existing graphical tests are very conservative. On the other hand, the empirical sizes of the proposed graphical tests based on the RTF and the RCHD plot appear to be reasonable. The test based on the LCHD plot has reasonable size at sample size 500, though the achieved size is considerably higher than the nominal value for smaller sample sizes.

Table 2

Achieved sizes of the existing and proposed graphical tests based on RTF, LCHD and RCHD plots, at nominal level of significance 0.05 (20,000 simulation runs)

Size	RTF		LCHD		RCHD	
$n_1 = n_2$	Existing	Proposed	Existing	Proposed	Existing	Proposed
25	0.000	0.048	0.002	0.146	0.000	0.047
50	0.000	0.051	0.003	0.125	0.000	0.050
100	0.000	0.049	0.002	0.094	0.000	0.049
200	0.000	0.051	0.001	0.071	0.000	0.051
500	0.000	0.050	0.000	0.050	0.000	0.050

Simulation studies with various degrees of censoring (not reported here) indicate that for heavier (up to 75%) censoring, larger sample sizes are needed for the empirical size to approach the nominal value, while the pattern observed for light censoring is similar to that for no censoring.

Tables 3 and 4 summarize the power results for data generated from the IHR and the DHR models, respectively, in the case of no censoring. The proposed tests based on the three plots are found to have much higher power compared to that of the corresponding conventional tests for any sample size.

Table 3

Achieved powers of the existing and proposed graphical tests based on RTF, LCHD and RCHD plots, when data are generated from the IHR model (20,000 simulation runs)

Size	RTF		LCHD		RCHD	
$n_1 = n_2$	Existing	Proposed	Existing	Proposed	Existing	Proposed
25	0.000	0.360	0.000	0.006	0.000	0.360
50	0.000	0.661	0.000	0.033	0.000	0.661
100	0.002	0.929	0.000	0.546	0.002	0.929
200	0.214	0.998	0.000	0.973	0.225	0.998
500	0.994	1.000	0.876	1.000	0.994	1.000

Table 4

Achieved powers of the existing and proposed graphical tests based on RTF, LCHD and RCHD plots, when data are generated from the DHR model (20,000 simulation runs)

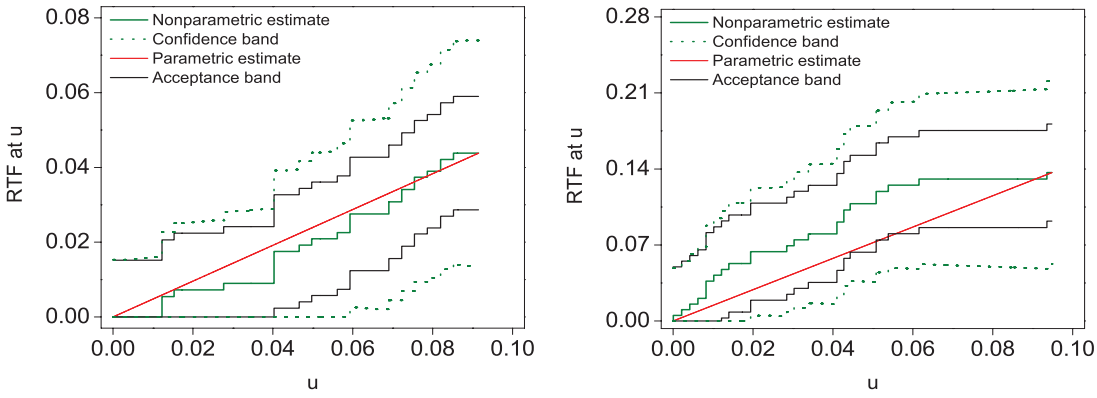
Size	RTF		LCHD		RCHD	
$n_1 = n_2$	Existing	Proposed	Existing	Proposed	Existing	Proposed
25	0.029	0.603	0.213	0.881	0.047	0.587
50	0.214	0.912	0.586	0.986	0.290	0.910
100	0.706	0.997	0.914	1.000	0.762	0.997
200	0.990	1.000	0.999	1.000	0.993	1.000
500	1.000	1.000	1.000	1.000	1.000	1.000

5 Data Analytic Illustrations

We consider a data set consisting of post operative survival times (in days) of 205 patients of malignant melanoma. In the period 1962-77, the patients were treated with tumor operation at the Department of Plastic Surgery, University Hospital of Odense, Denmark. Andersen et al. (1992) (p.709) provide the data set with detailed analysis throughout the book. In our analysis, the patients are categorized into two groups according to their sex. Group 1 consists of $n_1 = 79$ male patients, out of which 29 are observed to die from the disease, while Group 2 contains $n_2 = 126$ female patients with 28 observed deaths. Through analytical tests, Andersen et al. (1992) concluded that these two groups have proportional hazard rates.

As another example, we consider data arising from a multicenter trial of acute leukemia patients prepared for bone marrow transplantation with a radiation free conditioning regimen. Details of the study are given in Copelan et al. (1991). The data set is given in Klein and Moeschberger (1997) (pp. 464-467). The data consists of times (in days) to death of 137 bone marrow transplantation patients. We compare the survival times of the patients receiving no MTX (Group 1) with those receiving MTX as a graft-versus-host prophylactic (Group 2). Group 1 consists of 97 patients with 56 observed deaths within the period of study, while in group 2, there are 40 patients with 25 observed deaths.

Figure 4 exhibits the nonparametric estimates of RTF for the malignant melanoma survival data (part (a)) and the bone marrow transplantation data (part (b)), along with the parametric estimates represented by the straight lines through origin having slopes ($\widehat{\theta}_K$) 0.51 and 1.44, respectively. The 95% acceptance bands for the parametric estimates (thin curves) and the 95% asymptotic confidence bands for the RTF (dotted curves) are also shown.

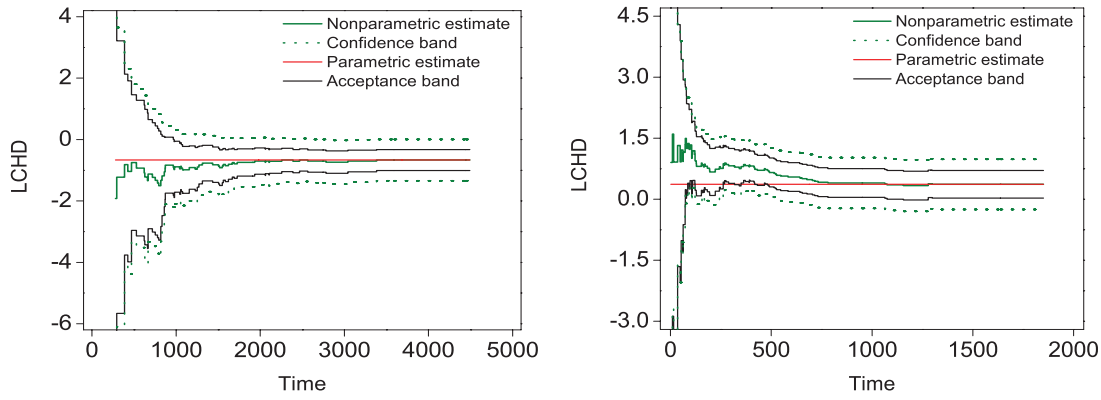


(a) Melanoma Survival Data (b) Bone Marrow Transplantation Data

Figure 4: Plots of parametric and nonparametric estimates of RTF, together with 95% asymptotic acceptance bands and 95% asymptotic confidence bands for two real data sets.

In Figure 4(a), the parametric estimate of RTF fits in its acceptance band as well as in the asymptotic confidence band of the plot of $r^k(\cdot)$. Thus, neither the conventional test, nor the proposed test, rejects the null hypothesis. This conclusion is in line with the findings of Andersen et al. (1992). In Figure 4(b), the nonparametric estimate of the RTF exhibits a concave pattern, suggesting a decrease hazard ratio between Groups 1 and 2. However, many straight lines through origin easily fit into the asymptotic confidence bands, making it impossible for the conventional test to reject the hypothesis of proportional hazards. In contrast, the parametric estimate does not fit in the acceptance band, leading to the rejection of this hypothesis by the proposed test, at level of significance 0.05. This conclusion supports the findings of Klein and Moeschberger (1997) based on analytical tests. The advantage of this particular graphical test over the analytical tests is that, in addition to formal rejection of the null hypothesis, one also gets an indication of the nature of departure of the underlying relative trend function from the null hypothesis. The test has the p-values of 0.313 and 0.034 for melanoma and bone marrow transplantation survival data, respectively.

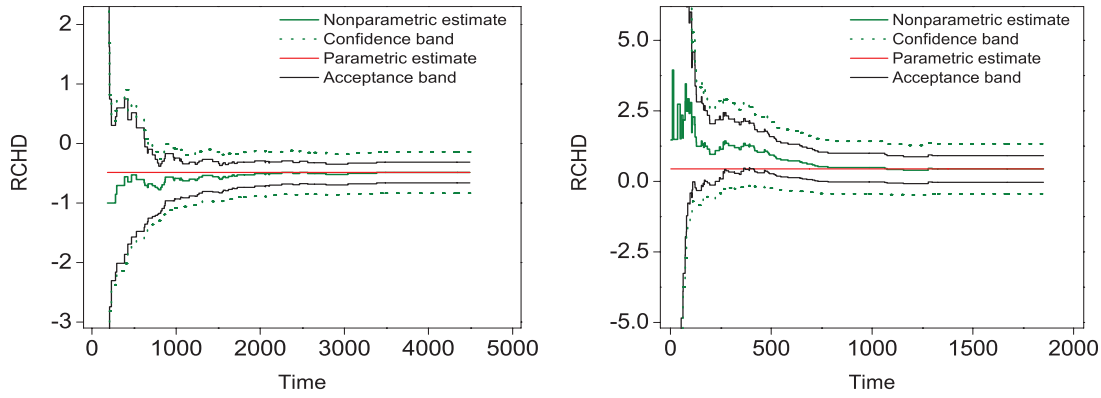
Figures 5 and 6 exhibit the LCHD and RCHD plots, respectively, for the two data sets mentioned above. In the case of the melanoma survival data, the parametric estimates of LCHD and RCHD fit into their respective acceptance bands, while the asymptotic confidence bands leave enough room for drawing horizontal straight lines through them. Thus, none of the tests reject the null hypothesis of proportional hazards. The p-values of the proposed test based on the LCHD and RCHD plots are 0.746 and 0.237, respectively. In the case of the bone marrow transplantation data, the nonparametric estimates of the LCHD and the RCHD exhibit a decreasing pattern, suggesting a decreasing ratio of cumulative hazards between Groups 1 and 2.



(a) Melanoma Survival Data (b) Bone Marrow Transplantation Data

Figure 5: Plots of parametric and nonparametric estimates of LCHD, together with 95% asymptotic acceptance bands and 95% asymptotic confidence bands for two real data sets.

The asymptotic confidence bands of LCHD and RCHD admit many horizontal straight lines, and the conventional tests based on these bands are unable to reject the null hypothesis. In contrast, the parametric estimates of LCHD and RCHD do not fit into the respective 95% acceptance bands, leading to the rejection of this hypothesis by the proposed tests. The proposed tests based on LCHD and RCHD plots have the p-values of 0.006 and 0.034, respectively, for this data set.



(a) Melanoma Survival Data (b) Bone Marrow Transplantation Data

Figure 6: Plots of parametric and nonparametric estimates of RCHD, together with 95% asymptotic acceptance bands and 95% asymptotic confidence bands for two real data sets.

These data analytic examples illustrate how the proposed graphical methods are able to combine the formality and power of an analytical test with the descriptive value of a plot (e.g., the indication of decreasing hazard ratio in the case of Figure 4(b), and decreasing cumulative hazard ratio in the cases of Figures 5(b) and 6(b)).

6 Concluding Remarks

Instead of using the supremum norm of the limiting Gaussian process (4) as a pivot for asymptotic inference, one could also use the supremum norm of a weighted version of the process with general weights, as in Miller and Siegmund (1982). If the inverse

of the estimated standard deviation function is used, that would produce what are called EP (equally probable) bands.

The graphical techniques mentioned in this paper easily extend to the comparison of the intensity functions of the two groups under the multiplicative intensity model of life history analysis (Anderson and Borgan, 1985), which include, for instance, comparison of the cause-specific hazards in a competing risks situation. The confidence bounds used by Sahoo and Sengupta (2011) in multivariate versions of the RTF, the LCHD and the RCHD plots can also be replaced by acceptance bands.

The idea of using the difference between nonparametric and parametric estimates of a function of two distributions may be used for constructing other graphical tests. For instance, a graphical two-sample test can be constructed on the basis of the two-sample Q-Q plot. The asymptotic arguments for such a test would have to be based on the theory of empirical processes (see Shorak and Wellner (1986)).

As far as the RTF plot is concerned, there is no reason to believe that the specified guidelines are any better than those obtained by interchanging the axes. The latter would have horizontal, rather than vertical, jumps. In this connection, it may be recalled that the human eye is equally sensitive to horizontal and vertical separations of two curves Cleveland and McGill (1984). Sengupta Sengupta (1996) considered what is essentially the perpendicular distance between the parametric and nonparametric estimates of the RTF plots. Acceptance bands corresponding to the perpendicular distance can be constructed along the lines of the bands provided in this paper.

Appendix: Proofs

Proof of Theorem 3.2. The difference process can be written as $\sqrt{M}\{\widehat{r}^K(u)-r^k(u)\}-u\sqrt{M}\{\widehat{\theta}_K-\theta\}$. By using arguments similar to those used by Dabrowska et al. (1989) in the case of unweighted plots, it can be shown that under H_0, the first part of the above difference process, $\sqrt{M}\{\widehat{r}^K(u)-r^k(u)\}$, has asymptotically the same distribution as

$$\sqrt{\widehat{\eta}_1}W_2^{(n_2)}\left(\widehat{\Lambda}_1^{K\,-1}(u)\right)-\theta\sqrt{\widehat{\eta}_2}W_1^{(n_1)}\left(\widehat{\Lambda}_1^{K\,-1}(u)\right),$$

where

$$W_i^{(n_i)}(t)=\sqrt{n_i}\left(\widehat{\Lambda}_i^K(t)-\Lambda_i^k(t)\right),\ i=1,2.$$

It was shown in Gill and Schumacher (1987) that for $i=1,2$, $W_i^{(n_i)}(t)$ converges in distribution to a zero mean Gaussian process $W_i(t)$, having variance function

$$\int_0^t \frac{k(s)d\Lambda_i^k(s)}{\overline{F}_i(s)\overline{F}_i^c(s)}.$$

Therefore, the first part of the difference process converges to the mean zero Gaussian process

$$\sqrt{\eta_1}\,W_2\left(\Lambda_1^{k\,-1}(u)\right)-\theta\sqrt{\eta_2}\,W_1\left(\Lambda_1^{k\,-1}(u)\right)$$

in the space $D[0, u_\tau]$. The second part, under H_0, can be written as

$$u\sqrt{M}\{\widehat{\theta}_K - \theta\} = \frac{u}{\Lambda_1^k(\tau)}\sqrt{M}\{\widehat{r}^K\left(\Lambda_1^k(\tau)\right) - r^k\left(\Lambda_1^k(\tau)\right)\}.$$

Thus, the difference process is a continuous function of the process $\sqrt{M}\{\widehat{r}^K(u) - r^k(u)\}$. The stated result follows from an application of the continuous mapping theorem.

The proof of Part (iii) follows along similar lines, by adapting the arguments of Dabrowska et al. (1989) to the case of weighted plots. The proof of part (ii) follows from an application of the delta method to the latter result and using the fact that under H_0,

$$\frac{\Lambda_2^k(t)}{\Lambda_2^k(\tau)} = \frac{\Lambda_1^k(t)}{\Lambda_1^k(\tau)}.$$

□

References

Andersen PK. Testing goodness-of-fit of Cox's regression and life model. *Biometrics* 1982; **38** : 67-77.

Andersen PK. Comparing survival distributions via hazard ratio estimates. *Scandinavian Journal of Statistics* 1983; **10** : 77-85.

Anderson PK, Borgan Ø. Counting process models for life history data: A review. *Scandinavian Journal of Statistics* 1985; **12** : 97-158.

Andersen PK, Borgan Ø, Gill RD, Keiding N. *Statistical Models Based on Counting Processes*. Springer Verlag: New York, 1992.

Arjas E. A graphical method for assessing goodness-of-fit in Cox's proportional hazard model. *Journal of the American Statistical Association* 1988; **83** : 204-212.

Bie O, Borgan Ø, Lestøl K. Confidence intervals and confidence bands for the cumulative hazard rate function and their small sample properties. *Scandinavian Journal of Statistics* 1987; **14** : 221-233.

Breslow NE, Edler L, Berger J. A two-sample censored-data rank test for acceleration. *Biometrics* 1984; **40** : 1049-1062.

Cleveland WS, McGill R. Graphical perception: theory, experimentation and application to the development of graphical methods. *Journal of the American Statistical Association* 1984; **79** : 531-554.

Copelan EA, Biggs JC, Thompson JM, Crilley P, Szer J, Klein JP, Kapoor N, Avalos BR, Cunningham I, Atkinson K, Downs K, Harmon GS, Daly MB, Brodsky I, Bulova SI, Tutschka PJ. Treatment for acute Myelocytic leukemia with allogeneic bone marrow transplantation following preparation with Bu/Cy. *Blood* 1991; **78** : 838-843.

Cox DR. Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B* 1972; **34** : 187-202.

Dabrowska D, Doksum K, Feduska NJ, Husing R, Neville P. Methods for comparing cumulative hazard functions in a semi-parametric hazard model. *Statistics in Medicine* 1992; **11** : 1465-1476.

Dabrowska D, Doksum K, Song JK. Graphical comparison of cumulative hazards for two populations. *Biometrika* 1989; **76** : 763-773.

Deshpande JV, Sengupta D. Testing for the hypothesis of proportional hazards in two populations. *Biometrika* 1995; **82** : 251-261.

Gill RD, Schumacher M. A simple test of the proportional hazards assumption. *Biometrika* 1987; **74** : 289-300.

Hart JD. *Nonparametric Smoothing and Lack-of-Fit Tests*, Springer Series in Statistics. Springer: New York, 1997.

Kay R. Proportional hazards regression models and the analysis of censored survival data. *Applied Statistics* 1977; **26** : 227-237.

Klein JP, Moeschberger ML. *Survival Analysis: Techniques for Censored and Truncated Data*, Springer Series in Statistics for Biology and Health. Springer: New York, 1997.

Lee L, Pirie WR. A graphical method for comparing trends in series of events. *Communications in Statistics, Theory and Methods* 1981; **10** : 827-848.

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

Miller R, Siegmund D. Maximally selected chi-square statistics. *Biometrics* 1982; **38** : 1011-1016.

Sahoo S, Sengupta D. Some diagnostic plots and corrective adjustments for the proportional hazards regression model. *Journal of Computational and Graphical Statistics* 2011; **20** : 375-394.

Schumacher M. Two-Sample Tests of Cramér-Von Mises and Kolmogorov-Smirnov Type for Randomly Censored Data. *International Statistical Review* 1984; **52** : 263-281.

Sengupta D. Graphical tools for censored survival data. In *Analysis of Censored Data*, edited by HL Koul, JV Deshpande, IMS Lecture Notes **27**. Institute of Mathematical Statistics : Hayward, CA, USA, 1996; 193-217.

Sengupta D, Bhattacharjee A, Rajeev B. Testing for the proportionality of hazards in two samples against the increasing cumulative hazard ratio alternative. *Scandinavian Journal of Statistics* 1998; **25** : 637-647.

Sengupta D, Deshpande JV. Some results on the relative aging of two life distributions, *Journal of Applied Probability* 1994; **31** : 991-1003.

Shorak GR, Wellner JA. *Empirical Processes with Applications to Statistics*. Wiley: New York, 1986.

Therneau TM, Grambsch PM, Fleming TR. Martingale-based residuals for survival models. *Biometrika* 1990; **77** : 147-160.

Wei LJ. Testing goodness of fit for proportional hazards model with censored observations. *Journal of the American Statistical Association* 1984; **79** : 649-652.