Testing for the Hypothesis of Increasing Hazard Ratio in Two Samples

Technical Report No. ASU/2014/9

Dated: 25 August 2014

Shyamsundar Sahoo, Haldia Government College, Haldia

and

Debasis Sengupta, Indian Statistical Institute, Kolkata

Indian Statistical Institute

Applied Statistics Unit

Kolkata 700 108

Testing for the Hypothesis of Increasing Hazard Ratio in Two Samples

S. Sahoo[1] and D. Sengupta[2]

[1]Department of Statistics, Haldia Government College, Purba Medinipur, Haldia - 721657.
E-mail: sssahoo.stats@gmail.com

[2]Applied Statistics Unit, Indian Statistical Institute, 203 B.T. Road, Kolkata - 700108.
E-mail: sdebasis@isical.ac.in

Abstract

Tests designed to detect increasing hazard ratio against the proportional hazards hypothesis are generally consistent for other alternatives also. This article provides a test of the null hypothesis of increasing hazard ratio. The test is based on the separation between an empirical version of the relative trend function and its greatest convex minorant. The proportional hazards model, the least favorable null model for large samples, is used to produce simulation based cutoffs. A simulation study shows reasonable performance of the proposed test in small samples. The analytical test, together with a graphical version, is illustrated through two real life examples.

Key Words: Proportional hazards model; Increasing failure rate; Greatest convex minorant; Increasing hazard ratio; Two-sample problem.

1 Introduction

The proportional hazards (PH) assumption has been used widely for modeling and analysis of survival data. A test of this assumption is not only an important two-sample problem, it is also relevant as a diagnostic check for Cox's regression model (Cox, 1972). Several graphical and analytical tests have been proposed for this purpose; see Kay (1977), Lee and Pirie (1981), Wei (1984), Breslow et al. (1984), Gill and Schumacher (1987), Dabrowska et al. (1989), Dabrowska et al. (1992), Deshpande and Sengupta (1995), Sengupta et al. (1998) and Sahoo and Sengupta (2014).

In many real life situations, the hazard ratio of two samples is observed to vary with time. Pocock et al. (1982) observed the phenomenon of crossing hazards while studying the prognostic relevance in the long run in the treatment of breast cancer patients. Champlin et al. (1983) and Begg et al. (1984) found the superiority of a treatment over another only in the short term period. Ng'andu (1997) observed the phenomenon of non-constant

hazard ratio while comparing the survival probabilities of heroin addicts treated in two different methadone clinics. Ng'andu (1997) also reported an instance of crossing hazards among gastric carcinoma patients treated in a clinical trial that compared chemotherapy and chemotherapy combined with radiation therapy. The phenomenon of crossing hazards is sometimes modeled with increasing hazard ratio (IHR). Several authors (see Gill and Schumacher, 1987 and Deshpande and Sengupta, 1995) have proposed analytical tests for assessing the PH hypothesis for two-samples against the monotone hazard ratio alternative. Sengupta et al. (1998) proposed a test procedure against the weaker alternative of increasing cumulative hazard ratio.

In case the hazard ratio in two samples is neither constant nor monotonically increasing, a procedure for testing the PH hypothesis against the IHR alternative would produce a meaningless decision. Therefore, whether or not the PH hypothesis is rejected in favor of the IHR alternative for a particular data set, one cannot conclude that the hazard ratio is increasing. A definite conclusion can be reached if rejection of the PH hypothesis by such a test is complemented by non-rejection by another test, for which the null hypothesis is IHR.

In this paper, we propose an omnibus goodness-of-fit test to check the validity of the IHR assumption. We develop the procedure for complete data (with no censoring), and later indicate how it can be extended to the case of randomly right-censored data.

Consider two distributions F_X and F_Y having hazard functions λ_X and λ_Y, and cumulative hazard functions (CHF) $\Lambda_X = -\ln(1-F_X)$ and $\Lambda_Y = -\ln(1-F_Y)$, and density functions f_X and f_Y, respectively. The IHR model for the ordered pair of distributions (F_X, F_Y) means that the ratio λ_Y/λ_X is a monotone increasing function over $[0,\infty)$, while the PH model is such that λ_Y/λ_X is equal to some positive constant for all $t > 0$. Let $X_1, X_2, \ldots, X_m$ and $Y_1, Y_2, \ldots, Y_n$ be sets of independent samples from distributions F_X and F_Y, respectively. We assume that each set of samples is arranged in increasing order, and denote them by the vectors $\mathbf{X}$ and $\mathbf{Y}$, respectively.

A typical test of PH vs. IHR concerns the hypotheses

$$\begin{aligned} \text{Null hypothesis,} \quad & H_0 : \quad \lambda_Y/\lambda_X \text{ is a positive constant,} \\ \text{Alternative hyppothesis,} \quad & H_1 : \quad \lambda_Y/\lambda_X \text{ is an increasing function.} \end{aligned} \tag{1}$$

The testing problem considered here concerns the hypotheses

$$\begin{aligned} \text{Null hypothesis,} \quad & H_1 : \quad \lambda_Y/\lambda_X \text{ is an increasing function,} \\ \text{Alternative hypothesis,} \quad & H_2 : \quad \lambda_Y/\lambda_X \text{ is not an increasing function.} \end{aligned} \tag{2}$$

This article is organized as follows: In Section 2, we present a solution to the hypothesis testing problem (2) on the basis of data $\mathbf{Y}$, assuming that F_X is known. In Section 3, we develop a procedure for the testing problem (2), on the basis of data $\mathbf{X}$ and $\mathbf{Y}$. In Section 4,

we present a graphical version of this test. In Section 5, we present the results of a simulation study on the small sample performance of the proposed tests. In Section 6, we illustrate the use of the proposed tests through the analysis of two data sets.

2 A test for the IHR hypothesis: F_X is known

Let $Z_i = \Lambda_X(Y_i)$ for $i = 1, 2, \ldots, n$, and $\mathbf{Z} = (Z_1, Z_2, \ldots, Z_n)^T$. The vector $\mathbf{Z}$ may be said to contain a set of n ordered samples from the distribution $F_Z = F_Y \circ \Lambda_X^{-1}$, having CHF $\Lambda_Z = \Lambda_Y \circ \Lambda_X^{-1}$. Sengupta and Deshpande (1994) showed that λ_Y/λ_X is a monotonically increasing function if and only if $\Lambda_Y \circ \Lambda_X^{-1}$ is convex. The convexity of Λ_Z, in turn, is equivalent to the increasing failure rate (IFR) property of the distribution function F_Z. Therefore, the testing problem (2) reduces to the simplified problem

$$\begin{array}{rrl} \text{Null hypothesis,} & H_1^* : & F_Z \text{ is an IFR distribution,} \\ \text{Alternative hypothesis,} & H_2^* : & F_Z \text{ is not an IFR distribution,} \end{array} \tag{3}$$

based on the data $\mathbf{Z}$.

A solution to the above testing problem was proposed by Tenga and Santner (1984). We show the key steps in the development of a modification of that test, with the intention of following a similar route in Section 3, when F_X is unknown and represented by the data $\mathbf{X}$.

Let us assume $Y_1 < Y_2 < \cdots < Y_n$, i.e., the elements of $\mathbf{Y}$ are distinct. Note that, under the assumed model, this assumption is likely to hold with probability 1. Let $\widehat{\Lambda}_Z$ be the Nelson-Aalen estimator based on $\mathbf{Z}$, defined as

$$\widehat{\Lambda}_Z(t) = \sum_{i=1}^{n} \frac{I(Z_i \le t)}{\sum_{j=1}^{n} I(Z_i \le Z_j)}.$$

Let $\widehat{G}_Z$ denote the greatest convex minorant (GCM) of $\widehat{\Lambda}_Z$, i.e., the largest convex function dominated by $\widehat{\Lambda}_Z$. It can be shown that $\widehat{G}_Z$ is the piecewise linear function obtained by linearly interpolating in between its values at $Z_1, Z_2, \ldots, Z_n$, which are

$$g_j = \widehat{G}_Z(Z_j) = \begin{cases} h_1, & j = 1, \\ \min\left\{h_j, \min\limits_{D(j)} \left\{ \dfrac{Z_k - Z_j}{Z_k - Z_i} h_i + \dfrac{Z_j - Z_i}{Z_k - Z_i} h_k \right\}\right\}, & 1 < j < n, \\ h_n, & j = n, \end{cases}$$

where $h_j = \widehat{\Lambda}_Z(Z_j-)$ for $j = 1, 2, \ldots, n$ and $D(j) = \{(i,k) : 1 \le i < j < k \le n\}$ for $j = 1, 2, \ldots, n$. This fact is established by following the arguments used by Tenga and

Santner (1984) to prove a similar result (their Theorem 2.1), where they used the negative logarithm of the empirical survival function instead of the Nelson-Aalen estimator. The Kolmogorov-Smirnov (K-S) distance statistic for testing (3) is the maximum separation between the graphs of $\widehat{\Lambda}_Z$ and $\widehat{G}_Z$, which simplifies to

$$KSD_n(\mathbf{Z}) = \max_{1<j<n} \{w_j(h_j - g_j)\}, \tag{4}$$

where w_j is a weight function with $w_j > 0$ for $1 \leq j \leq n$. The weight function may be suitably chosen to counter the instability of the unweighted difference, arising from large variability of $\widehat{\Lambda}_Z$ in the right tail, because of smaller risk sets at large t (Gill and Schumacher, 1987).

It can be shown along the lines of Tenga and Santner (1984, Theorem 3.1) that $KSD_n(\mathbf{Z})$ is stochastically smaller than $KSD_n(\mathbf{U})$, where $\mathbf{U} = (\Lambda_Z(Z_1), \Lambda_Z(Z_2), \ldots, \Lambda_Z(Z_n))^T$. Since the latter vector consists of ordered samples from the unit exponential distribution, the small sample distribution of $KSD_n(\mathbf{U})$ can be obtained through simulations, and be used for obtaining a conservative test.

3 A test for the IHR hypothesis: F_X is unknown

Let us denote the Nelson-Aalen estimator based on the $\mathbf{X}$-sample by $\widehat{\Lambda}_X(t)$. Then, for fixed $Y_1 < Y_2 < \cdots < Y_n$, define $Z_{mj} = \widehat{\Lambda}_X(Y_j)$ for $j = 1, 2, \ldots, n$, and $\mathbf{Z}_m = (Z_{m1}, Z_{m2}, \ldots, Z_{mn})^T$. Note that the elements of $\mathbf{Z}_m$ may not be unique. Let Λ_{mn} be the Nelson-Aalen estimator computed from the data $\mathbf{Z}_m$.

The function Λ_{mn} is related to the relative trend function (RTF), defined by Lee and Pirie (1981) as $\Lambda_Y \circ \Lambda_X^{-1}$. Gill and Schumacher (1987) had worked with the empirical RTF (ERTF), $\widehat{\Lambda}_Y \circ \widehat{\Lambda}_X^{-1}$, where $\widehat{\Lambda}_X^{-1}$ is the left-continuous function defined by $\widehat{\Lambda}_X^{-1}(u) = \inf\{t : t > 0, \widehat{\Lambda}_X(t) \geq u\}$. It can be verified that $\Lambda_{mn}(t) = \widehat{\Lambda}_Y \circ \widehat{\Lambda}_X^{-1}(t+)$ for $t \in [0, Z_{mn})$. If one uses the right-continuous inverse defined by $\widehat{\Lambda}_X^{-1}(u) = \sup\{t : t > 0, \widehat{\Lambda}_X(t) \leq u\}$, then the function $\widehat{\Lambda}_Y \circ \widehat{\Lambda}_X^{-1}$ coincides with Λ_{mn} over the set $[0, Z_{mn})$.

Let us denote the GCM of Λ_{mn} by G_{mn}. It can be shown that this function has a form similar to that of $\widehat{G}_Z$ defined in Section 2, viz. it is the piecewise linear function obtained by linearly interpolating in between its values at $Z_{m1}, Z_{m2}, \ldots, Z_{mn}$, which are

$$g_{mj} = \begin{cases} h_{m1}, & 1 \leq j \leq c_{mn}, \\ \min\left\{h_{mj}, \min\limits_{D(j)} \left\{ \dfrac{Z_{mk} - Z_{mj}}{Z_{mk} - Z_{mi}} h_{mi} + \dfrac{Z_{mj} - Z_{mi}}{Z_{mk} - Z_{mi}} h_{mk} \right\} \right\}, & c_{mn} < j \leq n - d_{mn}, \\ h_{mn}, & n - d_{mn} < j \leq n, \end{cases}$$

where $c_{mn} = \sum_{j=1}^{n} I(Z_{mj} = Z_{m1})$, $\quad d_{mn} = \sum_{j=1}^{n} I(Z_{mj} = Z_{mn})$,

$$h_{mj} = \Lambda_{mn}(Z_{mj}-), \quad 1 \le j \le n,$$
$$D(j) = \{(i,k) :\ c_{mn} \le i < j < k \le n - d_{mn} + 1;\ Z_{mi} < Z_{mj} < Z_{mk}\}, \quad c_{mn} < j \le n - d_{mn}.$$

Note that linear interpolation is needed only when $g_{mj} \neq g_{ml}$, which happens only when $Z_{mj} \neq Z_{ml}$.

For testing the hypotheses in (2), we define a statistic analogous to the maximum distance statistic mentioned in Section 2, as

$$KSD_n(\mathbf{Z}_m) = \max_{c_{mn} < j < n - d_{mn} + 1} \{w_j(h_{mj} - g_{mj})\}, \tag{5}$$

where the weight w_j is positive for each j.

We now prove a result that enables us to approximate conservatively the null distribution of the statistic $KSD_n(\mathbf{Z}_m)$ by a universal distribution when the size m of the first sample is large.

THEOREM 3.1: Let $\mathbf{Z}_m = (Z_{m1}, Z_{m2}, \ldots, Z_{mn})$ be the vector of n-order statistics which is generated from two independent samples $\mathbf{X}$ and $\mathbf{Y}$ of sizes m and n from the distributions F_X and F_Y, respectively, as described above, such that the ordered pair of distributions $(F_X,\ F_Y)$ is IHR and $F_Y \circ F_X^{-1}(1) = 1$. Then

$$\lim_{m \to \infty} P\{KSD_n(\mathbf{Z}_m) \ge t\} \le P\{KSD_n(\mathbf{U}) \ge t\} \qquad \text{for all } t \ge 0, \tag{6}$$

where $\mathbf{U} = (U_1, U_2, \ldots, U_n)$ is the vector of order statistics of a sample of size n from the unit exponential distribution. The above result holds with equality if F_X and F_Y have proportional hazards.

PROOF: It can be shown that $\widehat{\Lambda}_X$ converges uniformly with probability 1 to Λ_X as $m \to \infty$ on $[0, \tau]$, for any positive number τ strictly less than $F_X^{-1}(1)$ (Shorack and Wellner, 1985). For such a number τ, let A_τ denote the event that $Y_n \le \tau$ and B_τ denote the event that $\widehat{\Lambda}_X$ converges to Λ_X uniformly on $[0, \tau]$ as $m \to \infty$. Then $P(A_\tau \cap B_\tau) = P(A_\tau)P(B_\tau) = P(A_\tau) = F_Y^n(\tau)$.

Let C_τ denote the event that $\mathbf{Z}_m \to \mathbf{Z}$ as $m \to \infty$. It is easy to see that $A_\tau \cap B_\tau$ implies

$$\begin{aligned} \|\mathbf{Z}_m - \mathbf{Z}\|_2 &\le \max_{1 \le j \le n} |Z_{mj} - Z_j|\sqrt{n} = \max_{1 \le j \le n} |\widehat{\Lambda}_X(Y_j) - \Lambda_X(Y_j)|\sqrt{n} \\ &\le \sup_{1 \le y \le \tau} |\widehat{\Lambda}_X(y) - \Lambda_X(y)|\sqrt{n} \\ &\to 0 \qquad \text{as } m \to \infty. \end{aligned}$$

Thus, $A_\tau \cap B_\tau \Rightarrow A_\tau \cap C_\tau$. Hence, $P(A_\tau \cap B_\tau) \leq P(A_\tau \cap C_\tau) \leq P(A_\tau)$, i.e., $P(A_\tau \cap C_\tau) = F_Y^n(\tau)$.

Let D_τ denote the event that there is a number M such that the vector $\mathbf{Z}_m$ has n distinct elements for $m > M$. We shall show that $A_\tau \cap C_\tau \Rightarrow D_\tau$. To see this implication, suppose the event $A_\tau \cap C_\tau$ has taken place. It follows that there is a number M such that for $m > M$, we have $\|\mathbf{Z}_m - \mathbf{Z}\|_2 < \epsilon$, where $\epsilon = \frac{1}{3}\min_{1\leq i<j\leq n}|Z_j - Z_i|$. Therefore, for $m > M$ and $1 \leq i < j \leq n$, we can write

$$\begin{aligned} |Z_{mj} - Z_{mi}| &= |(Z_{mj} - Z_j) - (Z_{mi} - Z_i) + (Z_j - Z_i)| \\ &\geq |Z_j - Z_i| - |Z_{mj} - Z_j| - |Z_{mi} - Z_i| \\ &> 3\epsilon - \epsilon - \epsilon = \epsilon > 0, \end{aligned}$$

i.e., the event D_τ has also occurred. Thus, $P(A_\tau \cap C_\tau \cap D_\tau) = P(A_\tau \cap C_\tau) = F_Y^n(\tau)$.

If the event D_τ takes place, then the vectors $(h_{m1}, h_{m2}, \ldots, h_{mn})^T$, $(g_{m1}, g_{m2}, \ldots, g_{mn})^T$ are both continuous functions of $\mathbf{Z}_m$ for sufficiently large m. Therefore, $KSD_n(\mathbf{Z}_m)$ is also a continuous function of $\mathbf{Z}_m$ for sufficiently large m. Thus, $A_\tau \cap C_\tau \cap D_\tau$ implies that $KSD_n(\mathbf{Z}_m) \to KSD_n(\mathbf{Z})$, an event that we denote by E. Thus, $P(E) \geq F_Y^n(\tau)$.

This inequality holds for all $\tau < F_X^{-1}(1)$. By allowing τ to go arbitrarily close to $F_X^{-1}(1)$, we can arrange $F_Y^n(\tau)$ to be arbitrarily close to $\{F_Y \circ F_X^{-1}(1)\}^n$, which is equal to 1. We conclude that $P(E) = 1$. Since almost sure convergence implies convergence in distribution, we can write

$$\lim_{m\to\infty} P\{KSD_n(\mathbf{Z}_m) \geq t\} = P\{KSD_n(\mathbf{Z}) \geq t\} \qquad \text{for all } t \geq 0.$$

Since Z has an IFR distribution, Theorem 3.1 of Tenga and Santner (1984) implies that $P\{KSD_n(\mathbf{Z}) \geq t\} \leq P\{KSD_n(\mathbf{U}) \geq t\}$ for all $t \geq 0$, with equality holding for all $t \geq 0$ if Z has the exponential distribution. The statement of the theorem follows. □

The above theorem shows that when m is large, the PH model gives rise to the least favorable configuration under H_1 in (2) for testing this hypothesis against H_2 through the statistic $KSD_n(\mathbf{Z})$. A test of H_1 vs. H_2 is then given by

$$\phi(\mathbf{Z}_m) = \begin{cases} 1, & \text{if } KSD_n(\mathbf{Z}_m) > c_{\alpha,n}, \\ 0, & \text{otherwise}, \end{cases} \tag{7}$$

where $c_{\alpha,n}$ is chosen to satisfy $P\{KSD_n(\mathbf{U}) > c_{\alpha,n}\} = \alpha$, $\mathbf{U} = (U_1, U_2, \ldots, U_n)$ being a vector of n-order statistics of a sample of size n from the unit exponential distribution. According to Theorem 3.1, the level of the test (7) goes to α as $m \to \infty$. The value $c_{\alpha,n}$ may be obtained through simulation as the $(1-\alpha)$ quantile of the distribution of $KSD_n(\mathbf{U})$.

One can also bypass Theorem 3.1 and use the following variation of the test (7):

$$\phi^*(\mathbf{Z}_m) = \begin{cases} 1, & \text{if } KSD_n(\mathbf{Z}_m) > c^*_{\alpha,m,n}(\theta), \\ 0, & \text{otherwise,} \end{cases} \tag{8}$$

where the alternative cutoff $c^*_{\alpha,m,n}(\theta)$ is chosen as the $1-\alpha$ quantile of the simulated distribution of $KSD_n(\mathbf{Z}_m)$, where $\mathbf{X}$ is sampled from the unit exponential distribution and $\mathbf{Y}$ is sampled from the exponential distribution with parameter θ (i.e., mean $1/\theta$), θ being a suitable hazard ratio that leads to a conservative cut-off.

The above test can be extended to the case of randomly right-censored data by using the censored-data version of the Nelson Aalen estimators,

$$\begin{aligned} \widehat{\Lambda}_X(t) &= \sum_{\substack{i\in\{1,2,\ldots,m\} \\ X_i \text{ is uncensored}}} \frac{I(X_i \le t)}{\sum_{j=1}^n I(X_i \le X_j)}, \\ \Lambda_{mn}(t) &= \sum_{\substack{i\in\{1,2,\ldots,n\} \\ Y_i \text{ is uncensored}}} \frac{I(Z_{mi} \le t)}{\sum_{j=1}^n I(Z_{mi} \le Z_{mj})}. \end{aligned}$$

While Theorem 3.1 cannot be used for censored data, a cut-off determined by simulation of two samples can be used.

4 A related graphical test

Sahoo and Sengupta (2014) had introduced acceptance bands for a constrained estimate of the relative trend function under the PH assumption, by making use of the supremum of the difference between this estimate and an unconstrained estimate. The idea was that if the hazards are indeed proportional, then the constrained estimate should lie within a neighborhood of the unconstrained estimate (an acceptance band). Non-containment within the band corresponds to rejection of the null hypothesis.

Using the above idea, the test proposed in Section 3 can be redesigned as a graphical test of the IHR assumption. Let w be a piecewise constant function defined over the interval $[Z_{m1},\ Z_{mn})$ as

$$w(t) = w_j \qquad \text{if } \max\{Z_{mi} :\ Z_{mi} < Z_{mj}\} \le t < Z_{mj} \text{ for some } j,\ c_{mn} < j \le n,$$

where w_j is the weight used in (5). Then, under the null hypothesis, the GCM of Λ_{mn} should lie, with probability $1-\alpha$, above $\Lambda_{mn} - c^*_{\alpha,m,n}(\theta)/w$ or $\Lambda_{mn} - c_{\alpha,n}/w$, depending on whichever cut-off is used. Therefore, this curve may be plotted as a lower acceptance

band for the GCM of Λ_{mn}. The GCM strays beyond the acceptance band if and only if the hypothesis is rejected. If there is rejection, the plot would give a visual indication as to where the GCM crosses the band, i.e., deviates most from the IHR assumption.

5 Simulation studies

We run an initial set of simulations to determine the thresholds for the proposed test, and examine whether to use the threshold $c_{\alpha,n}$ or the threshold $c^*_{\alpha,m,n}(\theta)$ with a suitable θ. We then perform further simulations to study the empirical size and power of the tests.

For each set of simulations reported below, we use two choices of the weights w_j (see Tenga and Santner, 1984) in the statistic (5),

$$w_j^{(1)} = \frac{1}{h_{mj}}, \tag{9}$$

$$w_j^{(2)} = \frac{1}{n(h_{mj} - \max\{h_{mi} : \ h_{mi} < h_{mj}\})}. \tag{10}$$

A total of 10,000 replicates are used for obtaining each entry in Tables 1-6 that follow.

5.1 Determining Conservative threshold

In order to determine $c_{\alpha,n}$, we simulate the random vector $\mathbf{U}$ as an ordered set of n samples from the unit exponential distribution, compute $KSD_n(\mathbf{U})$, and then use many replicates to determine the $1-\alpha$ quantile. We use $\alpha = 0.05$, $n = 50$, 100, 200 and 500.

In order to determine $c^*_{\alpha,m,n}(\theta)$, we simulate the random vector $\mathbf{X}$ as an ordered set of m samples from the unit exponential distribution and the random vector $\mathbf{Y}$ as an ordered set of n samples from the exponential distribution with mean $1/\theta$. Subsequently, we compute $KSD_n(\mathbf{Z}_m)$, and then use many replicates to determine the $1-\alpha$ quantile. We use $\alpha = 0.05$, $m, n = 50$, 100, 200 and 500, $\theta = 1/4$, $1/2$, 1, 2 and 4. The results for the weight functions $w_j^{(1)}$ and $w_j^{(2)}$, based on 10,000 runs, are summarized in Tables 1 and 2, respectively.

It is observed from Table 1 that as θ increases, the quantile $c^*_{0.05,n,n}(\theta)$ of the test decreases gradually. For the range of values of θ considered here, the values of $c^*_{0.05,n,n}(\theta)$ are much smaller than $c_{0.05,n}$. This finding indicates that the cut-off $c_{0.05,n}$, meant for large m, is too conservative for $m = n$ and the choices of n used here. Therefore, when the weight function is as in (9), we decide to use the test (8) with $\theta = 0.25$, rather than using the test (7).

On the other hand, Table 2 shows that when the weight function $w_j^{(2)}$ is used, the quantile $c^*_{0.05,n,n}(\theta)$ initially increases and then decreases with θ. The highest point is generally reached at $\theta = 2$. However, the maximum value is not necessarily smaller than the corresponding

Table 1. Empirical thresholds $c_{0.05,n}$ and $c^*_{0.05,n,n}(\theta)$ for test statistic (5) with weights (9)

Threshold	Value of threshold for			
	$n = 50$	$n = 100$	$n = 200$	$n = 500$
$c_{0.05,n}$	0.9634	0.9665	0.9657	0.9627
$c^*_{0.05,n,n}(\theta),\ \theta = 1/4$	0.8828	0.8825	0.8818	0.8835
$c^*_{0.05,n,n}(\theta),\ \theta = 1/2$	0.8529	0.8502	0.8499	0.8479
$c^*_{0.05,n,n}(\theta),\ \theta = 1$	0.8343	0.8273	0.8209	0.8182
$c^*_{0.05,n,n}(\theta),\ \theta = 2$	0.8139	0.8139	0.8093	0.8005
$c^*_{0.05,n,n}(\theta),\ \theta = 4$	0.7950	0.7918	0.7920	0.7892

Table 2. Empirical thresholds $c_{0.05,n}$ and $c^*_{0.05,n,n}(\theta)$ for test statistic (5) with weights (10)

Threshold	Value of threshold for			
	$n = 50$	$n = 100$	$n = 200$	$n = 500$
$c_{0.05,n}$	0.2142	0.1896	0.1678	0.1475
$c^*_{0.05,n,n}(\theta),\ \theta = 1/4$	0.1450	0.1315	0.1192	0.1039
$c^*_{0.05,n,n}(\theta),\ \theta = 1/2$	0.1776	0.1588	0.1391	0.1174
$c^*_{0.05,n,n}(\theta),\ \theta = 1$	0.2012	0.1827	0.1618	0.1381
$c^*_{0.05,n,n}(\theta),\ \theta = 2$	0.2152	0.2013	0.1822	0.1579
$c^*_{0.05,n,n}(\theta),\ \theta = 4$	0.1968	0.1934	0.1806	0.1569

quantile $c_{0.05,n}$ for the same value of n. Therefore, when the weight function is as in (10), we decide to use the test (8) with $\theta = 2$, rather than using the test (7).

5.2 Empirical Size

Since the cut-offs $c^*_{0.05,m,n}(0.25)$ and $c^*_{0.05,m,n}(2)$ for weight functions (9) and (10) were determined conservatively after considering several combinations of exponential distributions for the two samples, we now check the empirical size by considering some other combinations of distributions under the null hypothesis. Let $W(\gamma, \delta)$ denote the Weibull distribution having hazard rate $\gamma\delta t^{\delta-1}$. In our first experiment, we choose $W(1, 2)$ and $W(\theta, 2)$ as the distributions for the first and second samples, respectively. Here, the hazard ratio is constant and is equal to θ, which is chosen as 1/4, 1/2, 1, 2 and 4. In all cases, we choose equal sample sizes ($m = n = 50$, 100, 200, 500) and 10,000 simulation runs.

The results given in Table 3 show that for both weight functions, the empirically observed size from simulations is sometimes about the same as the nominal size, and often below it.

In our second experiment, we choose $W(1, 1)$ and $W(\gamma, \delta)$ as the distributions for the

Table 3. Empirical size of test (8) with weights (9) and (10) under PH models with distributions $W(1,2)$ and $W(\theta,2)$ (nominal size 0.05)

θ	Size of test with weight $w_j^{(1)}$ for				Size of test with weight $w_j^{(2)}$ for			
	$n=50$	$n=100$	$n=200$	$n=500$	$n=50$	$n=100$	$n=200$	$n=500$
1/4	0.0454	0.0502	0.0521	0.0463	0.0003	0.0001	0.0000	0.0001
1/2	0.0261	0.0205	0.0188	0.0182	0.0058	0.0024	0.0009	0.0006
1	0.0138	0.0103	0.0104	0.0104	0.0261	0.0176	0.0134	0.0093
2	0.0078	0.0069	0.0060	0.0057	0.0466	0.0485	0.0510	0.0513
4	0.0032	0.0043	0.0030	0.0030	0.0233	0.0372	0.0461	0.0500

first and second samples, respectively, where δ is chosen as 1.2, 1.4, 1.6, 1.8 and 2, and γ is chosen to ensure that the two distributions have the same median. Note that in this case, the hazard ratio is $\gamma\delta t^{\delta-1}$, which is an increasing function of t.

The results are summarized in Table 4. It is seen that for both the weight functions, the empirically observed size is smaller than the nominal size. For the test with weight $w_j^{(2)}$, the size is smaller when δ is away from 1, i.e., the hazard ratio is away from the PH model.

Our third experiment involves censored data. We consider the pairs of distributions $W(1,1)$ and $W(\theta,1)$ with $\theta = 0.25$, 1 and 2. The censoring distribution for each sample is also chosen as exponential, with parameter adjusted to produce a specified censoring fraction, i.e., a theoretically specified probability ϕ of a particular observation being censored. We use the same censoring fraction in the two samples, which is allowed to have values 0.1, 0.25 and 0.50. Since the study on thresholds had indicated that a lower sample size produces a higher threshold, we use the expected number of complete data samples while selecting the appropriate complete-data threshold for size calculations. Specifically, we calculate the empirical size of the tests for sample sizes m and and n (expected to include $100\phi\%$ censored

Table 4. Empirical size of test (8) with weights (9) and (10) under IHR models with equi-median distributions $W(1,1)$ and $W(\gamma,\delta)$ (nominal size 0.05)

δ	Size of test with weight $w_j^{(1)}$ for				Size of test with weight $w_j^{(2)}$ for			
	$n=50$	$n=100$	$n=200$	$n=500$	$n=50$	$n=100$	$n=200$	$n=500$
1.2	0.0015	0.0026	0.0025	0.0041	0.0061	0.0013	0.0009	0.0001
1.4	0.0012	0.0019	0.0037	0.0071	0.0026	0.0004	0.0000	0.0000
1.6	0.0008	0.0013	0.0049	0.0084	0.0006	0.0000	0.0001	0.0000
1.8	0.0006	0.0024	0.0043	0.0137	0.0002	0.0001	0.0000	0.0000
2.0	0.0004	0.0019	0.0073	0.0166	0.0000	0.0000	0.0000	0.0000

Table 5. Empirical size of test (8) with weights (9) and (10) for censored data from PH models with distributions $W(1,1)$ and $W(\theta,1)$ (censoring fraction ϕ, nominal size 0.05)

ϕ	θ	Size of test with weight $w_j^{(1)}$ for $[n(1-\phi)] =$				Size of test with weight $w_j^{(2)}$ for $[n(1-\phi)] =$			
		50	100	200	500	50	100	200	500
0.10	0.25	0.0496	0.0467	0.0495	0.0511	0.0000	0.0001	0.0000	0.0000
0.10	1	0.0141	0.0099	0.0099	0.0078	0.0109	0.0064	0.0028	0.0030
0.10	2	0.0067	0.0084	0.0049	0.0039	0.0196	0.0206	0.0208	0.0225
0.25	0.25	0.0483	0.0487	0.0545	0.0490	0.0000	0.0000	0.0000	0.0000
0.25	1	0.0129	0.0132	0.0095	0.0123	0.0168	0.0004	0.0000	0.0003
0.25	2	0.0082	0.0067	0.0063	0.0050	0.0029	0.0027	0.0026	0.0044
0.50	0.25	0.0546	0.0547	0.0528	0.0538	0.0000	0.0000	0.0000	0.0000
0.50	1	0.0165	0.0126	0.0129	0.0094	0.0000	0.0000	0.0000	0.0000
0.50	2	0.0078	0.0059	0.0064	0.0041	0.0000	0.0000	0.0000	0.0000

data), by comparing with the threshold obtained for complete data with sample sizes $[m(1-\phi)]$ and $[n(1-\phi)]$. For both of the latter numbers, we use the values 50, 100, 200 and 500. The results reported in Table 5.

The findings of Table 5 indicate that the empirical size of the test with weight $w_j^{(2)}$ is generally smaller than the nominal size, while the size of the test with weight $w_j^{(1)}$ marginally exceeds the nominal size when there is heavy censoring.

5.3 Empirical Power

We study the power properties of the tests for the decreasing hazard ratio (DHR) alternative, obtained through the distributions $W(1,1)$ and $W(\gamma,\delta)$. The parameter δ is varied in steps from 0.8 to 0.2, producing progressive departure from the PH assumption (included in the null hypothesis). The parameter γ is adjusted so that $W(\gamma,\delta)$ has the same median as $W(1,1)$. We use complete sample of size $m = n = 50$, 100, 200 and 500.

The results are shown in Table 6. As expected, the empirical power is found to increase with δ, as one moves further away from the null hypothesis. The increase in power happens faster when the sample size is larger. The test with weight $w_j^{(2)}$ is found to have slightly larger power when the hazard ratio does not decrease fast (i.e., the model is not very far from the null hypothesis) and the sample size is large. The test with weight $w_j^{(1)}$ has higher power when the sample size is small but the departure from the null hypothesis is substantial.

Table 6. Empirical power of test (8) with weights (9) and (10) under IHR models with equi-median distributions $W(1,1)$ and $W(\gamma,\delta)$ (nominal size 0.05)

δ	Size of test with weight $w_j^{(1)}$ for				Size of test with weight $w_j^{(2)}$ for			
	$n=50$	$n=100$	$n=200$	$n=500$	$n=50$	$n=100$	$n=200$	$n=500$
0.8	0.0768	0.1093	0.1621	0.2279	0.1114	0.1446	0.2403	0.4851
0.7	0.1768	0.3006	0.4677	0.6631	0.2097	0.3474	0.6002	0.9371
0.6	0.3917	0.6132	0.8346	0.9805	0.3815	0.6271	0.9149	0.9997
0.5	0.6635	0.9035	0.9904	1.0000	0.5832	0.8658	0.9932	1.0000
0.4	0.8964	0.9943	1.0000	1.0000	0.7595	0.9718	1.0000	1.0000
0.3	0.9863	0.9999	1.0000	1.0000	0.8875	0.9959	1.0000	1.0000
0.2	0.9961	1.0000	1.0000	1.0000	0.9366	0.9997	1.0000	1.0000

6 Data analysis

In this section, we illustrate the utility of the tests proposed here by analyzing two real data sets.

The first data set consists of the survival times (in days) of gastric cancer patients in a clinical trial that compared chemotherapy and chemotherapy combined with radiation therapy in the treatment of locally advanced unresectable gastric carcinoma. The trial was conducted by the Gastrointestinal Tumor Study Group (Schein et al., 1982), and in the trial, 45 patients were randomized to each of the two treatment groups. There are two cases of censoring in the first group, and six cases of censoring in the second group. The data are analyzed by Stablein and Koutrouvelis (1985) and also by Klein and Moeschberger (1997). These studies reveal the situation of crossing hazards between the two groups. We refer to this data set as the gastric cancer data (GCD).

The second data example concerns patients with primary biliary cirrhosis (PBC) of the liver. The Mayo Clinic conducted a randomized clinical trial of PBC patients between January, 1974 and May, 1984, to compare the drug D-penicillamine (DPCA) with a placebo. There were 312 participants for whom complete information was available. Of these randomized patients, 125 had died in the trial by July, 1986. The data set is given in Fleming and Harrington (1991, Appendix D.1) with detailed analysis. While DPCA turned out to produce no significant improvement over the placebo, the covariate `protime` (prothrombin time) was found to have a bearing on the survival time. Graphical checks indicated a non-monotone hazard ratio over time for groups of patients stratified by this covariate (see Fleming and Harrington, 1991 and Therneau et. al., 1990). For the present analysis, we use the threshold 11 for `protime`, to split the 312 subjects into two groups with approximately equal number

Table 7. P-values of the various tests for GCD and PBC data sets

Data set	Tests of PH			Tests of IHR	
	T_{GL}	T_{PL}	U_c	KSD_{mn} with (9)	KSD_{mn} with (10)
GCD	< 0.0001	< 0.0001	0.004	0.868	0.623
PBC	< 0.0001	< 0.0001	< 0.0001	0.007	0.021

of observed deaths. Group 1 with `protime` less than or equal to 11 has 229 subjects including 66 deaths, while Group 2 with `protime` greater than 11 has 83 subjects including 59 that were observed to die.

The p-values of three analytical tests of the PH assumption, all of which are consistent against the IHR alternative, are reported in Table 7. The test statistics T_{GL} and T_{PL} are those proposed by Gill and Schumacher (1987), with the Gehan versus log-rank weight functions and the Prentice versus log-rank weight functions, respectively. The statistic U_c is the U-statistic proposed by Deshpande and Sengupta (1995), with standard deviation for scaling determined by 5000 simulation runs. The table shows that one-sided p-values of all the tests indicate rejection of the PH hypothesis for both the data sets.

The last two columns of Table 7 show the p-values of the proposed test (8) with weight functions given by (9) and (10). The p-values are computed by simulation. Specifically, for each data set, simulated values of $KSD_n(\mathbf{Z}_m)$ were generated by using uncensored samples drawn from the exponential distributions with rate parameters 1 and θ, respectively. For reasons explained in Section 5.1, θ was chosen as 0.25 in the case of weight function (9) and as 2 in the case of weight function (10). The sample sizes of the simulated data were selected to match the number of uncensored observations in the actual data set. It may be observed that the proposed test with either weight function clearly indicates acceptance of the IHR hypothesis in the case of the GCD data and rejection of that hypothesis in the case of the PBC data. This conclusion complements the findings of the existing analytical tests, which are able to reject the PH hypothesis but are unable to distinguish between the IHR and non-IHR situations.

As a follow-up study, we use the graphical test described in Section 4. The test is based on the function Λ_{mn}, which is an empirical version of the relative trend function (see the discussion in the second paragraph of Section 3). For the sake of simplicity, we refer to this function as the RTF in the present discussion. Figure 1 shows, for the GCD data, the plot of the RTF, its GCM and two lower acceptance bands of the GCM computed at level $\alpha = 0.05$ by using the weight function $w_j^{(1)}$ given by (9) and the weight function $w_j^{(2)}$ given by (10).

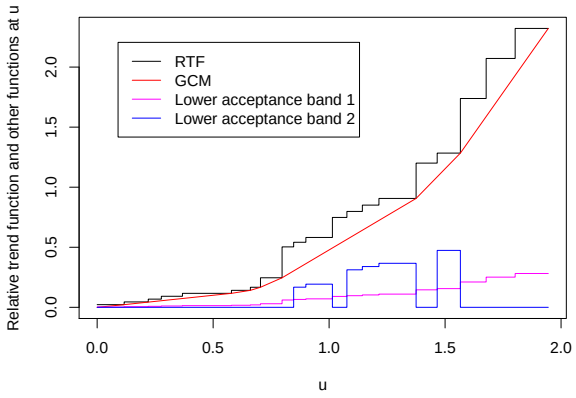

Figure 1. Plots of RTF, its GCM and 95% lower acceptance bands of GCM, computed with two weight functions, for the GCD data

The latter curves are referred to as lower acceptance bands 1 and 2, respectively. Figure 2 Shows the corresponding plot for the PBC data.

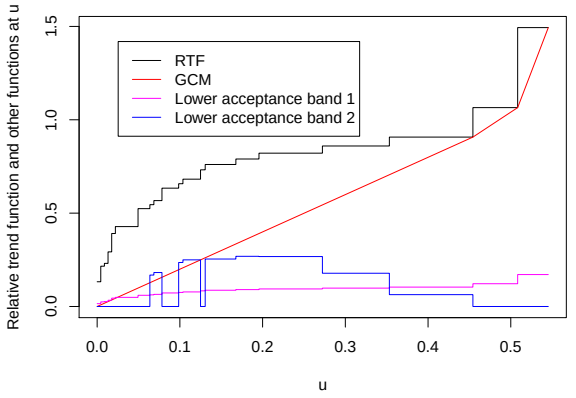

Figure 2. Plots of RTF, its GCM and 95% lower acceptance bands of GCM, computed with two weight functions, for the PBC data

The GCM in the case of the GCD data stays well above the lower acceptance bands all the way. In the case of the PBC data however, the GCM crosses the acceptance band. The crossing occurs rather early in the case of the band 1, as the first weight function gives much weight to the initial portion of the RTF. In the case of the second weight function, less weight is given to regions of sudden jump in the RTF. Consequently, the crossing of band 2

occurs in regions of small jump of the RTF. Crossings of both the bands occur in a region where the RTF is by and large concave. These are indeed regions of clear departure from the ideal (convex) shape of the RTF under the null hypothesis.

Acknowledgement

This research is partially sponsored by the project "Optimization and Reliability Modeling" funded by the Indian Statistical Institute, Kolkata.

References

Begg, C. B., McGlave, P. B., Bennet, J. M., Cassileth, P. A., and Oken, M. M. (1984). A critical comparison of allogeneic bone marrow transplantation and conventional chemotherapy as treatment for acute non-lymphomytic leukemia. *Journal of Clinical Oncology*, **2**, 369–378.

Breslow, N. E., Edler, L., and Berger, J. (1984). A two-sample censored-data rank test for acceleration. *Biometrics*, **40**, 1049–1062.

Champlin, R., Mitsuyasu, R., Elashoff, R., and Gale, R. P. (1983). Recent advances in bone-marrow transplantation. In *UCLA Symposia on Molecular and Cellular Biology*, Ed. R. P. Gale, **7**, pp. 141–158. New York: Alan R. Liss.

Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of Royal Statistical Society, Series B*, **34**, 187–202.

Dabrowska, D., Doksum, K., and Song, J. K. (1989). Graphical comparison of cumulative hazards for two populations. *Biometrika*, **76**, 763–773.

Dabrowska, D., Doksum, K., Feduska, N. J., Husing, R., and Neville, P. (1992). Methods for comparing cumulative hazard functions in a semi-parametric hazard model. *Statistics in Medicine*, **11**, 1465–1476.

Deshpande, J. V., and Sengupta, D. (1995). Testing for the hypothesis of proportional hazards in two populations. *Biometrika* **82**, 251–261.

Fleming, T.R., and Harrington, D.P. (1991). *Counting Processes and Survival Analysis*. Wiley Series in Probability and Mathematical Statistics, Wiley, New York.

Schein, P. D., Stablein, D. M., Bruckner, H. W., Douglass, H. O., Mayer, R., et al. (1982).

A comparison of combination chemotherapy and combined modality therapy for locally advanced gastric carcinoma. *Cancer*, **49**, 1771–1777.

Gill, R., and Schumacher, M. (1987). A simple test of the proportional hazards assumption. *Biometrika*, **74**, 289–300.

Kay, R. (1977). Proportional hazards regression models and the analysis of censored survival data. *Applied Statistics*, **26**, 227–237.

Klein, J. P., and Moeschberger, M. L. (1997). *Survival Analysis: Techniques for Censored and Truncated Data*. Series in Statistics for Biology and Health, Springer, New York.

Lee, L., and Pirie, W. R. (1981). A graphical method for comparing trends in series of events. *Communications in Statistics, Theory and Methods*, **10**, 827–848.

Ng'andu, N. H. (1997). An empirical comparison of statistical tests for assessing the proportional hazards assumption of Cox's model. *Statistics in Medicine*, **16**, 611–626.

Pocock, S. J., Gore. S. M., and Kerr, G. R. (1982). Long-term survival analysis: The curability of breast cancer. *Statistics in Medicine*, **1**, 93–104.

Sahoo, S., and Sengupta, D. (2014). On graphical tests for proportionality of hazards in two samples. *Technical Report* No. ASU/2014/5, Applied Statistics Unit, Indian Statistical Institute, 2014 (available in `http://www.isical.ac.in/~asu/TR/TechRepASU201405.pdf`).

Sengupta, D., and Deshpande, J. V. (1994). Some results on the relative aging of two life distributions. *Journal of Applied Probability*, **31**, 991–1003.

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

Stablein, D. M., and Koutrouvelis, I. A. (1985). A two-sample test sensitive to crossing hazards in uncensored and singly censored data. *Biometrics*, **41**, 643–652.

Tenga, R., and Santner, T. J. (1984). Testing goodness of fit to the increasing failure rate family. *Naval Research Logistics*, **31**, 617–630.

Therneau, T.M. and Grambsch, P.M. (2000). *Modelling Survival Data: Extending the Cox Model*. Springer Series in Statistics for Biology and Health, Springer, New York.

Wei, L. J. (1984). Testing goodness of fit for proportional hazards model with censored observations. *Journal of the American Statistical Association*, **79**, 649–652.