STATISTICS IN MEDICINE
*Statist. Med.* (in press)
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.2544

# Comparing two crossing hazard rates by Cox proportional hazards modelling

Kejian Liu[1,‡] Peihua Qiu[2,*,†] and Jun Sheng[2,§]

[1]*Department of Biostatistics, Novartis Pharmaceuticals Corporation, One Health Plaza, East Hanover, NJ 07936, U.S.A.*
[2]*School of Statistics, University of Minnesota, 224 Church Street SE, Minneapolis, MN 55455, U.S.A.*

SUMMARY

Motivated by a clinical trial of zinc nasal spray for the treatment of the common cold, we consider the problem of comparing two crossing hazard rates. A comprehensive review of the existing methods for dealing with the crossing hazard rates problem is provided. A new method, based on modelling the crossing hazard rates, is proposed and implemented under the Cox proportional hazards framework. The main advantage of the proposed method is the utilization of the Box–Cox transformation which covers a wide range of hazard crossing patterns. Simulation studies are conducted for comparing the performance of the existing methods and the proposed one, which show that the proposed method outperforms some of its peers in certain cases. Applications to a kidney dialysis patients data and the zinc nasal spray clinical trial data are discussed. Copyright © 2006 John Wiley & Sons, Ltd.

KEY WORDS: Box–Cox transformation; covariates; likelihood ratio test; power; proportional hazards regression; resampling techniques; survival analysis; treatment effects; zinc nasal spray

## 1. INTRODUCTION

In clinical trials, the primary outcome is often defined as the time to occurrence of a clinically important event. For instance, for life-threatening diseases, such as cancer, cardiovascular diseases, and AIDS, the most relevant endpoint is often to evaluate treatment effects by investigating patients' survival times. For infectious diseases, with which patients' full recovery is possible, we are often interested in the time to resolution of the disease. In this kind of survival data, some type(s) of data censoring, such as the right censoring, interval censoring, and so forth, is often involved. See, e.g. Reference [1] for a detailed discussion.

*Correspondence to: Peihua Qiu, School of Statistics, University of Minnesota, 224 Church Street SE, Minneapolis, MN 55455, U.S.A.
†E-mail: qiu@stat.umn.edu
‡E-mail: kejian.liu@novartis.com
§E-mail: junsheng@stat.umn.edu

*Received 14 March 2005*
*Accepted 30 January 2006*

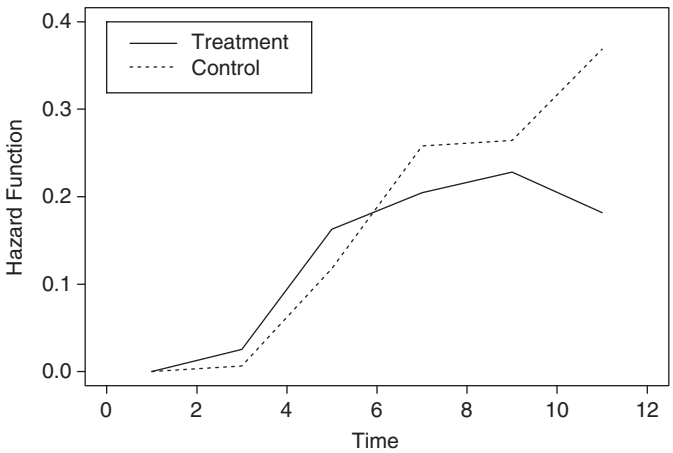


Figure 1. Life-table estimates of the two hazard rates of cold resolution for the treatment and control groups in the zinc nasal spray study. Time is in unit of days.

To evaluate treatment effects in these cases, the logrank, Gehan–Wilcoxon, and Peto–Peto tests, among several others, are routinely used in practice for handling censored data (cf., e.g. Reference [1], Chapter 7). It is well known in the literature that the logrank test is optimal when the two hazard rates of the treatment and control groups are proportional. However, the assumption of proportional hazard rates is obviously violated when the two hazard rates cross each other at some unknown time point. Therefore, new statistical methodologies are needed for evaluating treatment effects in cases with crossing hazard rates. This problem is the focus of the current paper.

The crossing hazard rates phenomenon is often seen when the treatment has quite different effects during different stages of a disease. An immediate example is provided by a recent clinical trial study for evaluating the effect of zinc nasal spray on curing common cold, which was performed by a research group in Marshfield Clinic at Wisconsin. Since the data from this study will be further analysed in Section 4, we provide a brief description of it here. Interested readers can see Reference [2] for a detailed description. This study was carried out as a randomized, double-blinded, placebo-controlled clinical trial, for investigating whether zinc nasal spray could shorten cold duration, and/or lessen cold symptom severity. A total of 160 patients were recruited to this study. For each patient, his or her daily cold symptom scores were recorded, and then the cold duration was determined from them. Statistical analysis included in Reference [2] did not find any significant treatment effect regarding cold duration; further investigation of the daily symptom scores suggested a transient reduction of symptom severity in the early stage of the medical treatment. The life-table estimates of the two hazard rates of cold resolution for the treatment and control groups are shown in Figure 1, from which it can be seen that they cross each other around the sixth day. Consequently, both the logrank and Gehan–Wilcoxon tests used by Belongia *et al.* [2] have little power in testing their difference, because these conventional tests could not usually handle the crossing hazard rates problem properly. Therefore, it is interesting to investigate whether the non-significance results from the conventional tests are because of the ineffective treatment, or if they are mainly

 

caused by the insensitivity of the statistical tests to the crossing pattern. In addition, in this study, the treatment group showed somewhat lower symptom scores at baseline. Therefore, it is also interesting to evaluate treatment effects after adjusting the baseline symptom scores, which can be treated as a covariate in this study.

For the crossing hazard rates problem, people may think that it is more convenient to check whether the two corresponding survival functions cross each other by looking at their Kaplan–Meier estimates. Since the survival function is a monotone transformation of the cumulative hazard function, it is easy to establish the following interesting relationship between the crossing of survival functions and the crossing of hazard rates. If two continuous survival functions cross each other once, then the corresponding hazard rates must cross each other at least once. However, the reverse is not true. It is possible that two survival functions do not cross, but their hazard rates cross very often. These results indicate that we still need to check whether the two hazard rates cross each other, even if their survival functions do not cross. That is, the former problem cannot be replaced by the latter one.

In the literature, there are several existing methods for testing whether two hazard rates cross each other. We conduct a comprehensive survey of these methods in Section 2, after the problem of crossing hazard rates is formally formulated. Then, in the same section, our proposed method and its major features are discussed. In Section 3, the proposed method is compared with several existing ones in some simulation examples. All the related methods are then applied to a kidney dialysis patients data and the zinc clinical trial data in Section 4. Finally, several remarks conclude the article in Section 5.

## 2. CROSSING HAZARD RATES: THE PROBLEM, EXISTING METHODS, AND PROPOSED METHOD

In this section, we first formulate the crossing hazard rates problem in statistical terms (Section 2.1), then provide a comprehensive review of existing methods that can be used for handling this problem (Section 2.2), and finally describe in details our proposed method and its major features (Section 2.3).

### *2.1. The problem*

The crossing hazard rates problem has been discussed by several authors (e.g. References [3–6]). In the literature, people are usually interested in testing the equality of two hazard rates against the specific alternative of crossing hazard rates. In this paper, we follow this tradition. Let $h_0$ and $h_1$ be the hazard rates of the survival times of the subjects in the control and treatment groups, respectively. Based on two censored samples of sizes $n_0$ and $n_1$ from the two groups, we want to test

$$\begin{aligned} &H_0 : h_1(t) = h_0(t), \quad \text{for all } t \in [0, \tau] \\ &\text{vs} \quad H_a : h_1 \quad \text{and} \quad h_0 \text{ cross each other at one point } \gamma \in [0, \tau] \end{aligned} \tag{1}$$

where $\gamma$ is the crossing time point, $[0, \tau]$ is the time range of interest, and $\tau$ is usually taken to be the largest observed survival time in the data (e.g. Reference [1], Chapter 7). For the crossing time point $\gamma$, we also want to construct its confidence interval.

 

In this article, we assume that the two hazard rates are both continuous, which is a convention in the survival analysis literature. This convention is also adopted by most authors in handling the crossing hazard rates problem. See next subsection for introduction about some existing methods on this topic. Then, the alternative hypothesis in (1) has two possibilities. One is that $h_1(t) < h_0(t)$ when $t < \gamma$, $h_1(t) = h_0(t)$ when $t = \gamma$, and $h_1(t) > h_0(t)$ when $t > \gamma$. In such cases, the treatment has benefits only in the early stage of the disease; it does not have any long-term advantages. As an example, radiation and chemotherapy can usually improve short-term patients' survivals; but they have little or no long-term medical benefits. The other possibility is that $h_1(t) > h_0(t)$ when $t < \gamma$, $h_1(t) = h_0(t)$ when $t = \gamma$, and $h_1(t) < h_0(t)$ when $t > \gamma$. Treatments of this type have benefits in the long run; but they may increase the risk in the early stage of the disease. Surgery is a good example of this type of medical treatments—due to infection and other short-term risks, it may cause high mortality in a short period after surgery. But, in the long run, most surgeries could improve patients' health conditions.

Besides testing the hypotheses in (1) and constructing a confidence interval for the crossing time point $\gamma$, we are also interested in these statistical inferences after adjusting the effect of some confounding covariates. In the zinc clinical trial example described in Section 1, one important confounding covariate is the baseline cold symptom score. As an aside, in the current paper, we are not interested in the case when two hazard rates cross at two or more places, since this situation is not common in practice.

### 2.2. *The existing methods*

The crossing hazard rates problem has received much attention from some statisticians since 1980s. Fleming *et al.* [7] proposed a modified Kolmogorov–Smirnov statistic for comparing two crossing hazard rates. Gill [8] suggested a supremum version of the weighted logrank tests, which is often called the Renyi-type test in the literature. Three other tests, which are all analogues of common non-parametric tests for uncensored data, including the Cramer–Von Mises test, were discussed by Klein and Moeshberger ([1], Section 7.7). Recently, Lin and Wang [9] suggested a test statistic, by measuring the squared difference between the treatment and control groups, for testing whether the two hazard rates are equivalent. All these tests mentioned above are expected to have greater sensitivity to crossings of two hazard rates, compared to the logrank test and other conventional tests, because they avoid early differences between the two hazard rates being cancelled out by late differences of opposite signs in the hazard rates, which occurs when a conventional test is applied to a case with crossing hazard rates. However, due to their omnibus nature, these tests, which are referred to as the *first class* of methods in the remaining parts of the article, would have reduced power in testing hypotheses with specific alternatives, such as the one in (1).

The *second class* of methods handle the crossing hazard rates problem by choosing special weights in the weighted logrank test. The conventional weights, such as the logrank, Gehan, Peto–Peto, and Fleming–Harrison weights (see Reference [1], Chapter 7), are all positive. Consequently, when two hazard rates cross, the early differences in favour of one group would be cancelled out by the late differences in favour of the other group, as mentioned above. A natural solution to this problem is to use a weighting scheme that changes its sign before and after the crossing point. This is the major idea behind the method proposed by Mantel and Stablein [10]. More specifically, when the crossing point $\gamma$ is given, Mantel and Stablein suggested using weights $w_i = 1$ when $t_i < \gamma$ and $w_i = -1$ when $t_i > \gamma$. The resulting

 

statistic is denoted by $W_\gamma$. If the crossing point $\gamma$ is unknown, then they suggested using the test statistic $W = \sup_\gamma W_\gamma$. To use this test, the null distribution of $W$ should be determined by simulation, since its theoretical expression and/or asymptotic results are not available yet. An alternative weighting scheme was proposed by Moreau *et al.* [11], which is defined by $w_i = 1 + \log(-\log(\widehat{S}_i))$, where $\widehat{S}_i$ is the estimate of the survival function at $t_i$ used in the Peto–Peto test [12]. It can be checked that, using this weighting scheme, weights are initially negative and then become positive after the value of $\widehat{S}_i$ falls below 0.69. It has been shown that this weighting scheme provides an optimal test against the specific alternative hypothesis

$$H_a : \Lambda_1(t) = (\Lambda_0(t))^b$$

where $\Lambda_0(t)$ and $\Lambda_1(t)$ are the cumulative hazard functions of the control and treatment groups, respectively, and $b \neq 1$ is a positive constant. The null hypothesis corresponds to the case when $b = 1$. The above alternative hypothesis implies that

$$h_1(t) = bh_0(t)(\Lambda_0(t))^{b-1}$$

Therefore, when $b > 1 (< 1)$, it is actually assumed that the hazard ratio is below (above) 1 up to a certain time point, and then above (below) 1 beyond that time point. The two hazard rates cross each other at the time point $\gamma$ satisfying the condition that $b(\Lambda_0(\gamma))^{b-1} = 1$.

The *third class* of methods employs the modelling approach. Anderson and Senthilselvan [3] considered the following model:

$$h_1(t) = h_0(t) \exp(\beta_1 I(t \leqslant \gamma) - \beta_2 I(t > \gamma)) \tag{2}$$

where $\beta_1$ and $\beta_2$ are two coefficients. Model (2) assumes that the log hazard ratio equals $\beta_1$ when $t \leqslant \gamma$, and changes to $-\beta_2$ when $t > \gamma$. To implement this approach, they first estimated $\gamma$ using the maximum-likelihood method, and then tested for $\beta$'s after replacing $\gamma$ by its estimate and treating the estimate as fixed. Obviously, this procedure is only approximately valid, because the substitution of $\gamma$ by its estimate would actually change the distribution of $\widehat{\beta}$'s. The main theoretical difficulty of this problem is that $\gamma$ is un-identifiable under the null hypothesis. To get rid of this difficulty, O'Quigley and Pessione [5] and O'Quigley [6] suggested some modifications using resampling techniques, which will be further discussed in the next subsection. Related research in this direction includes the general non-proportional hazard modelling approach suggested by O'Quigley and Pessione [13], of which the crossing hazard rates problem is a special case, and the recent procedure for testing erosion of regression effect suggested by O'Quigley and Natarajan [14]. An alternative modelling approach was suggested by Breslow *et al.* [4]. Their model is defined by

$$h_1(t) = h_0(t) \exp(\alpha + \beta z(t))$$

where $z(t_i) = i$ (i.e. rank score), or, $z(t_i) = \sum_{j \leqslant i} 1/n_j$ (i.e. cumulative hazard score). Based on this model, a test was suggested for testing the acceleration of the hazard function. Together with the logrank test, this test can be used for testing the difference between two crossing hazard rates.

Comparing the three classes of methods described above for handling the crossing hazard rates problem, we would expect that the second (i.e. those by choosing special weights) and the third (i.e. those based on models) classes of methods are more powerful for testing differences between two crossing hazard rates, because they are designed for testing the specific

 

crossing hazard rates alternative (i.e. the alternative in (1)), instead of some more general alternatives considered by the first class of methods. Between the second and third classes of methods, those model-based methods would have the advantage in accommodating covariates relatively easily. However, the two existing models described above have some obvious drawbacks. First, the assumptions required by these models might be problematic from theoretical viewpoint. For instance, the Anderson and Senthilselvan's model assumes that the hazard ratio changes its value abruptly at the crossing point, which may occur only in very limited situations. In most cases, the more conventional 'continuous hazards' assumption might be more appropriate (note that the hazard ratio is also continuous under this assumption). In Breslow *et al.*'s model, the function $z(t)$ is random, which may not be appropriate for describing a non-random hazard ratio. Second, both models assume very special parametric forms for the hazard ratio; these forms may not be flexible enough to include other crossing patterns that may be possible in applications.

### 2.3. The proposed method

Our method adopts the modelling approach because of its major benefits mentioned above, including the flexibility in accommodating covariate effects which is necessary in analysing the zinc nasal spray data. To this end, the following model is suggested for describing the hazard ratio:

$$h_1(t) = h_0(t)\exp\{\beta[\mathrm{BC}_\alpha(t) - \mathrm{BC}_\alpha(\gamma)]\} \tag{3}$$

where $\mathrm{BC}_\alpha(t)$ is a modified Box–Cox transformation of $t$, defined by

$$\mathrm{BC}_\alpha(t) = \begin{cases} t^\alpha & \text{if } \alpha \neq 0 \\ \log(t) & \text{if } \alpha = 0 \end{cases}$$

$\alpha$ and $\beta$ are two coefficients, $\gamma \in [0, \tau]$ is the crossing time point, and $[0, \tau]$ is the time range of interest. Note that the conventional Box–Cox transformation defines $\mathrm{BC}_\alpha(t) = (t^\alpha - 1)/\alpha$ when $\alpha \neq 0$. In the expression $\beta[\mathrm{BC}_\alpha(t) - \mathrm{BC}_\alpha(\gamma)]$ of (3), the constant term $-1/\alpha$ of $(t^\alpha - 1)/\alpha$ is cancelled out and the denominator $\alpha$ can be absorbed into $\beta$. Thus, the above more concise expression is adopted here, without losing any flexibility of the model.

Depending on the values of $\alpha$ and $\beta$, model (3) assumes that the hazard ratio is either below 1 when $t < \gamma$ and above 1 when $t > \gamma$, or the other way around. Clearly, this model allows the hazard ratio to change continuously over time, which is an advantage over the Anderson and Senthilselvan's model, as discussed at the end of Section 2.2. Unlike the Breslow *et al.*'s model, there are no random items used in (3). So, this model is well defined. In addition, this model allows more functional forms for the hazard ratio, compared to the two existing ones. Therefore, it can be applied to more applications. As an aside, it can be checked that model (3) with $\alpha = 0$ is equivalent to the model specified in the alternative hypothesis of the testing problem considered by Moreau *et al.* [11] in the case when the distribution of the survival time is Weibull.

Model (3) can be re-written as

$$h_x(t) = h_0(t)\exp\{\beta[\mathrm{BC}_\alpha(t) - \mathrm{BC}_\alpha(\gamma)]x\}$$

where $x = 0$ or 1, and it is a group indicator with 0 denoting the control group and 1 denoting the treatment group. This form of the model is convenient to use, especially when covariates

 

need to be considered. In the case when a $p$-dimensional covariate vector $\mathbf{Z}$ is involved, for instance, it can be easily accommodated into the model as follows:

$$h_x(t) = h_0(t)\exp\{\beta[\mathrm{BC}_\alpha(t) - \mathrm{BC}_\alpha(\gamma)]x + \theta'\mathbf{Z}\} \tag{4}$$

where $\theta$ is a $p \times 1$ coefficient vector.

Model (3) assumes that the hazard ratio has the same functional form before and after the crossing point. If for some reason we believe that the functional form of the hazard ratio could be different on two different sides of the crossing point, then model (3) can be generalized to

$$h_1(t) = h_0(t)\exp\{\beta_1[\mathrm{BC}_{\alpha_1}(t) - \mathrm{BC}_{\alpha_1}(\gamma)]I(t \leqslant \gamma) + \beta_2[\mathrm{BC}_{\alpha_2}(t) - \mathrm{BC}_{\alpha_2}(\gamma)]I(t > \gamma)\}$$

where $\alpha_1, \alpha_2, \beta_1$ and $\beta_2$ are coefficients. Of course, possible covariates can also be included in this model, as in model (4). Note that the above model includes the case when the two hazard rates are different before the crossing point and equal after the crossing point (e.g. the case when $\beta_2 = 0$) and the case when they are equal before the crossing point and different after the crossing point (e.g. the case when $\beta_1 = 0$).

To estimate parameters in model (3), we follow the idea proposed in Reference [15] to find the maximum-likelihood estimates in two steps. First, for given $\alpha$, we estimate $\beta$ and $\gamma$ values that maximize the likelihood, which can be accomplished by the Cox proportional hazards modelling incorporated in some standard statistical software packages (e.g. the function coxph() in S-plus or R, and PROC PHREG in SAS). In the second step we plot the maximized likelihood against $\alpha$ to identify the maximizer of $\alpha$. To ease computer implementation, the grid search algorithm can also be employed for finding the grid point of $\alpha$ resulting in the maximum likelihood, as the final estimate of $\alpha$. Then, the point estimates of $\beta$ and $\gamma$ can be obtained simultaneously.

It can be seen that hypothesis testing of (1) is equivalent to testing of

$$H_0 : \beta = 0 \quad \text{vs} \quad H_a : \beta \neq 0 \tag{5}$$

If $\alpha$ and $\gamma$ in model (3) are known, then testing for (5) can be accomplished by some standard tests, such as the Wald test. In applications, however, $\alpha$ and $\gamma$ are usually unknown. In such cases, we can adopt the strategy of Reference [3], by performing the Wald test for $\beta$ after $\alpha$ and $\gamma$ are replaced by their estimates. However, according to several authors, including [6], and our own numerical experience, this approach fails to control the type I error properly, mainly because the variability of $\widehat{\beta}$ is underestimated after $\alpha$ and $\gamma$ are replaced by their estimates. Theoretically, the real problem is that, under $H_0$, all parameters disappear from the model, and thus, $\alpha$ and $\gamma$ are unidentifiable. In such cases, the likelihood ratio test cannot be used either, because the number of constraints under $H_0$ and the number of parameters disappearing from the model when $H_0$ is true are different, and consequently it is hard to figure out the appropriate degrees of freedom for the test.

Davies [16, 17] proposed a method for handling a similar problem, by which the related test statistic was first computed based on fixed values of the nuisance parameter, and then the maximum of the test statistic values over the range of possible values of the nuisance parameter was used for testing purposes. This method was applied to the crossing hazard rates problem by O'Quigley and Pessione [5] when they tried to derive an appropriate testing procedure for Anderson and Senthilselvan's modelling approach. In addition, O'Quigley and Pessione suggested a direct bootstrap method and showed that it was appropriate for testing

 

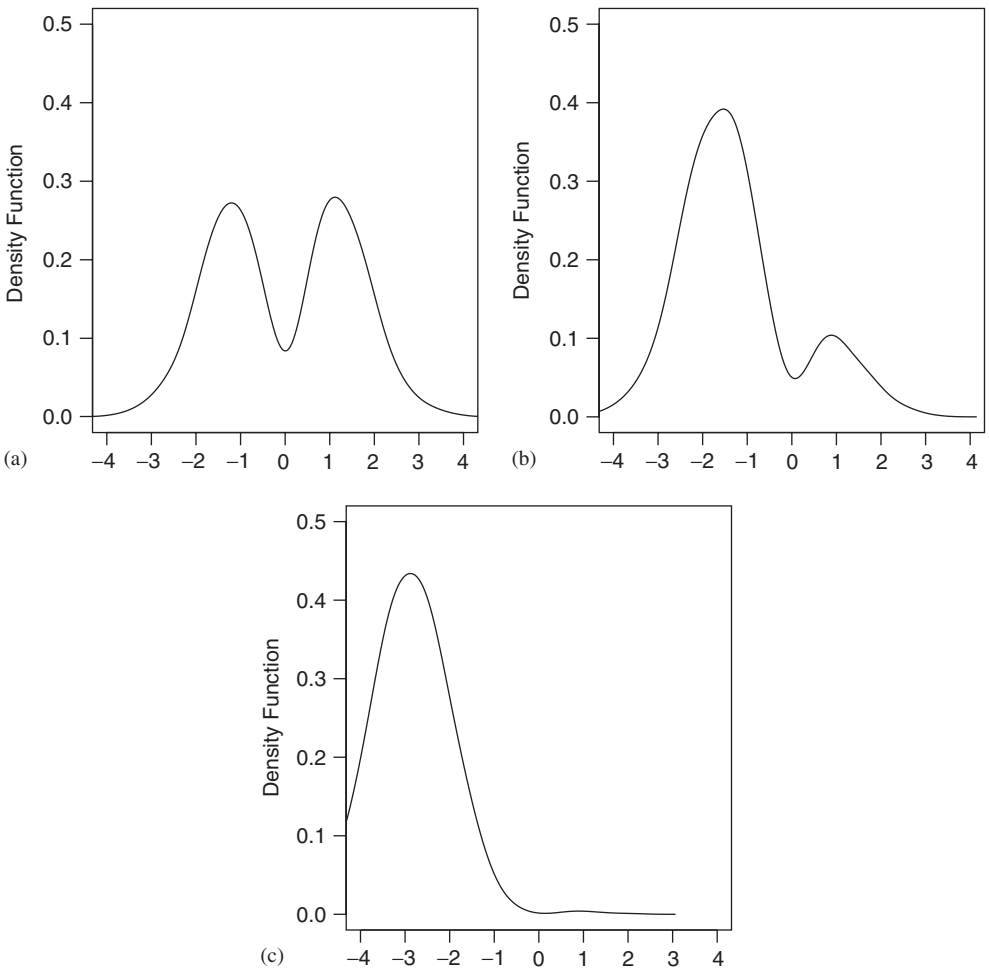


Figure 2. Plots (a)–(c) show the bootstrap estimates of the distribution of $\widehat{\beta}$ when the control distribution is Weibull$(2.0, 1.0)$ and the treatment distribution is Weibull$(2.0, 1.0)$, Weibull$(1.8, 1.0)$ and Weibull$(1.5, 1.0)$, respectively.

the crossing hazard rates problem formulated by Anderson and Senthilselvan's model. We tried Davies's method in our case; but, the results were not satisfactory. However, from simulations, we found that the direct bootstrap method performed well in our case after it is adapted to our modelling approach.

The adapted direct bootstrap method works as follows. From bootstrap samples obtained from the original data, we can obtain a bootstrap estimate of the distribution of $\widehat{\beta}$. As O'Quigley and Pessione [5] observed, the limit distribution of this bootstrap estimate is bimodal and symmetric about zero when $H_0$ is true, which is demonstrated by Figure 2. In this figure, plot (a) shows the density of 1000 bootstrap estimates of $\beta$ when the treatment and control populations has the same distribution Weibull$(2.0, 1.0)$. In this case, $H_0$ holds and

 

it can be seen that the distribution of $\widehat{\beta}$ is indeed bimodal and symmetric about zero. Plots (b) and (c) show the same results when the control distribution is Weibull$(2.0, 1.0)$ and the treatment distribution is Weibull$(1.8, 1.0)$ and Weibull$(1.5, 1.0)$, respectively. It can be seen that, when moving away from $H_0$, bimodality of the distribution of $\widehat{\beta}$ is still maintained; but, the distribution becomes skewed. This finding is the major motivation for O'Quigley and Pessione to propose a non-parametric test of $H_0$ based on the test statistic $T = \min\{B^*(0), 1-B^*(0)\}$, where $B^*(u) = \int_{-\infty}^{u} b^*(w)\,\mathrm{d}w$ and $b^*$ is the empirical bootstrap density of $\widehat{\beta}$. It can be easily checked that $T$ reaches its maximum when $H_0$ holds, and it becomes smaller when moving away from $H_0$. Thus, by this testing procedure, the null hypothesis is rejected when $T$ is small. Simulations from both O'Quigley and Pessione [5] and ours show that, under $H_0$, $T$ is uniformly distributed over the range $[0, 0.5]$. Therefore, $p$-value of the direct bootstrap procedure is simply $2t$, where $t$ is the observed value of the test statistics $T$. In other words, if we resample from the treatment and control samples, respectively, for $n_B$ times, we can compute the $p$-value by $2\min(n_B^+, n_B^-)/n_B$, where $n_B^+$ and $n_B^-$ denote the numbers of positive and negative values of the test statistic computed from the $n_B$ bootstrap samples.

The confidence interval for $\gamma$ can also be constructed by the bootstrap method as follows. From the given data, we draw $n_B$ bootstrap samples, as described above. Then, $n_B$ values of the point estimate of $\gamma$ can be computed, from which a confidence interval for $\gamma$ can be constructed. In order to ensure that the estimate of $\gamma$ is non-negative, we suggest that in constructing the confidence interval for $\gamma$, $\gamma$ is transformed to $\log(\gamma)$ first, prior to application of the bootstrap method.

At the end of this section, we would like to point out that, like other modelling procedures for handling the crossing hazard rates problem, the proposed procedure only considers the null hypothesis of equal hazard rates and the alternative hypothesis of crossing hazard rates in (5). In other words, some realistic cases, e.g. the cases when the two hazard rates are parallel to each other or when they are neither parallel nor crossing, are not covered by its null and alternative hypotheses. Therefore, this procedure is appropriate to use only in cases when our major goal is to test the existence of crossing in the two hazard rates and construct a confidence interval for the crossing time point when there is evidence for a crossing point based on visual display of the two hazard rates. One such case is the zinc nasal spray example mentioned in Section 1, in which the proposed procedure is appropriate to use, because Figure 1 suggests that the two hazard rates may cross each other. In such cases, our procedure can be used to test whether the crossing is real or it is caused by random variation.

## 3. A SIMULATION STUDY

In this section, we present some simulation results to compare the performance of the proposed procedure with some existing ones. In the literature, most existing methods for comparing crossing hazard rates are discussed when there is no covariate involved. For this reason, the proposed procedure (PP) is compared to the existing methods in this setting first. The existing procedures considered here include: the modified Kolmogorov–Smirnov procedure (KS), the Renyi-type test (RE), the procedure by Mantel and Stablein ([10], MS), the procedure by Lin and Wang ([9], LW), the modelling approach by Anderson and Senthilselvan ([3], AS), and the modelling approach by Breslow *et al.* ([4], BR). All these procedures are

 

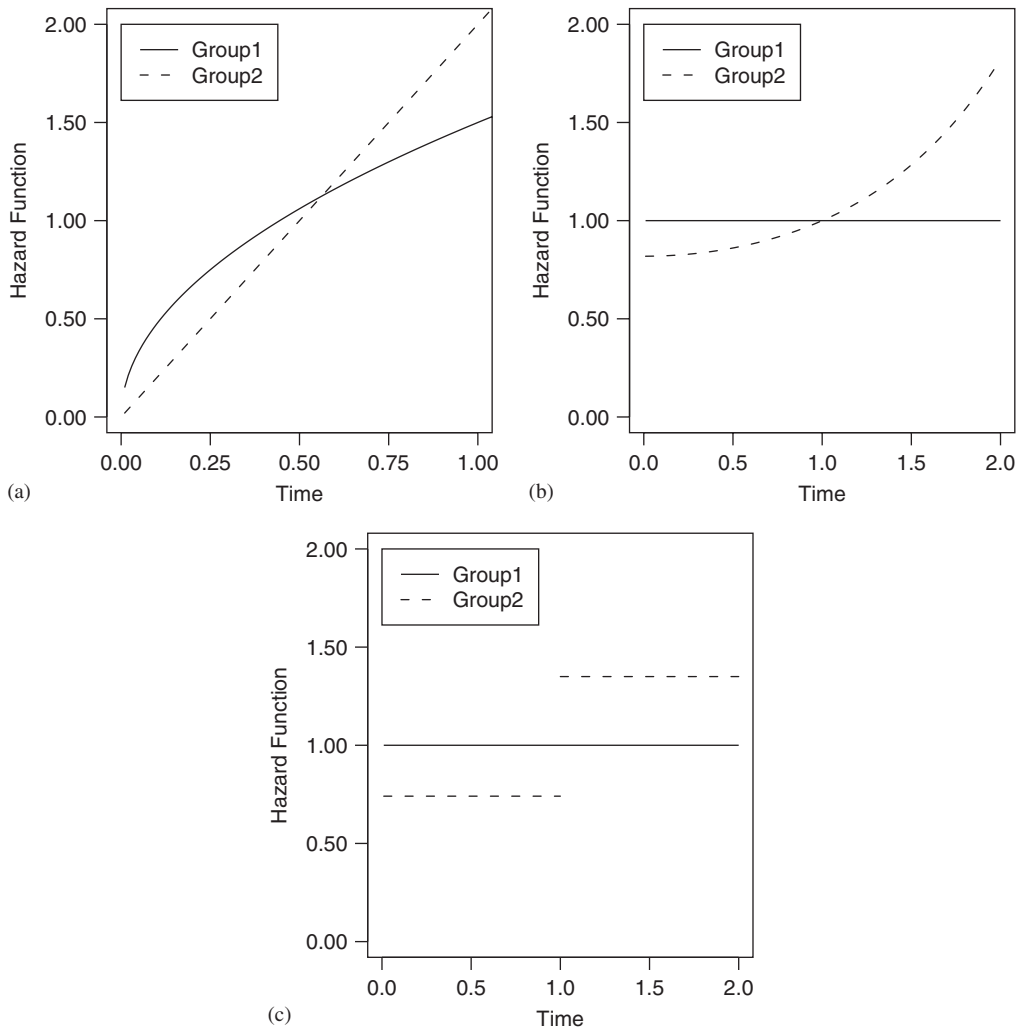


Figure 3. Plots (a)–(c) show the hazard rates in cases I–III, respectively.

described in Section 2.2; they cover all three classes of methods for handling crossing hazard rates. Breslow *et al.*'s procedure has two versions, i.e. the one based on rank score and the one based on cumulative hazard score. They are denoted by BR1 and BR2, respectively. To see how traditional testing procedures perform when two hazard rates cross, three routinely used ones are also included here, which are the logrank (LR), Gehan–Wilcoxon (GW) and Peto–Peto (PE) procedures. For all the procedures mentioned above, only procedures PP, AS and MS provide point estimates of the crossing time $\gamma$ explicitly.

We consider three different patterns of crossing hazard rates here, which are illustrated by Figure 3(a)–(c), respectively. In case I, the survival time has Weibull$(1.5, 1.0)$ distribution in the control group and Weibull$(2.0, 1.0)$ distribution in the treatment group, and the true

 

crossing point is $\gamma = 0.563$. In this case, the shape of the two hazard rates look similar to that in the zinc data (cf. Figure 1). In case II, the survival time has Exponential(1.0) distribution in the control group and the hazard of the treatment group is defined by model (3) with $\alpha = 2.0$, $\beta = 0.2$ and $\gamma = 1.0$. In case III, the survival time still has Exponential(1.0) distribution in the control group; but, in the treatment group, its hazard rate is defined by the Anderson and Senthilselvan's model (2) with $\beta_1 = \beta_2 = -0.3$ and $\gamma = 1.0$. These three cases are chosen under the following considerations. The first case is similar to what happened in the zinc clinical trial, which motivates the current research. The second case can be described well by our model (3), but not the two existing models described in Section 2.2. So, this case is favourable to the proposed procedure. The third case is specified by Anderson and Senthilselvan's [3] model (2). Although this case may not be realistic in applications due to a jump in hazard rate at $t = 1.0$, it is considered here to see how the proposed procedure performs in cases favourable to Anderson and Senthilselvan's model.

In each case, two groups of survival data are generated as follows. Each group has 100 observations with the censoring rate chosen as either 0 per cent (i.e. no censoring) or 20 per cent. Censoring time is uniformly distributed on $[0, \tau]$, where $\tau$ is the maximum survival time observed in the simulated data, and the censoring time is independent of the survival time.

For each method, we computed its actual size (i.e. probability of type I error) and power. For the three methods of PP, AS and MS, MSE and Bias of their point estimates of the crossing point are also provided. Note that point estimates of procedures MS and AS can only be presented in intervals, since these two procedures search for their point estimates at the observed survival times only and all points in the interval of two consecutive survival times would give the same value of the related test statistic. When we compute the MSE and Bias values for these two procedures, middle points of the obtained intervals are used as their point estimates. Computation of the size and power of the procedure PP is based on the direct bootstrap method with $n_B = 1000$, as discussed at the end of Section 2.3.

Tables I–III summarize the results in cases I–III, respectively, based on 1000 replications. From Table I, we can observe that: (1) the third class methods tend to have the best performance, followed by the second class methods, and then by the first class methods; (2) the proposed procedure PP is superior to the other modelling procedures in terms of power, MSE, and Bias for this particular case; (3) the three traditional tests LR, GW, and PE perform noticeably worse than the other methods considered when the two hazard rates cross; and (4) it seems that type I error is controlled well by all methods considered. Results in Table II show similar findings. As expected, in case III, we can see from Table III that the proposed procedure PP performs slightly worse than procedures AS and MS; it has similar power to those of procedures BR1 and BR2; and it performs better than the remaining procedures.

Next, we consider the case when one binary covariate $Z$ is included in model (4) with $\theta$ as its coefficient. When $Z = 0$, the two hazard rates are chosen to be those shown in Figure 3(a). Parameter $\theta$ is chosen to be $\log(1.5), \log(2.0)$, or $\log(2.5)$. Please note that the two hazard rates cross at the same position in the two cases when $Z = 1$ and 0, because inclusion of the covariate $Z$ does not change the ratio of the two hazard rates. As before, the censoring rate is chosen to be either 0 per cent or 20 per cent in this example, and the sample size $n$ is fixed at 100. In each group of the data, the first half observations are assigned the value of $Z = 0$, and the second half are assigned the value of $Z = 1$. Results based on 1000 replications are summarized in Table IV. From the table, it can be seen that both the power of the proposed

 

Table I. Powers and sizes of various methods for comparing two hazard rates in case I. For procedures PP, MS, and AS, MSEs and Biases of the point estimates of the crossing time are also provided.

| Method | Censoring (%) | Power | Size | MSE | Bias |
|---|---|---|---|---|---|
| PP | 0 | 0.937 | 0.053 | 0.017 | 0.003 |
| | 20 | 0.796 | 0.049 | 0.022 | 0.003 |
| KS | 0 | 0.305 | 0.050 | | |
| | 20 | 0.318 | 0.048 | | |
| RE | 0 | 0.179 | 0.046 | | |
| | 20 | 0.156 | 0.051 | | |
| MS | 0 | 0.888 | 0.053 | 0.038 | 0.016 |
| | 20 | 0.722 | 0.052 | 0.042 | 0.003 |
| LW | 0 | 0.612 | 0.045 | | |
| | 20 | 0.367 | 0.044 | | |
| AS | 0 | 0.850 | 0.037 | 0.034 | 0.012 |
| | 20 | 0.693 | 0.046 | 0.038 | −0.002 |
| BR1 | 0 | 0.904 | 0.054 | | |
| | 20 | 0.796 | 0.052 | | |
| BR2 | 0 | 0.806 | 0.049 | | |
| | 20 | 0.696 | 0.049 | | |
| LR | 0 | 0.165 | 0.050 | | |
| | 20 | 0.071 | 0.053 | | |
| GW | 0 | 0.116 | 0.051 | | |
| | 20 | 0.202 | 0.050 | | |
| PE | 0 | 0.116 | 0.051 | | |
| | 20 | 0.151 | 0.051 | | |

test for $\beta$ and the accuracy of the point estimate of $\gamma$ are affected just slightly by the inclusion of the covariate.

## 4. APPLICATIONS TO A KIDNEY DIALYSIS PATIENTS DATA AND THE ZINC NASAL SPRAY DATA

In this section, we apply the methods for handling the crossing hazard rates problem to two real-data examples. One is the kidney dialysis patients data described in details by Klein and Moeschberger ([1], Section 1.4), and the other is the zinc nasal spray data as described in Section 1.

We first discuss the kidney dialysis patients data which were taken from a study designed to assess the time to first exit-site infection (in months) in 119 patients with renal insufficiency, among which 43 patients utilized a surgically placed catheter (Group 1) and 76 patients utilized a percutaneous placement of their catheter (Group 2). Catheter failure was the primary reason for censoring. There are 27 censored observations in Group 1 and 65

 

Table II. Powers and sizes of various methods for comparing two hazard rates in case II. For procedures PP, MS, and AS, MSEs and Biases of the point estimates of the crossing time are also provided.

| Method | Censoring (%) | Power | Size | MSE | Bias |
|---|---|---|---|---|---|
| PP | 0 | 0.987 | 0.052 | 0.108 | −0.012 |
| | 20 | 0.804 | 0.050 | 0.122 | −0.016 |
| KS | 0 | 0.261 | 0.047 | | |
| | 20 | 0.205 | 0.047 | | |
| RE | 0 | 0.170 | 0.049 | | |
| | 20 | 0.135 | 0.046 | | |
| MS | 0 | 0.727 | 0.050 | 0.194 | −0.113 |
| | 20 | 0.478 | 0.047 | 0.212 | −0.133 |
| LW | 0 | 0.876 | 0.040 | | |
| | 20 | 0.522 | 0.039 | | |
| AS | 0 | 0.631 | 0.038 | 0.182 | −0.115 |
| | 20 | 0.403 | 0.039 | 0.198 | −0.128 |
| BR1 | 0 | 0.821 | 0.050 | | |
| | 20 | 0.552 | 0.045 | | |
| BR2 | 0 | 0.951 | 0.046 | | |
| | 20 | 0.761 | 0.046 | | |
| LR | 0 | 0.092 | 0.050 | | |
| | 20 | 0.045 | 0.049 | | |
| GW | 0 | 0.136 | 0.051 | | |
| | 20 | 0.167 | 0.049 | | |
| PE | 0 | 0.136 | 0.051 | | |
| | 20 | 0.151 | 0.049 | | |

censored observations in Group 2. These data were also analysed by Lin and Wang [9], from which it can be seen that the two survival functions cross at a quite early time. The methods discussed in the previous section are then applied to this example to test for equality of the two hazard rates. The results are presented in Table V. For methods PP, MS, and AS, point estimates (for procedures MS and AS) or 90 per cent confidence interval (for procedure PP) of the crossing point $\gamma$ are also provided, besides $p$-values of the related hypothesis tests. Point estimates of the procedures MS and AS can only be given by intervals, as pointed out in Section 3. O'Quigley and Pessione's [5] direct bootstrap method is used in computing the $p$-value of procedure PP, as described in Section 2.3. As observed by Lin and Wang [9], the two survival curves in this example appear to be quite different. Thus, it is not surprising to see from Table V that almost all methods, except the three traditional tests and the first class method RE, yield significant results at the 0.05 level. However, the proposed method provides the smallest $p$-value.

We now turn to the zinc nasal spray data. The estimates of two hazards of the treatment and control groups are presented in Figure 1, which shows that the two hazard rates cross each other around the sixth day. Based on some conventional tests without taking the crossing

 

Table III. Powers and sizes of various methods for comparing two hazard rates in case III. For procedures PP, MS, and AS, MSEs and Biases of the point estimates of the crossing time are also provided.

| Method | Censoring (%) | Power | Size | MSE | Bias |
|---|---|---|---|---|---|
| PP | 0 | 0.697 | 0.052 | 0.122 | −0.035 |
| | 20 | 0.554 | 0.050 | 0.152 | −0.055 |
| KS | 0 | 0.514 | 0.047 | | |
| | 20 | 0.496 | 0.047 | | |
| RE | 0 | 0.349 | 0.049 | | |
| | 20 | 0.381 | 0.046 | | |
| MS | 0 | 0.726 | 0.050 | 0.061 | −0.004 |
| | 20 | 0.600 | 0.047 | 0.088 | −0.015 |
| LW | 0 | 0.145 | 0.040 | | |
| | 20 | 0.100 | 0.039 | | |
| AS | 0 | 0.701 | 0.038 | 0.049 | −0.004 |
| | 20 | 0.583 | 0.039 | 0.067 | −0.011 |
| BR1 | 0 | 0.658 | 0.050 | | |
| | 20 | 0.525 | 0.045 | | |
| BR2 | 0 | 0.600 | 0.046 | | |
| | 20 | 0.540 | 0.046 | | |
| LR | 0 | 0.083 | 0.050 | | |
| | 20 | 0.126 | 0.049 | | |
| GW | 0 | 0.387 | 0.051 | | |
| | 20 | 0.419 | 0.049 | | |
| PE | 0 | 0.387 | 0.051 | | |
| | 20 | 0.403 | 0.049 | | |

Table IV. We consider the case when model (4) includes a binary covariate $Z$ with coefficient $\theta$ fixed at $\log(1.5), \log(2.0)$, and $\log(2.5)$, respectively. This table presents the power of the proposed testing procedure for $\beta$, and MSEs and Biases of the point estimates of $\gamma$ and $\theta$.

| $\theta$ | Censoring (%) | Power | MSE of $\hat{\gamma}$ | Bias of $\hat{\gamma}$ | MSE of $\hat{\theta}$ | Bias of $\hat{\theta}$ |
|---|---|---|---|---|---|---|
| log(1.5) | 0 | 0.808 | 0.023 | −0.017 | 0.020 | 0.029 |
| | 20 | 0.689 | 0.025 | −0.031 | 0.027 | 0.040 |
| log(2.0) | 0 | 0.785 | 0.023 | −0.009 | 0.023 | 0.030 |
| | 20 | 0.647 | 0.026 | −0.030 | 0.028 | 0.041 |
| log(2.5) | 0 | 0.796 | 0.023 | −0.011 | 0.024 | 0.025 |
| | 20 | 0.650 | 0.026 | −0.018 | 0.030 | 0.039 |

 

Table V. $p$-values for testing crossing hazard rates, and point estimates (for procedures MS and AS) or 90 per cent confidence interval (for procedure PP) of the crossing time point $\gamma$, provided by various methods when they are applied to the kidney dialysis patients data.

| Method | $p$-value | Cross point estimate |
|---|---|---|
| PP | 0.001 | (1.171, 4.693) |
| KS | 0.031 | |
| RE | 0.220 | |
| MS | 0.006 | $\widehat{\gamma} \in [2.5, 3.5]$ |
| LW | 0.012 | |
| AS | 0.004 | $\widehat{\gamma} \in [2.5, 3.5]$ |
| BR1 | 0.007 | |
| BR2 | 0.026 | |
| LR | 0.112 | |
| GW | 0.964 | |
| PE | 0.237 | |

Table VI. $p$-values for testing crossing hazard rates, and point estimates (for procedures MS and AS) or 90 per cent confidence interval (for procedures PP and PP(b)) of the crossing time point $\gamma$, provided by various methods when they are applied to the zinc nasal spray data.

| Method | $p$-value | Cross point estimate |
|---|---|---|
| PP | 0.024 | (4.555, 7.956) |
| PP(b) | 0.028 | (3.714, 7.611) |
| KS | 0.948 | |
| RE | 0.750 | |
| MS | 0.508 | $\widehat{\gamma} \in [5.0, 6.0]$ |
| LW | 0.206 | |
| AS | 0.053 | $\widehat{\gamma} \in [5.0, 6.0]$ |
| BR1 | 0.072 | |
| BR2 | 0.084 | |
| LR | 0.953 | |
| GW | 0.534 | |
| PE | 0.466 | |

hazards into consideration, Belongia *et al.* [2] did not find any significant difference of cold duration between the two hazards. Here, we re-analyse these data by using various methods developed for comparing the crossing hazards along with three traditional methods LR, GW, and PE. The results are presented in Table VI, which also includes the $p$-value for hazard comparison after adjustment of the baseline symptom score when using method PP (denoted as PP(b)). From Table VI, we can see that method PP provides quite significant results, while method AS gives only marginally significant result, and the remaining methods all fail to detect the difference between the two hazard rates. Results obtained from our proposed approach provide evidence that zinc nasal spray may have some antiviral effect early on, which is consistent with the conclusions drawn from the comparison of daily symptom scores in Belongia *et al.* [2]. A possible explanation of early treatment effect is as follows. The common cold is mainly caused by a viral infection in the nose, and it has been shown in

 

some studies that zinc has a direct antiviral effect. Thus, applying zinc directly to the entrance of nose would be more effective in the early stage of cold when only small amount of viruses enter the nasal area, especially when most viruses have not penetrated into host cells. After viruses overcome the body's defence system and penetrate into host cells, application of zinc nasal spray would have little or no effect. Thus, the method proposed in this paper could help us formulate new research hypotheses for further investigation about the potential effect of zinc nasal spray on the common cold, which may be missed otherwise.

## 5. CONCLUDING REMARKS

We have proposed a model-based approach for testing equality of two crossing hazard rates and for constructing a confidence interval for the crossing point when there is evidence for such a crossing point based on visual display of the two hazard rates. This method can incorporate possible effects of covariates as well. The Box–Cox transformation embedded in our model enables us to capture different crossing patterns. On the other hand, more parameters are included in our model, compared to some existing methods handling this problem, which may cause some complexity in implementation and lead to a loss in power in certain situations.

The crossing hazard rates problem considered in this paper can be further explored. For instance, the hypotheses in (1) do not cover all possible cases in applications, such as cases when the two hazard rates are neither crossing nor equivalent within the interested time interval, or when one hazard rate is greater than the other initially and then both become equal. Appropriate procedures are needed to test for treatment effect under these situations. In some cases, we may also be interested in testing whether the two hazard rates cross each other more than once, which is completely excluded from the discussion of the current paper. When the hazard rates cross more than once, the corresponding survival functions should be close to each other, and the testing problem would become much more complex as the number of crossings increases. In such cases, it is also unclear whether the bootstrap procedure described in Section 2.3 is still appropriate. All these issues require much future research.

The proposed procedure still has room for improvement. For instance, although model (3) is much more flexible than the ones used by Anderson and Senthilselvan [3] and Breslow *et al.* [4], its parametric form may still exclude some possible crossing patterns in applications. As an example, if the true log hazard ratio increases at the exponential rate after the crossing time point, then the current model (3) is unable to capture this crossing pattern. In such a case, Manly's [18] exponential transformation might be preferred, compared to the Box–Cox power transformation used in model (3). All these issues should be addressed in the future research.

ACKNOWLEDGEMENTS

We thank two anonymous reviewers for many constructive comments and suggestions, which greatly improve the quality of the paper.

REFERENCES

1. Klein JP, Moeschberger ML. *Survival Analysis*. Springer: New York, 1997.
2. Belongia EA, Berg R, Liu K. A randomized trial of zinc nasal spray for the treatment of upper respiratory illness in adults. *The American Journal of Medicine* 2001; **111**:103–108.

 

3. Anderson JA, Senthilselvan A. A two-step regression model for hazard functions. *Applied Statistician* 1982; **31**:44–51.
4. Breslow NE, Edler L, Berger J. A two–sample censored-data rank test for acceleration. *Biometrics* 1984; **40**:1049–1062.
5. O'Quigley J, Pessione F. The problem of a covariate-time qualitative interaction in a survival study. *Biometrics* 1991; **47**:101–115.
6. O'Quigley J. On a two-sided test for crossing hazard rates. *The Statistician* 1994; **43**:563–569.
7. Fleming TR, O'Fallon JR, O'Brien PC, Harrington DP. Modified Kolmogorov–Smirnov test procedures with application to arbitrarily right-censored data. *Biometrics* 1980; **36**:607–625.
8. Gill RD. Censoring and stochastic integrals. *Mathematical Centre Tracts*, vol. 124. Mathematical Centrum: Amsterdam, 1980.
9. Lin X, Wang H. A new testing approach for comparing the overall homogeneity of survival curves. *Biometrical Journal* 2004; **46**:489–496.
10. Mantel N, Stablein DM. The crossing hazard function problem. *The Statistician* 1988; **37**:59–64.
11. Moreau T, Maccario J, Lellouch J, Huber C. Weighted log rank statistics for comparing two distributions. *Biometrika* 1992; **79**:195–198.
12. Peto R, Peto J. Asymptotically efficient rank invariant test procedures. *Journal of the Royal Statistical Society* (*Series A*) 1972; **135**:185–206.
13. O'Quigley J, Pessione F. Score test for homogeneity of regression effect in the proportional hazards model. *Biometrics* 1989; **45**:135–144.
14. O'Quigley J, Natarajan L. Erosion of regression effect in a survival study. *Biometrics* 2004; **60**:344–351.
15. Box GEP, Cox DR. An analysis of transformation. *Journal of the Royal Statistical Society* (*Series A*) 1964; **26**:211–252.
16. Davies RB. Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika* 1977; **64**:247–254.
17. Davies RB. Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika* 1987; **74**:33–43.
18. Manly BFJ. Exponential data transformations. *The Statistician* 1976; **25**:37–42.

 