A Two-Stage Procedure For Comparing Hazard Rate Functions

Peihua Qiu[1] and Jun Sheng[2]

[1]School of Statistics, University of Minnesota, 313 Ford Hall,
224 Church St. SE, Minneapolis, MN 55455, USA.

[2]Bristol West Insurance Group, 5301 W Cypress St., Suite 100,
Tampa, FL 33607, USA.

Abstract

Comparison of two hazard rates is important in applications related to times to occurrence of a specific event. Conventional comparison procedures, such as the logrank, Gehan-Wilcoxon, and Peto-Peto tests, are powerful only when the two hazard rates do not cross each other. Because crossing hazard rates are common in practice, a number of procedures have been proposed in the literature for comparing such rates. However, most of these procedures only consider the alternative hypothesis with crossing hazard rates; many other realistic cases, including those when the two hazard rates run parallel to each other, are excluded from consideration. In this paper, we propose a two-stage procedure that considers all possible alternatives, including ones with crossing or running parallel hazard rates. To define its significance level and p-value properly, a new procedure for handling the crossing hazard rates problem is suggested, which has the property that its test statistic is asymptotically independent of the test statistic of the logrank test. We show that the two-stage procedure, with the logrank test and the suggested procedure for handling the crossing hazard rates problem used in its two stages, performs well in applications in comparing two hazard rates.

Key Words: Additive tests; Censoring; Crossing hazard rates; Power; Proportional hazards regression; Resampling; Sequential tests; Significance level; Survival analysis.

1 Introduction

To evaluate treatment effect in cases of survival data, we often need to compare two hazard rate functions of the treatment and control groups (cf., e.g., Lawless 1982, Bain and Engelhardt 1991, Klein and Moeschberger 1997). For this purpose, the logrank, Gehan-Wilcoxon, and Peto-Peto tests, among several others, are routinely used in practice (Klein and Moeschberger 1997, Chapter

7). It is well known that the logrank test is optimal when the two hazard rates are proportional. However, the assumption of proportional hazard rates is violated when the two hazard rates cross each other. In such cases, it has been well demonstrated that the conventional procedures would perform poorly (see e.g., Breslow *et al.* 1984, O'Quigley and Pessione 1989, 1991, Moreau *et al.* 1992, O'Quigley 1994, Lin and Wang 2004, Liu *et al.* 2006). This paper suggests a two-stage procedure for comparing two hazard rate functions efficiently in both cases when they cross and when they are different but not crossing.

The phenomenon of crossing hazard rates is common in applications. In some cases, the treatment has benefits only in the early stage of a disease and it does not have any long-term advantages. As an example, radiation and chemotherapy can usually improve patients' prospects for short-term survival; but they have little or no long-term medical benefits. In some other cases, treatments have benefits in the long run; they may increase the risk in the early stage after the treatment is applied. Surgery is a good example of this type of medical treatment. Due to infection and other short-term risks, it may cause high mortality in a short period after surgery; but, in the long run, surgery will often improve a patient's long-term health. In all these cases, the two related hazard rates often cross each other. One other real example with crossing hazard rates will be discussed in Section 5.

In the literature, there are a number of procedures for handling the crossing hazard rates problem, which can be roughly classified into the following three groups. The first group of methods tries to avoid early differences between the two hazard rates being canceled out by later differences of opposite sign, which often occurs when a traditional test such as the logrank test is applied to a case with crossing hazard rates. To this end, they usually define their test statistics using absolute or squared differences between the two hazard rates (cf., e.g., Fleming *et al.* 1980, Lin and Wang 2004). The second group of methods handles the crossing hazard rates problem by choosing special weights in the weighted logrank test, which change signs before and after a potential crossing point. See, e.g., Mantel and Stablein (1988) and Moreau *et al.* (1992) for different weighting schemes. The third group of methods employs the modeling approach, by including explicitly the crossing structure of the hazard rates in a model (cf., e.g., Anderson and Senthilselvan 1982, Breslow *et al.* 1984, Liu *et al.* 2007). Comparing the three groups of methods described above, one would expect that the second and third groups of methods are more powerful in testing differences between two crossing hazard rates, because they are designed specifically for testing the crossing hazard rates

alternative, instead of some more general alternatives considered by the first group of methods. Between the second and third groups of methods, those model-based methods have the advantage that they can accommodate covariates easily. For recent development on estimation of hazard rates, see Cheng *et al.* (2006) and the references cited therein.

In most procedures for handling the crossing hazard rates problem, the null hypothesis considered is that the two hazard rates are the same, and the alternative hypothesis is that they cross each other at an unknown crossing point. Obviously, this formulation of the hypotheses excludes some important cases, e.g., cases when two hazard rates are different but not crossing. Furthermore, because these procedures are designed specifically for detecting crossings, they are not as powerful as we would expect, for detecting other differences between two hazard rates (e.g., the difference between two running parallel hazard rates).

To detect arbitrary difference between two hazard rates, we suggest a procedure with two stages. In the first stage, a conventional procedure (e.g., the logrank test) is applied, to detect all kinds of differences between the two hazard rates, except certain crossings. In the second stage, a procedure for detecting crossings is applied. The overall significance level of the two-stage procedure is controlled at a given level α. In this procedure, if an arbitrary conventional procedure and an arbitrary procedure for handling the crossing hazard rates problem are used in its two stages, then the two individual tests are usually correlated. Consequently, its significance level and p-value are difficult to define and compute. To overcome this difficulty, in this paper, we suggest a new procedure for handling the crossing hazard rates problem, which has the property that its test statistic is asymptotically independent of the test statistic of the logrank test.

The remaining part of the article is organized as follows. The two-stage procedure is described in detail in Section 2. Definitions and computation of its significance level and p-value are discussed in Section 3. A simulation study is presented in Section 4. Applications to two real datasets are discussed in Section 5. Several remarks conclude the article in Section 6. The proof of a theorem is given in Appendix. Computer codes in R that are used in the numerical studies of this paper are available from the first author; they can also be downloaded from the Datasets Website of the Royal Statistical Society, at the address http://www.blackwellpublishing.com/rss/SeriesB.htm.

2 The Two-Stage Procedure

2.1 Framework of the two-stage procedure

Let h_0 and h_1 be the hazard rate functions of survival times of subjects in the control and treatment groups, respectively, and let $[0, \tau]$ be the time range of interest. Then, we are interested in testing the following hypotheses:

$$\begin{aligned} & H_0 : h_1(t) = h_0(t), \text{ for all } t \in [0, \tau] \\ \text{vs.} \quad & H_a : h_1(t) \neq h_0(t), \text{ for some } t \in [0, \tau]. \end{aligned} \tag{2.1}$$

In (2.1), the alternative hypothesis includes all possible patterns that h_0 and h_1 are different, including the crossing and running parallel patterns.

As described in Section 1, the proposed procedure for testing the hypotheses in (2.1) is sequential and it consists of two stages. In stage one, a conventional procedure, such as the logrank, Gehan-Wilcoxon, and Peto-Peto tests, is applied. If we obtain a rejection of the null hypothesis in stage one, then the entire procedure ends, and we conclude that the two hazard rates are significantly different. Otherwise, we proceed to stage two, by applying a procedure for handling the crossing hazard rates problem, from which we can distinguish cases when the two hazard rates are identical from cases when they cross each other. The entire two-stage testing procedure is demonstrated in Figure 2.1.

It should be pointed out that the order of the two stages in the proposed two-stage testing procedure might be altered. The ordering demonstrated by Figure 2.1 is chosen here based on the following consideration. In applications, it might be true that the majority number of cases when the two hazard rates are different does not have crossings involved (i.e., the "different but not crossing" cases), although the crossing hazard rates phenomenon is also quite common. By using the proposed two-stage procedure, these "different but not crossing" cases can mostly be distinguished in the first stage alone, and consequently the whole testing procedure would end immediately after stage one; thus, the testing task is simplified in such cases using the proposed proposal, compared to the alternative proposal with the order of the two stages reversed.

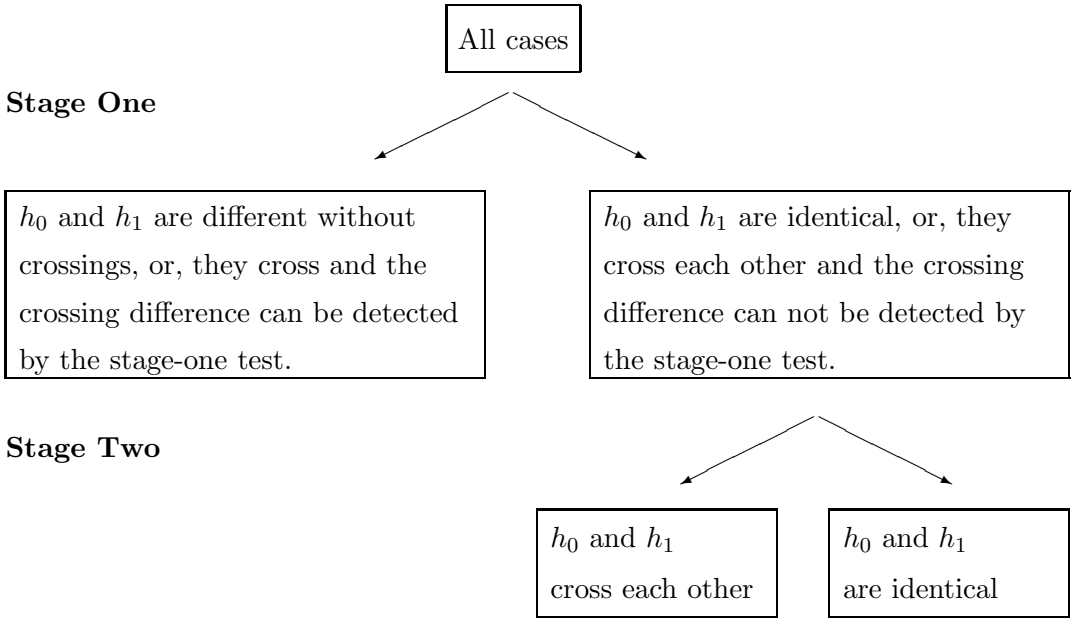

Figure 2.1: Diagram for demonstrating the proposed two-stage testing procedure for comparing two hazard rates.

2.2 A new procedure for handling the crossing hazard rates problem

Theoretically speaking, any conventional test can be used in the first stage of the proposed two-stage procedure, and any procedure for handling the crossing hazard rates problem can be used in its second stage. However, when the two tests are not independent, it is often difficult to define the significance level and p-value of the entire two-stage procedure, as mentioned in Section 1. To overcome this difficulty, in this part, we propose a new procedure for handling the crossing hazard rates problem, which has the property that it is asymptotically independent of the conventional logrank test.

For $j = 1, 2$, let n_j be the original number of subjects in group j, and $n = n_1 + n_2$. Suppose that $\{t_1, t_2, ..., t_D\}$ is the set of D distinct ordered event times in the pooled sample. For $i = 1, ..., D$, define d_{ij} as the observed number of events out of Y_{ij} individuals at risk in the jth group at time t_i. Let $d_i = d_{i1} + d_{i2}$ and $Y_i = Y_{i1} + Y_{i2}$. Then, the test statistic of the conventional logrank test can be written as

$$U = \frac{\sum_{i=1}^{D} w_{i1} \left(d_{i1} - Y_{i1} \frac{d_i}{Y_i} \right)}{\sqrt{\sum_{i=1}^{D} w_{i1}^2 \frac{Y_{i1}}{Y_i} \frac{Y_{i2}}{Y_i} \frac{Y_i - d_i}{Y_i - 1} d_i}}, \tag{2.2}$$

where $w_{i1} = 1$, or any other finite positive constant, for all i.

As described in Section 1, the major idea of the second group of methods for handling the

crossing hazard rates problem is to choose special weights in the weighted logrank test, which change signs before and after a potential crossing point. However, if the weights are chosen as in the existing methods, the resulting test would not be asymptotically independent of the logrank test. To make the testing procedure for handling the crossing hazard rates problem asymptotically independent of the logrank test, we propose a new weighting scheme, which is described below.

First, we introduce some notation. For $j = 1, 2$ and $k = 1, 2, ..., n_j$, let T_{kj} be the event time of the kth subject in group j with c.d.f. F_j, C_{kj} be the censoring time with c.d.f. G_j, and

$$S_j(s) = 1 - F_j(s), \;\; L_j(s) = 1 - G_j(s), \;\; X_{kj} = \min(T_{kj}, C_{kj}),$$
$$\delta_{kj} = I_{\{T_{kj} < C_{kj}\}}, \;\; \pi_j(s) = P\left(X_{kj} > s\right) = S_j(s)L_j(s).$$

Note that, in the above expression for $\pi_j(s)$, we have made a conventional assumption that event times T_{kj} and censoring times C_{kj} are independent of each other. Also, under H_0 in (2.1), $F_1 = F_2 = F$ and $S_1 = S_2 = S$.

Let $0 < \epsilon < 0.5$ be a small number, and $D_\epsilon = [D \cdot \epsilon]$ be the integer part of $D \cdot \epsilon$. For any $r \in [\epsilon, 1 - \epsilon]$, define $m = [D \cdot r]$. Then, for a possible crossing at time $D_\epsilon \le m \le D - D_\epsilon$, we consider a weighted logrank test with new weights

$$w_{i2}^{(m)} = \begin{cases} -1, & \text{if } i = 1, 2, \ldots, m \\ c_m, & \text{otherwise,} \end{cases} \tag{2.3}$$

where c_m is a positive quantity. In Appendix (cf., expressions (A.8) and (A.9)), we show that, under H_0, the asymptotic covariance between the logrank test statistic U (cf., (2.2)) and the new weighted logrank test statistic is

$$-\int_0^{F^{-1}(r)} \frac{L_1(s)L_2(s)}{p_1L_1(s) + p_2L_2(s)} dF(s) + \int_{F^{-1}(r)}^{u} c_m \frac{L_1(s)L_2(s)}{p_1L_1(s) + p_2L_2(s)} dF(s), \tag{2.4}$$

where $u = \inf\{s : \min(\pi_1(s), \pi_2(s)) = 0\}$ and $p_j = \lim_{n\to\infty} n_j/n$, for $j = 1, 2$. Therefore, this asymptotic covariance would be zero if $c_m = \int_0^{F^{-1}(r)} \frac{L_1(s)L_2(s)}{p_1L_1(s)+p_2L_2(s)} dF(s) / \int_{F^{-1}(r)}^{u} \frac{L_1(s)L_2(s)}{p_1L_1(s)+p_2L_2(s)} dF(s)$. A reasonable estimator of c_m is

$$\widehat{c}_m = \frac{\sum_{i=1}^{m} \frac{\widehat{L}_1(t_i)\widehat{L}_2(t_i)}{(n_1/n)\widehat{L}_1(t_i)+(n_2/n)\widehat{L}_2(t_i)} \cdot \Delta\widehat{S}(t_i)}{\sum_{i=m+1}^{D} \frac{\widehat{L}_1(t_i)\widehat{L}_2(t_i)}{(n_1/n)\widehat{L}_1(t_i)+(n_2/n)\widehat{L}_2(t_i)} \cdot \Delta\widehat{S}(t_i)}, \tag{2.5}$$

where $\widehat{L}_j$ and $\widehat{S}$ are the Kaplan-Meier estimates of the survival functions of the censoring and event times, respectively. Note that formulas for computing the Kaplan-Meier estimates $\widehat{L}_j$ are the same

as those for computing the Kaplan-Meier estimates $\widehat{S}_j$, except that δ_{kj} should be replaced by $1-\delta_{kj}$ in the former case. Note also that $\widehat{S}$ is computed from the pooled sample.

Then, our proposed test statistic for handling the crossing hazard rates problem is defined by

$$V = \sup_{D_\epsilon \le m \le D-D_\epsilon} V_m, \tag{2.6}$$

where

$$V_m = \frac{\sum_{i=1}^{D} \widehat{w}_{i2}^{(m)} \left(d_{i1} - Y_{i1} \frac{d_i}{Y_i} \right)}{\sqrt{\sum_{i=1}^{D} \left(\widehat{w}_{i2}^{(m)} \right)^2 \frac{Y_{i1}}{Y_i} \frac{Y_{i2}}{Y_i} \frac{Y_i - d_i}{Y_i - 1} d_i}}, \tag{2.7}$$

and $\widehat{w}_{i2}^{(m)}$ is defined by (2.3) with c_m replaced by $\widehat{c}_m$.

Mantel and Stablein (1988) defined their weighted log-rank test statistic for comparing two crossing hazard rates by using $w_{i2}^{(m)} = -1$ if $i \le m$ and $w_{i2}^{(m)} = 1$ otherwise. Compared to this test statistic, the test statistic V defined by expressions (2.3) and (2.5)–(2.7) has two major differences. One is that, in Mantel and Stablein's statistic, weights change sign before and after a potential crossing point, but they all have the same magnitude; the proposed weights in (2.3) have different magnitudes and signs before and after a potential crossing point, and thus their values depend on the position of the potential crossing point. The other major difference is that, in Mantel and Stablein's procedure, the potential crossing point considered ranges from 0 to D, while it ranges from D_ϵ to $D-D_\epsilon$ in the proposed procedure. Therefore, Mantel and Stablein's procedure has certain ability to detect differences between two hazard rates that are different but not crossing, while the proposed procedure does not have this property. However, since the cases when two hazard rates are different but not crossing have already been considered in the first stage of the two-stage procedure by a conventional procedure, which is usually more powerful for that purpose, the two-stage procedure would not lose any power in testing hypotheses in (2.1), by using the proposed method (2.5)–(2.7) in its second stage. This will be further demonstrated by some numerical examples in Section 4. Because the test statistic of Anderson and Senthilselvan's (1982) procedure is the same as that of Mantel and Stablein's when the data have no censoring, the above comment is also valid when Mantel and Stablein's procedure is replaced by Anderson and Senthilselvan's procedure and when the observations have no censoring.

Theorem 2.1 *For $j = 1, 2$, suppose that the event time has cdf F_j with continuous pdf, the censoring time has cdf G_j, observations in the two groups (i.e., treatment and control groups) are independent of each other, and censoring times are independent of event times in each group.*

Then, under the assumption that $F_1 \equiv F_2$, the two statistics U and V defined in (2.2) and (2.6) are asymptotically independent of each other.

According to Theorem 2.1, using U and V in the two stages of the proposed two-stage procedure, the two individual tests are asymptotically independent of each other under H_0. The proof of this theorem is provided in the Appendix.

In the proposed test statistic V, there is a positive parameter ϵ involved. From expression (2.4), it can be seen that, if ϵ was chosen 0, then it is not guaranteed that we can find c_m, for all $D_\epsilon \leq m \leq D - D_\epsilon$, so that V_m and U are asymptotically independent of each other. For instance, when $m = 0$ (or, $r = 0$), the asymptotic covariance between V_0 and U would be positive under some regularity conditions. The major purpose of using ϵ in V is to exclude such cases from consideration. In applications, ϵ should be chosen such that the potential crossing point is included in $[D_\epsilon, D - D_\epsilon]$. To this end, a figure showing the life-table estimates of the two hazard rates should be helpful (cf., Figure 5.1 in Section 5). Theoretically speaking, $\epsilon > 0$ could be arbitrarily small. In applications, if ϵ is chosen too small, then variability of V could be quite large. Based on our numerical experience, results would be quite stable if ϵ is chosen such that $D_\epsilon \geq 5$.

3 Significance Level and p-Value

3.1 Significance level

Suppose that the significance level of the proposed two-stage procedure is fixed at α, and the significance levels of its two stages are α_1 and α_2, respectively. Then, by their definitions, we have

$$\alpha_1 + P_{H_0}(\text{reject in stage two} \mid \text{fail to reject in stage one})(1 - \alpha_1) = \alpha. \tag{3.1}$$

Because the tests in the two individual stages of the proposed two-stage procedure are asymptotically independent under H_0, (3.1) implies that it is asymptotically true that

$$\alpha_1 + \alpha_2(1 - \alpha_1) = \alpha. \tag{3.2}$$

It can be seen from equation (3.1) that, if the tests in the two individual stages are not asymptotically independent of each other, then it is generally difficult to control the asymptotic significance level of the two-stage procedure at a certain level, because the conditional probability term in (3.1) is often difficult to estimate.

Therefore, in order to guarantee that the whole two-stage procedure has an asymptotic significance level of α, α_1 and α_2 should satisfy equation (3.2), or, for a given $\alpha_1 \leq \alpha$, α_2 should be chosen $\alpha_2 = (\alpha - \alpha_1)/(1 - \alpha_1)$. In applications, if we have some prior information about the pattern of the two hazard rates, then α_1 and α_2 can be determined accordingly. In the two extreme cases that we believe the two hazard rates can not cross each other or they can not be different but not crossing, α_1 can be simply chosen α and 0, respectively. If we do not have such prior information, then we can simply let $\alpha_1 = \alpha_2$, and choose them to be

$$\alpha_1 = \alpha_2 = 1 - \sqrt{1 - \alpha}. \tag{3.3}$$

After α_1 and α_2 are determined, the critical value of the first individual test can be computed, using the asymptotic standard Normal distribution of U under H_0. Regarding the second individual test, Davies (1987) provided an upper bound for the tail probability of the null distribution of a similar statistic, based on some results about the partial likelihood ratio test for survival data. But, as pointed out by O'Quigley and Pessione (1991) and based on our own numerical experience, critical value based on this upper bound may not be accurate. In this paper, the critical value of the null distribution of V is estimated using bootstrap (cf., e.g., O'Quigley and Pessione 1991, Efron and Tibshirani 1993, Davison and Hinkley 1997, Liu *et al.* 2007).

3.2 p-value

After α_1 and α_2 are determined (e.g., by expressions (3.2) and (3.3)), the rejection region of the two-stage procedure is well defined. However, its p-value is still not defined yet. We notice that the proposed two-stage procedure can be treated as a *two-stage adaptive test* in the literature (c.f., e.g., Posch and Bauer 1999, Brannath *et al.* 2002). Generally speaking, the p-value of a two-stage adaptive test can be defined in several different ways. For the proposed two-stage procedure, since the tests in its two individual stages are asymptotically independent of each other under H_0, one convenient definition of the p-value is

$$p - \text{value} = \begin{cases} p_1, & \text{if } p_1 \leq \alpha_1 \\ \alpha_1 + p_2(1 - \alpha_1), & \text{otherwise.} \end{cases} \tag{3.4}$$

where p_1 and p_2 denote the p-values of the two stages, and α_1 is the significance level of the first stage.

From equation (3.4), it can be seen that the test using that p-value would reject the null hypothesis when (i) the test in the first stage rejects the null hypothesis (i.e., $p_1 \leq \alpha_1$), or (ii) the test in the first stage fails to reject the null hypothesis, but the test in the second stage rejects the null hypothesis (i.e., $p_1 > \alpha_1$ and $p_2 \leq \alpha_2$). The latter result is a direct conclusion of equations (3.2) and (3.4), because $p_2 \leq \alpha_2$ is equivalent to $\alpha_1 + p_2(1 - \alpha_1) \leq \alpha$ under the restriction of (3.2). Therefore, the test defined by the p-value in (3.4) is equivalent to the test defined by α_1 and α_2 discussed in Subsection 3.1.

4 A Simulation Study

In the simulation study, it is assumed that the hazard rate function of the control group is $h_0(t) = 1$, and the hazard rate function of the treatment group is $h_1(t) = a_0 + a_1 t$. The following four cases of $h_1(t)$ are considered: (i) $a_0 = 1$ and $a_1 = 0$, (ii) $a_0 = 2$ and $a_1 = 0$, (iii) $a_0 = 0.3$ and $a_1 = 1$, and (iv) $a_0 = 1.2$ and $a_1 = 0.6$. In case (i), $h_0(t) \equiv h_1(t)$ and thus H_0 holds. In case (ii), $h_0(t)$ and $h_1(t)$ are parallel to each other. They cross each other at $t = 0.7$ in case (iii); they are different, but neither parallel nor crossing, in case (iv). Therefore, the above four cases, demonstrated by Figure 4.1, represent four typical patterns of the two hazard rate functions.

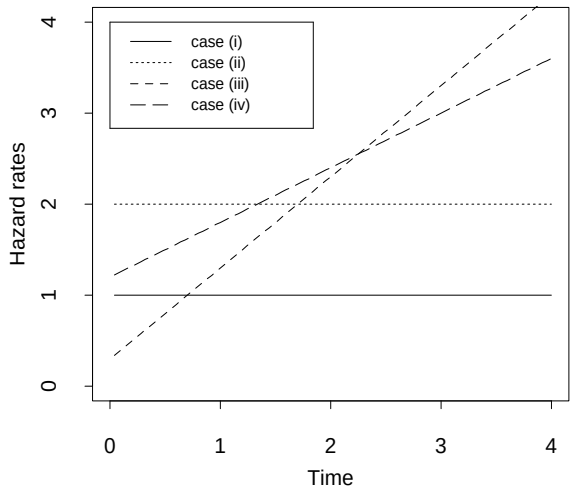

Figure 4.1: Four cases of $h_1(t)$ considered in the simulation study. In all cases, $h_0(t) = 1$.

In each case, either 50 or 100 survival times are generated for each of the treatment and control groups. For each sample, we consider two scenarios: there is no censoring in the data, or the censoring time has a Uniform distribution on $[0, 2]$. The following procedures are performed for testing hypotheses in (1.1): logrank (LR), Gehan-Wilcoxon (GW), Peto-Peto (PP), Anderson

and Senthilselvan's (1982) modeling approach using bootstrap (AS), proposed new procedure for handling the crossing hazard rates problem (NP, cf., expressions (2.3) and (2.5)–(2.7)), and the proposed two-stage procedure using LR and NP in its first and second stages respectively (LR+NP). The first three procedures are conventional, and they should be powerful in handling the cases when the two hazard rates are different but not crossing (i.e., cases (ii) and (iv)); the fourth and fifth procedures are designed for handling the crossing hazard rates problem (i.e., case (iii)); and the two-stage procedure are designed for handling all possible alternatives.

For each method, its power, defined as the proportion of rejections, is recorded based on 1000 replicated simulations in each scenario. The nominal significance level of each method is fixed at the conventional level 0.05. For procedure NP, D_ϵ is fixed at 5. For bootstrap procedures used in AS and NP, the bootstrap sample size is fixed at 1000. The results are presented in Table 4.1. For readers' convenience, the censoring rates in various cases when the censoring time has a Uniform distribution on $[0, 2]$ are presented in Table 4.2. It should be pointed out that, theoretically speaking, censoring rates in the treatment and control groups in case (i) should be the same. They are slightly different in Table 4.2 due to randomness, since we use two different sets of random numbers for the two groups in our computer programs.

Table 4.1: Powers of various methods for comparing two hazard rate functions in several different cases. In cases with censoring, the censoring time has a Uniform distribution on $[0, 2]$.

sample size		cases without censoring				cases with censoring			
in each group	Methods	(i)	(ii)	(iii)	(iv)	(i)	(ii)	(iii)	(iv)
50	LR	0.052	0.923	0.050	0.593	0.046	0.810	0.173	0.301
	GW	0.048	0.842	0.222	0.329	0.040	0.688	0.371	0.190
	PP	0.048	0.842	0.222	0.329	0.044	0.733	0.307	0.225
	AS	0.038	0.072	0.670	0.215	0.036	0.075	0.234	0.072
	NP	0.045	0.130	0.692	0.040	0.047	0.049	0.339	0.063
	LR+NP	0.049	0.861	0.552	0.507	0.047	0.725	0.320	0.240
100	LR	0.049	0.997	0.054	0.901	0.049	0.976	0.263	0.543
	GW	0.051	0.981	0.332	0.616	0.048	0.943	0.599	0.317
	PP	0.051	0.981	0.332	0.616	0.047	0.961	0.499	0.392
	AS	0.046	0.082	0.956	0.370	0.030	0.075	0.511	0.117
	NP	0.043	0.287	0.936	0.026	0.042	0.037	0.629	0.102
	LR+NP	0.047	0.996	0.887	0.843	0.046	0.958	0.608	0.471

Let us first focus on the results when there is no censoring in the observations (i.e., results in the left half of Table 4.1). Since H_0 holds in case (i), power values of various procedures in this

Table 4.2: Censoring rates in four cases when the censoring time has a Uniform distribution on $[0, 2]$.

Groups	case (i)	case (ii)	case (iii)	case (iv)
Control	21.7%	21.7%	21.7%	21.6%
Treatment	21.6%	12.3%	24.5%	16.0%

case are their actual significance levels. From Table 4.1, it can be seen that their actual significance levels are quite close to the nominal level 0.05 in all cases. The small differences are mainly due to random variation. In case (ii) when the two hazard rate functions are different but parallel to each other, as expected, the three conventional procedures LR, GW, and PP perform well, and the two procedures AS and NP which are designed for comparing two crossing hazard rate functions are not as powerful as the three conventional procedures. The proposed two-stage procedure has a little less power in such a case, compared to procedure LR, it is a little more powerful than procedures GW and PP, and it is much more powerful than procedures AS and NP. In case (iii) when the two hazard rate functions cross each other, it can be seen that the conventional procedures LR, GW, and PP have small power in detecting the difference between the two hazard rate functions, and procedures AS and NP have larger powers. As a comparison, the two-stage procedure performs relatively well in such a case. In case (iv), the two hazard rates are different but not crossing. It can be seen that the two-stage procedure performs relatively well too, compared to its peers. When there is censoring in the observed data, similar results can be observed from Table 4.1, although the power values are generally smaller than those in the corresponding cases without data censoring.

In Table 4.1, we notice that the three conventional procedures LR, GW, and PP perform better in case (iii) when the data have censoring, compared to their performance in case (iii) when the data have no censoring, which can be intuitively explained as follows. In case (iii), the two hazard rate functions cross at $t = 0.7$. In the control group, the event time follows the exponential distribution with mean 1. It has 50.341% chance to be located before the crossing point and 49.659% chance to be located after the crossing point. On the other hand, the censoring time considered in this simulation study has a Uniform distribution on $[0, 2]$; it has 35% chance to be located before the crossing point and 65% chance to be located after the crossing point. So, the event time located after the crossing point has a much larger chance to be censored, compared to the event time located before the crossing point. Similar observations can be made regarding censoring of the event times in the treatment group. As a consequence, when computing the test statistics of the

three conventional procedures, there would be less cancellation of positive and negative differences between the two estimated hazard rate functions in the case when the data have censoring, compared to the case when the data have no censoring (cf., equation (2.2)). Therefore, these methods would be a little more powerful in comparing the two hazard rates in the former case.

The two test statistics U and V used by the proposed two-stage procedure are proved to be asymptotically independent in Theorem 2.1 when H_0 holds. To investigate their finite-sample distributional properties, we consider the following example. In case (i), assume that $n_1 = n_2 = 100$. In each of the treatment and control groups, the censoring time follows the Uniform distribution on [0,2], as before. Then, 5000 datasets are generated independently as described above, from which 5000 values of (U, V) can be computed. The density histogram of the 5000 values of U is presented in Figure 4.2(a), and the density histogram of the 5000 values of V is presented in Figure 4.2(b). It can be seen that the distribution of U is reasonably Normal. The distribution of V is bimodal and symmetric about zero, which is consistent with the results found by O'Quigley and Pessione (1991) about a similar statistic. The joint density histogram constructed from the 5000 values of (U, V) is presented in Figure 4.2(c). As a comparison, the joint density histogram of (U, V) based on the assumption that U and V are independent is shown in Figure 4.2(d). When U and V are assumed independent, the joint probability $P(u_1 < U \leq u_2, v_1 < V \leq v_2)$ is set to be $P(u_1 < U \leq u_2) \cdot P(v_1 < V \leq v_2)$, for any $u_1 < u_2$ and $v_1 < v_2$. It can be seen from plots (c) and (d) that the two joint distributions of (U, V) are very similar, which implies that U and V are reasonably independent of each other in such a case. Similar results were obtained in the case when $n_1 = n_2 = 200$.

At the end of this section, we would like to point out that the proposed two-stage procedure can also be used for handling cases with two crossing points. In such cases, the test statistic V defined in equation (2.6) can still be used in the second stage of the procedure, after weights $\widehat{w}_{i2}^{(m)}$ in (2.7) are replaced by new weights $\widehat{w}_{i3}^{(m)}$ defined by

$$\begin{aligned} \widehat{w}_{i3}^{(m)} &= \begin{cases} -1, & \text{if } 1 \leq i \leq m_1 \text{ or } m_2 + 1 \leq i \leq D \\ \widehat{b}_{m_1,m_2}, & \text{otherwise,} \end{cases} \\ \widehat{b}_{m_1,m_2} &= \frac{\left(\sum_{i=1}^{m_1} + \sum_{i=m_2+1}^{D}\right) \frac{\widehat{L}_1(t_i)\widehat{L}_2(t_i)}{(n_1/n)\widehat{L}_1(t_i)+(n_2/n)\widehat{L}_2(t_i)} \cdot \Delta\widehat{S}(t_i)}{\sum_{i=m_1+1}^{m_2} \frac{\widehat{L}_1(t_i)\widehat{L}_2(t_i)}{(n_1/n)\widehat{L}_1(t_i)+(n_2/n)\widehat{L}_2(t_i)} \cdot \Delta\widehat{S}(t_i)}, \end{aligned} \tag{4.1}$$

where $D_\epsilon \leq m_1 < m_2 \leq D - D_\epsilon$ are two integers, and $D_\epsilon, n, n_1, n_2, \widehat{L}_1, \widehat{L}_2$ and $\widehat{S}$ are defined in

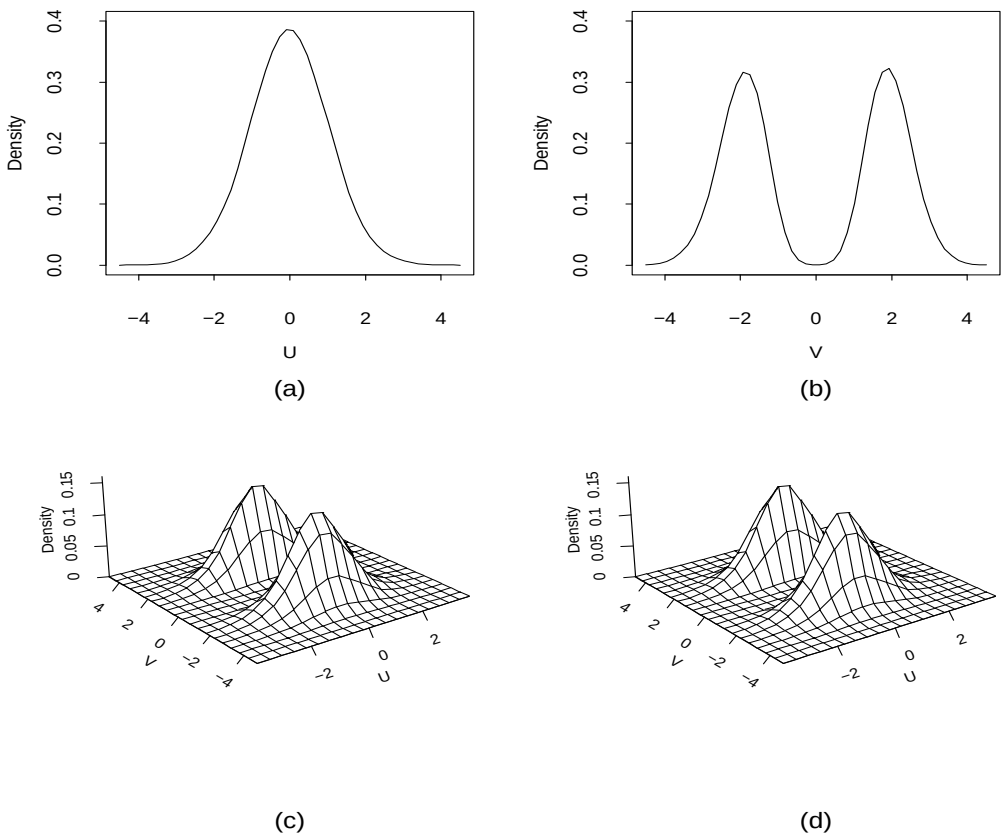

Figure 4.2: (a)-(b): Density histograms of U and V when $n_1 = n_2 = 100$. (c)-(d): Joint density histograms of (U, V) without and with the assumption that U and V are independent.

Section 2.2.

For demonstration, let us consider the two hazard rate functions shown in Figure 4.3, which cross each other at $t = 0.25$ and 0.925. For each of the treatment and control groups, 100 survival times are generated. For each sample, we consider two scenarios: there is no censoring in the data, or the censoring time has a Uniform distribution on $[0, 2]$, as in the previous example. Based on 1000 replications, computed power values of various methods considered in this section are shown in Table 4.3. From the table, it can be seen that the three conventional procedures LR, GW and PP have no power at all for comparing the two crossing hazard rate functions. Procedure AS is designed for handling cases with one crossing point only. So, its power is also low when the data have no censoring. When the data have censoring, it seems that its power is improved, which might be due to the reason that data censoring results in less cancellation of positive and negative differences between the two estimated hazard rate functions, as pointed out before regarding the performance of the three conventional procedures in case (iii) of Table 4.1. As a comparison, the proposed two-stage procedure has relatively large powers. Furthermore, the proposed procedure can provide point

estimates of the two crossing points, which can not be achieved by all other procedures considered here. When the data have censoring, the density histograms of 1000 replicated estimates of the two crossing points are shown in Figure 4.4. The corresponding density histograms when the data have no censoring are similar. From the figure, it can be seen that the point estimates perform reasonably well. Their sampling distributions are slightly skewed to the right, due mainly to the fact that distributions of the survival times in the control and treatment groups are both skewed to the right.

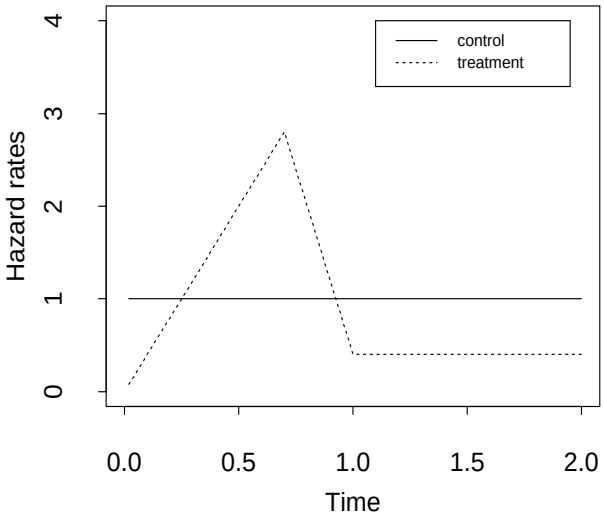

Figure 4.3: Solid and dotted lines denote $h_0(t)$ and $h_1(t)$, respectively. They cross at two time points $t = 0.25$ and 0.925.

Table 4.3: Powers of various methods for comparing two hazard rate functions shown in Figure 4.3. In the case with data censoring, the censoring time has a Uniform distribution on $[0, 2]$.

Methods	Case without censoring	Case with censoring
LR	0.063	0.069
GW	0.053	0.074
PP	0.053	0.047
AS	0.138	0.380
NP	0.659	0.540
LR+NP	0.558	0.409

5 Two Examples

We apply the related testing procedures considered in the previous section to two real datasets in this section. The first dataset is about kidney dialysis patients, which was taken from a study

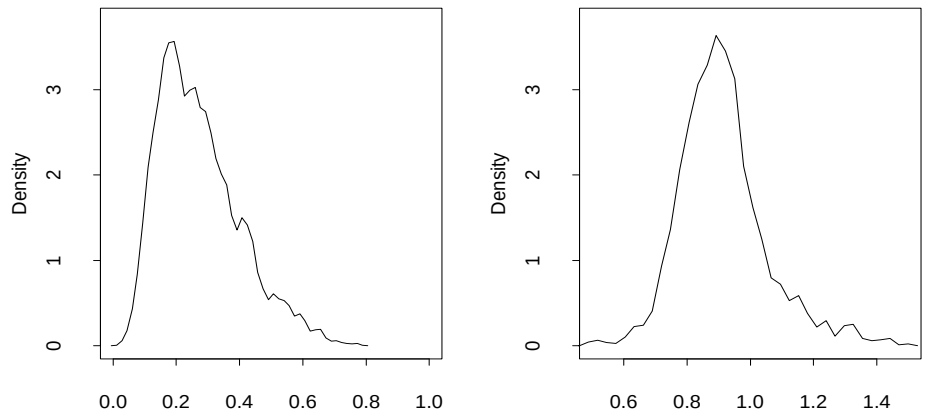

Figure 4.4: (a) Density histogram of 1000 replicated estimates of the first crossing point $t = 0.25$ when the data have censoring. (b) Density histogram of 1000 estimates of the second crossing point $t = 0.925$ when the data have censoring.

designed to assess the time to first exit-site infection (in months) in 119 patients with renal insufficiency. Among all patients, 43 of them utilized a surgically placed catheter (Group 1) and 76 of them utilized a percutaneous placement of their catheter (Group 2). Catheter failure was the primary reason for censoring. There were 27 censored observations in Group 1 and 65 censored observations in Group 2. This dataset was described in detail by Klein and Moeschberger (1997, Section 1.4). It was also analyzed by Lin and Wang (2004) recently, who found that the two hazard rates crossed at an early time, which can be seen from the two life-table estimators of the hazard rates shown in Figure 5.1(a) as well. Procedures LR, GW, PP, AS, NP, and LR+NP are then applied to this dataset. For each procedure, the conventional significance level 0.05 is used. For procedure NP, D_ϵ is fixed at 5, as before. For procedures AS and NP, the bootstrap sample size is fixed at 1000. For the two-stage procedure, we use $\alpha_1 = \alpha_2 = 0.0253$, as specified in equation (3.3). P-values of various procedures are listed in Table 5.1. It can be seen that the three conventional procedures LR, GW, and PP could not detect the crossing differences between the two hazard rates, and the remaining three procedures all detect such differences successfully.

Table 5.1: P-values of various procedures when they are applied to the kidney dialysis patients data (Kidney) and the rats data (Rat).

Datasets	LR	GW	PP	AS	NP	LR+NP
Kidney	0.112	0.964	0.237	0.005	<0.001	0.026
Rat	0.003	0.026	0.009	0.150	0.039	0.003

Next, we apply the related testing procedures to another dataset obtained from a study about

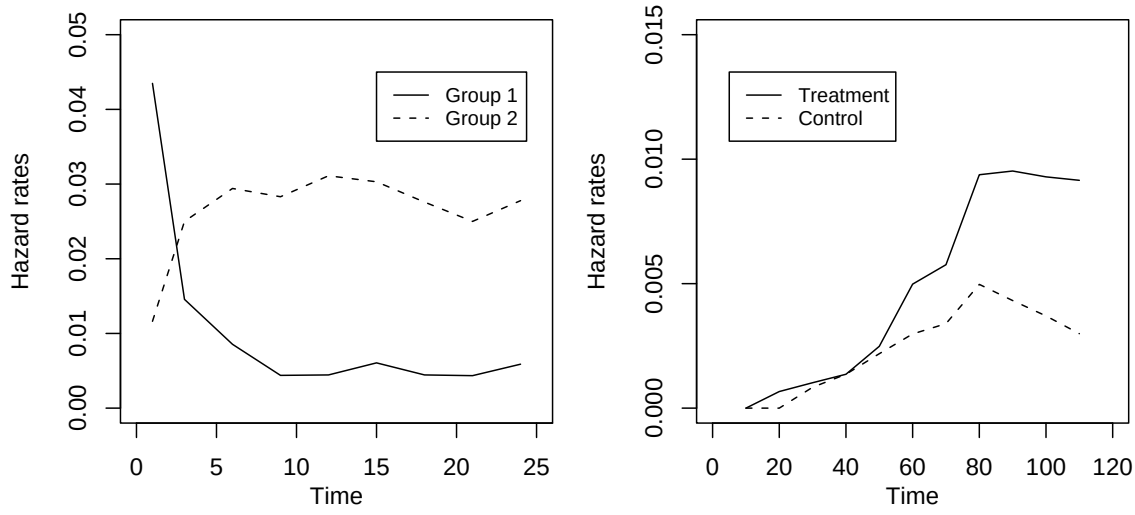

Figure 5.1: (a) Life-table estimators of the hazard rates for the kidney dialysis patients data. (b) Life-table estimators of the hazard rates for the rat data.

the tumorigenesis of a drug reported by Mantel *et al.* (1977). In the experiment, rats were taken from fifty distinct litters and one rat of the litter was randomly selected and given the drug. For each litter two rats were selected as controls and were given a placebo. All mice were females. In the treatment group, there are 29 censored observations of the times to tumor. The number of censored observations is 81 in the control group. The two life-table estimators of the hazard rates are shown in Figure 5.1(b), from which we can see that the two hazard rates touch briefly at an early time. The p-values of various procedures are reported in Table 5.1, from which we can see that the three conventional procedures LR, GW, and PP give significant results this time, procedure AS could not detect the differences between the two hazard rates, procedure NP gives a significant result but its p-value is relatively large, and the two-stage procedure detects the differences successfully once again.

6 Concluding Remarks

We have presented a two-stage procedure for comparing two hazard rates. A major advantage of this procedure is that it can handle all possible alternatives, including ones with crossing or different but not crossing hazard rates. To make specification of its significance level and computation of its p-value convenient, a new procedure for handling the crossing hazard rates problem is suggested, which has the property that its test statistic is asymptotically independent of the test statistic of the logrank test. Numerical examples show that the suggested two-stage procedure, with the

logrank test and the proposed procedure for handling the crossing hazard rates problem used in its two stages, works well in various cases.

Although our discussion in this paper focuses on the case when the logrank test and the suggested test for handling the crossing hazard rates problem are used in the two stages of the proposed two-stage procedure, at this moment it is unknown to us whether there exist some more powerful combinations of the conventional tests and tests for handling the crossing hazard rates problem in certain cases, which requires much future research.

This paper focuses on comparison of two hazard rate functions. In some applications, comparison of three or more hazard rate functions is also our concern (cf., Klein and Moeschberger 1997, Example 7.4). In such cases, there are many different crossing patterns, including crossings of two hazard rates only, crossings of three hazard rates only, and so forth; the relative positions of hazard rate functions can also change in many different ways before and after a given crossing point. For handling such cases, the log-rank test used in the first stage of the proposed two-stage procedure can be generalized in the way as described in Section 7.3 of Klein and Moeschberger (1997). However, it is not easy to generalize the proposed procedure (2.5)–(2.7) used in the second-stage of the two-stage procedure so that the generalized version is asymptotically independent of the generalized log-rank test. This problem also requires much future research.

In Section 4, we have investigated cases when two hazard rate functions cross once or twice. Theoretically speaking, as long as the number of crossing points is known, our proposed procedure can be adapted for handling cases with any number of crossing points, similarly to (4.1). To determine the number of possible crossing points, graphs displaying the life-table estimates of the two hazard rate functions are often helpful (cf., Figure 5.1). In applications, however, we may not have enough observed event times around each potential crossing point. Consequently, the proposed procedure may not have enough power in comparing the two hazard rate functions, especially when the number of potential crossing points is relatively large. In such cases, some alternative approaches might work better, verification of which is left for future research.

Acknowledgments: We thank the editor and two referees for providing many constructive comments and suggestions which greatly improved the quality of the paper.

Appendix: Proof of Theorem 2.1

Let $\mathbf{w} = (w_1, w_2, ..., w_D)^T$ denote a vector of weights; it represents either the weights in U or the weights in V (cf., expressions (2.2)–(2.7)). Then, we define

$$Z(\mathbf{w}) = h\sum_{i=1}^{D} w_i \left(d_{i1} - Y_{i1}\frac{d_i}{Y_i} \right), \;\; \widehat{\sigma}^2(\mathbf{w}) = h^2 \sum_{i=1}^{D} w_i^2 \frac{Y_{i1}}{Y_i}\frac{Y_{i2}}{Y_i}\frac{Y_i - d_i}{Y_i - 1} d_i,$$

where $h = \sqrt{n/(n_1 n_2)}$. We also define the following counting processes: for $j = 1, 2$,

$$\overline{Y}_j(s) = \sum_{k=1}^{n_j} I_{\{X_{kj} \geq s\}}, \;\; \overline{N}_j(s) = \sum_{k=1}^{n_j} I_{\{X_{kj} \leq s,\, \delta_{kj}=1\}}.$$

Note that, for group j, $\overline{Y}_j(s)$ is the at-risk process which is left continuous, and $\overline{N}_j(s)$ is the event process which is right continuous. Let $\widehat{S}(s)$ be the Kaplan-Meier estimator of the survival function $S(s)$, and $\mathcal{W}(s)$ be a bounded predictable function of $\widehat{S}(s-)$ satisfying $(\mathcal{W}(t_1), ..., \mathcal{W}(t_D))^T = \mathbf{w}$. Then, $Z(\mathbf{w})$ could be written as

$$Z(\mathbf{w}) = h\int_0^u \mathcal{W}(s)\frac{\overline{Y}_1(s)\overline{Y}_2(s)}{\overline{Y}_1(s) + \overline{Y}_2(s)} \left\{ \frac{d\overline{N}_1(s)}{\overline{Y}_1(s)} - \frac{d\overline{N}_2(s)}{\overline{Y}_2(s)} \right\}, \tag{A.1}$$

where $u = \inf\{s : \min(\pi_1(s), \pi_2(s)) = 0\}$. Let

$$K_{\mathcal{W}}(s) \;\; = \;\; h\mathcal{W}(s)\frac{\overline{Y}_1(s)\overline{Y}_2(s)}{\overline{Y}_1(s) + \overline{Y}_2(s)}.$$

By the facts that $\overline{Y}_1(s)/n_1$, $\overline{Y}_2(s)/n_2$, and $(\overline{Y}_1(s) + \overline{Y}_2(s))/n$ are consistent estimators of $\pi_1(s)$, $\pi_2(s)$, and $p_1\pi_1(s) + p_2\pi_2(s)$, respectively, we have

$$\frac{K_{\mathcal{W}}^2(s)}{\overline{Y}_j(s)} \xrightarrow{Pr} \mathcal{W}^2(s)\frac{p_1 p_2 \pi_1^2(s)\pi_2^2(s)}{p_j \pi_j(s)(p_1\pi_1(s) + p_2\pi_2(s))^2}, \text{ as } n \to \infty, \text{ for } j = 1, 2,$$

where $\xrightarrow{Pr}$ denotes convergence in probability, and $p_j = \lim_{n\to\infty} n_j/n$. It can be checked that the regularity conditions (1)–(3) of Corollary 7.2.1 in Fleming and Harrington (1991) are all satisfied in our case, therefore, by that result, we have

$$Z(\mathbf{w})/\sigma(\mathbf{w}) \xrightarrow{D} \mathrm{N}(0, 1), \tag{A.2}$$

where $\xrightarrow{D}$ denotes convergence in distribution, and

$$\begin{aligned}
\sigma^2(\mathbf{w}) \;\; &= \;\; \int_0^u \mathcal{W}^2(s)\frac{\pi_1(s)\pi_2(s)}{p_1\pi_1(s) + p_2\pi_2(s)} \cdot (1 - \Delta\Lambda(s))\, d\Lambda(s) \\
&= \;\; \int_0^u \mathcal{W}^2(s)\frac{\pi_1(s)\pi_2(s)}{p_1\pi_1(s) + p_2\pi_2(s)} \cdot \frac{1}{S(s)}\, dF(s) \\
&= \;\; \int_0^u \mathcal{W}^2(s)\frac{L_1(s)L_2(s)}{p_1 L_1(s) + p_2 L_2(s)}\, dF(s),
\end{aligned}$$

and $\Lambda(s)$ is the common cumulative hazard function of the event time under H_0 which is continuous because the common c.d.f. is assumed to have a continuous density function.

For any $r \in [\epsilon, 1-\epsilon]$, let $m = [D \cdot r]$ and $\mathcal{W}_2^r(s) = -I_{\{s<t_m\}} + \widehat{c}_m \cdot I_{\{s \geq t_m\}}$ where $\widehat{c}_m$ is defined in equation (2.5). Then, $\mathcal{W}_2^r(s)$ is a predictable function of $\widehat{S}(s-)$. Since $\widehat{c}_m < \infty$ for any $\epsilon \leq r \leq 1-\epsilon$, we have

$$\sigma^2\left(\mathbf{w}_2^r\right) = \int_0^u \left(\mathcal{W}_2^r\right)^2(s) \frac{L_1(s)L_2(s)}{p_1 L_1(s) + p_2 L_2(s)}\, dF(s) < \infty.$$

Therefore, there exists a constant $w_0 \in (0, \infty)$ such that, if we let $\mathcal{W}_1(s) \equiv w_0$, then we have

$$\sigma^2\left(\mathbf{w}_1\right) = \sigma^2\left(\mathbf{w}_2^r\right), \tag{A.3}$$

where $\mathbf{w}_1 = (\mathcal{W}_1(t_1), ..., \mathcal{W}_1(t_D))^T = w_0 \cdot \mathbf{1}_D$ and $\mathbf{w}_2^r = (\mathcal{W}_2^r(t_1), ..., \mathcal{W}_2^r(t_D))^T$.

Now, define

$$U^* = \frac{Z\left(\mathbf{w}_1\right)}{\sigma\left(\mathbf{w}_1\right)}, \;\; V_r^* = \frac{Z\left(\mathbf{w}_2^r\right)}{\sigma\left(\mathbf{w}_2^r\right)}.$$

We will show that

$$\begin{pmatrix} U^* \\ V_r^* \end{pmatrix} \xrightarrow{D} N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \mathbf{I}_2\right), \text{ as } n \to \infty. \tag{A.4}$$

To prove equation (A.4), let us consider the linear combination

$$aU^* + bV_r^* = \frac{Z\left(a\mathbf{w}_1 + b\mathbf{w}_2^r\right)}{\sigma\left(\mathbf{w}_1\right)}, \tag{A.5}$$

where a and b are two arbitrary constants and where equation (A.3) has been used. Then, by equations (A.1) and (A.2), we have

$$Z\left(a\mathbf{w}_1 + b\mathbf{w}_2^r\right) / \sigma\left(a\mathbf{w}_1 + b\mathbf{w}_2^r\right) \xrightarrow{D} \mathrm{N}(0,1), \text{ as } n \to \infty. \tag{A.6}$$

Therefore, if we can prove that

$$\frac{\sigma^2\left(a\mathbf{w}_1 + b\mathbf{w}_2^r\right)}{\sigma^2\left(\mathbf{w}_1\right)} \xrightarrow{Pr} a^2 + b^2, \text{ as } n \to \infty, \tag{A.7}$$

then after combining (A.5)–(A.7), we have

$$aU^* + bV_r^* \xrightarrow{D} \mathrm{N}(0, a^2 + b^2), \text{ as } n \to \infty.$$

Then, equation (A.4) follows.

To prove (A.7), we notice that

$$\begin{aligned}
& \sigma^2\left(a\mathbf{w}_1+b\mathbf{w}_2^r\right) \\
= & \int_0^u \left[a\mathcal{W}_1+b\mathcal{W}_2^r\right]^2(s)\frac{L_1(s)L_2(s)}{p_1L_1(s)+p_2L_2(s)}dF(s) \\
= & \int_0^u \left[a^2\mathcal{W}_1^2(s)+b^2\left(\mathcal{W}_2^r\right)^2(s)+2ab\mathcal{W}_1(s)\mathcal{W}_2^r(s)\right]\frac{L_1(s)L_2(s)}{p_1L_1(s)+p_2L_2(s)}dF(s) \\
= & a^2\sigma^2\left(\mathbf{w}_1\right)+b^2\sigma^2\left(\mathbf{w}_2^r\right)+2abw_0\int_0^u \mathcal{W}_2^r(s)\frac{L_1(s)L_2(s)}{p_1L_1(s)+p_2L_2(s)}dF(s) \\
= & \left(a^2+b^2\right)\sigma^2\left(\mathbf{w}_1\right)+2abw_0\int_0^u \mathcal{W}_2^r(s)\frac{L_1(s)L_2(s)}{p_1L_1(s)+p_2L_2(s)}dF(s). \qquad \text{(A.8)}
\end{aligned}$$

Now,

$$\begin{aligned}
& \int_0^u \mathcal{W}_2^r(s)\frac{L_1(s)L_2(s)}{p_1L_1(s)+p_2L_2(s)}dF(s) \\
= & -\int_0^{t_m}\frac{L_1(s)L_2(s)}{p_1L_1(s)+p_2L_2(s)}dF(s)+\int_{t_m}^u \widehat{c}_m\frac{L_1(s)L_2(s)}{p_1L_1(s)+p_2L_2(s)}dF(s).
\end{aligned}$$

If we can show that

$$\widehat{c}_m \xrightarrow{Pr} k_r, \text{ as n } \to\infty, \qquad \text{(A.9)}$$

where

$$k_r=\frac{\int_0^{F^{-1}(r)}\frac{L_1(s)L_2(s)}{p_1L_1(s)+p_2L_2(s)}dF(s)}{\int_{F^{-1}(r)}^u\frac{L_1(s)L_2(s)}{p_1L_1(s)+p_2L_2(s)}dF(s)},$$

then we can conclude that

$$\int_0^u \mathcal{W}_2^r(s)\frac{L_1(s)L_2(s)}{p_1L_1(s)+p_2L_2(s)}dF(s)\xrightarrow{Pr}0, \text{ as } n\to\infty,$$

since t_m converges to $F^{-1}(r)$ in probability. Therefore, result (A.7) is true.

To show (A.9), we notice that

$$\begin{aligned}
\sum_{i=1}^m \frac{\widehat{L}_1(s)\widehat{L}_2(s)}{(n_1/n)\widehat{L}_1(s)+(n_2/n)\widehat{L}_2(s)}\cdot\Delta\widehat{S}(s) &= \int_0^{t_m}\frac{\widehat{L}_1(s)\widehat{L}_2(s)}{(n_1/n)\widehat{L}_1(s)+(n_2/n)\widehat{L}_2(s)}\,d\widehat{S}(s) \\
&= -\int_0^{t_m}\frac{\widehat{L}_1(s)\widehat{L}_2(s)}{(n_1/n)\widehat{L}_1(s)+(n_2/n)\widehat{L}_2(s)}\,d\widehat{F}(s),
\end{aligned}$$

where $\widehat{F}(s)=1-\widehat{S}(s)$. Then,

$$\begin{aligned}
& \left|\int_0^{t_m}\frac{\widehat{L}_1(s)\widehat{L}_2(s)}{(n_1/n)\widehat{L}_1(s)+(n_2/n)\widehat{L}_2(s)}\,d\widehat{F}(s)-\int_0^{F^{-1}(r)}\frac{L_1(s)L_2(s)}{p_1L_1(s)+p_2L_2(s)}\,dF(s)\right| \\
\le & \left|\int_0^{t_m}\frac{\widehat{L}_1(s)\widehat{L}_2(s)}{(n_1/n)\widehat{L}_1(s)+(n_2/n)\widehat{L}_2(s)}\,d\widehat{F}(s)-\int_0^{F^{-1}(r)}\frac{\widehat{L}_1(s)\widehat{L}_2(s)}{(n_1/n)\widehat{L}_1(s)+(n_2/n)\widehat{L}_2(s)}\,d\widehat{F}(s)\right|+
\end{aligned}$$

$$
\begin{aligned}
& \left| \int_0^{F^{-1}(r)} \frac{\widehat{L}_1(s)\widehat{L}_2(s)}{(n_1/n)\widehat{L}_1(s)+(n_2/n)\widehat{L}_2(s)}\, d\widehat{F}(s) - \int_0^{F^{-1}(r)} \frac{L_1(s)L_2(s)}{p_1L_1(s)+p_2L_2(s)}\, d\widehat{F}(s) \right| + \\
& \left| \int_0^{F^{-1}(r)} \frac{L_1(s)L_2(s)}{p_1L_1(s)+p_2L_2(s)}\, d\widehat{F}(s) - \int_0^{F^{-1}(r)} \frac{L_1(s)L_2(s)}{p_1L_1(s)+p_2L_2(s)}\, dF(s) \right| \\
\leq \quad & \left| r - \widehat{F}\left(F^{-1}(r)\right) \right| + \int_0^{F^{-1}(r)} \left[\frac{1}{p_2} \sup_s \left| \widehat{L}_1(s) - L_1(s) \right| + \frac{1}{p_1} \sup_s \left| \widehat{L}_2(s) - L_2(s) \right| \right] d\widehat{F}(s) \\
& + \left| \int_0^{F^{-1}(r)} \frac{L_1(s)L_2(s)}{p_1L_1(s)+p_2L_2(s)}\, d\left(\widehat{F}(s) - F(s)\right) \right| \\
\xrightarrow{Pr} \quad & 0, \text{ as } n \to \infty.
\end{aligned}
$$

In the previous result, we have used the weak consistency of the Kaplan-Meier estimators. Similar results can be obtained about the denominator of $\widehat{c}_m$. So, (A.9) is proved. Consequently, both (A.7) and (A.4) are proved.

Now, the statistics U and V_m defined in Section 2 can be written as

$$U = U^* \frac{\sigma(\mathbf{w}_1)}{\widehat{\sigma}(\mathbf{w}_1)}, \quad V_m = V_r^* \frac{\sigma(\mathbf{w}_2^r)}{\widehat{\sigma}(\mathbf{w}_2^r)}.$$

By Corollary 7.2.1 in Fleming and Harrington (1991), under the conditions stated in the theorem, we have

$$\frac{\sigma^2(\mathbf{w})}{\widehat{\sigma}^2(\mathbf{w})} \xrightarrow{Pr} 1, \text{ as } n \to \infty$$

in both cases when $\mathbf{w} = \mathbf{w}_1$ and $\mathbf{w} = \mathbf{w}_2^r$, respectively. So, by Theorems 18.10 and 18.11 in van der Vaart (1998), it follows that (U, V_m) converges in distribution to a bivariate normal with zero mean and identity covariance matrix. Therefore, U and V_m are asymptotically independent of each other. So do U and V. The proof is now finished.

References

Anderson, J.A., and Senthilselvan, A. (1982), "A two-step regression model for hazard functions," *Applied Statistics*, **31,** 44–51.

Bain, L.J., and Engelhardt, M. (1991), *Statistical Analysis of Reliability and Life-testing Models: Theory and Methods (2nd ed.),* New York: Marcel Dekker, Inc.

Brannath, W., Posch, M., and Bauer, P. (2002), "Recursive combination tests," *Journal of the American Statistical Association*, **97**, 236–244.

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

Cheng, M.Y., Hall, P., and Tu, D. (2006), "Confidence bands for hazard rate under random censorship," *Biometrika*, **93**, 357–366.

Davies, R.B. (1987), "Hypothesis testing when a nuisance parameter is present only under the alternative," *Biometrika*, **74**, 33–43.

Davison, A.C., and Hinkley, D.V. (1997), *Bootstrap Methods and Their Application,* Cambridge University Press: Cambridge, UK.

Efron, B., and Tibshirani, R.J. (1993), *An Introduction to the Bootstrap,* Chapman & Hall/CRC: New York.

Fleming, T.R., and Harrington, D.P. (1991), *Counting Processes and Survival Analysis*, John Wiley & Sons: New York.

Fleming, T.R., O'Fallon, J.R., O'Brien, P.C., and Harrington, D.P. (1980), "Modified Kolmogorov-Smirnov test procedures with application to arbitrarily right-censored data," *Biometrics*, **36,**, 607–625.

Klein, J.P., and Moeschberger, M.L. (1997), *Survival Analysis,* New York: Springer-Verlag.

Lawless, J.F. (1982), *Statistical Models and Methods for Lifetime Data,* New York: John Wiley & Sons.

Lin, X., and Wang, H. (2004), "A new testing approach for comparing the overall homogeneity of survival curves," *Biometrical Journal*, **46**, 489–496.

Liu,K, Qiu, P., and Sheng, J. (2007), "Comparing two crossing hazard rates by Cox proportional hazards modeling," *Statistics in Medicine*, **26**, 375–391.

Mantel, N., Bohidar, N.R., and Ciminera, J.L. (1977), "Mantel-Haenszel analysis of litter-matched time-to-response data, with modifications for recovery of interlitter information" *Cancer Research*, **37**, 3863–3868.

Mantel, N., and Stablein, D.M. (1988), "The crossing hazard function problem," *The Statistician*, **37,** 59–64.

Moreau, T., Maccario, J., Lellouch, J., and Huber, C. (1992), "Weighted log rank statistics for comparing two distributions," *Biometrika*, **79,** 195–198.

O'Quigley, J. (1994), "On a two-sided test for crossing hazard rates," *The Statistician*, **43,** 563–569.

O'Quigley, J., and Pessione, F. (1989), "Score test for homogeneity of regression effect in the proportional hazards model," *Biometrics*, **45,** 135–144.

O'Quigley, J., and Pessione, F. (1991), "The problem of a covariate-time qualitative interaction in a survival study," *Biometrics*, **47,** 101–115.

Posch, M., and Bauer, P. (1999), "Adaptive two-stage designs and the conditional error function," *Biometrical Journal*, **41**, 689–696.

van der Vaart, A.W. (1998), *Asymptotic Statistics,* Cambridge, UK: Cambridge University Press.