Comparison of Multiple Hazard Rate Functions

Zhongxue Chen[1,*], Hanwen Huang[2], and Peihua Qiu[3]

[1]Department of Epidemiology and Biostatistics, School of Public Health, Indiana University Bloomington. 1025 E. 7th street, Bloomington, IN, 47405, USA. Email: zc3@indiana.edu.

[2]Department of Epidemiology and Biostatistics, College of Public Health, University of Georgia. Athens, GA 30602, USA. Email: huanghw@uga.edu.

[3]Department of Biostatistics, College of Public Health & Health Professions and College of Medicine, University of Florida. Gainesville, FL 32611, USA. Email: pqiu@ufl.edu.

*Corresponding author, Phone: 1-812-855-1163

Running title: Comparing hazard rate functions

SUMMARY

Many robust tests have been proposed in the literature to compare two hazard rate functions, however, very few of them can be used in cases when there are multiple hazard rate functions to be compared. In this paper, we propose an approach for detecting the difference among multiple hazard rate functions. Through a simulation study and a real-data application, we show that the new method is robust and powerful in many situations, compared with some commonly used tests.

Key Words: Asymptotically independent; counting process; crossing; survival data.

1. Introduction

In survival data analysis, people are often interested in comparing hazard rate functions or survival curves of different treatment groups. In the literature, a large number of tests have been proposed to compare two hazard rate functions. See, for example, the so-called $G^{\rho,\gamma}$ test statistics introduced by Fleming and Harrington (Fleming and Harrington, 1991) that include the commonly used log-rank test due to Mantel and Haenszel (Mantel and Haenszel, 1959), and the Gehan-Wilcoxon test due to Gehan (Gehan, 1965) as special cases. The $G^{\rho,\gamma}$ tests are weighted log-rank tests that are based on the weighted sum of the differences between the expected and observed numbers of events in one group at each failure time point. In the $G^{\rho,\gamma}$ tests, the weights are usually assigned using non-negative values. In different situations, it has been shown that improvement of the $G^{\rho,\gamma}$ tests can be achieved by choosing appropriate weights (Buyske, Fagerstrom and Ying, 2000; Yang and Prentice, 2010).

In practice, the two hazard rate functions may cross each other. If this is the case, then the $G^{\rho,\gamma}$ tests would not be optimal and they may even have little or no power at all. To circumvent this limitation, negative weights have been considered and assigned to some failure time points (Moreau et al., 1992; Qiu and Sheng, 2008). In Qiu and Sheng (2008), the authors proposed a two-stage approach to compare two hazard rate functions. In the first stage, they used the log-rank test, which has been shown to be the most powerful test when the hazard rates are proportional (Fleming and Harrington, 1991); in the second stage, a weighted log-rank test with possible negative weights was designed to detect the difference between two hazard rate functions in cases when they crossed each other. An overall p-value was then calculated based on

the two test statistics. This two-stage test was shown to be robust and it has a good power to detect the difference between two hazard rate functions regardless whether they cross each other or not. In the literature, some other methods designed specifically for detecting the difference between two crossing hazard rate functions have been proposed either based on other functions (e.g., absolute or squared) of the differences between the two hazard rates or some modeling techniques (Lin and Wang, 2004; Liu, Qiu and Sheng, 2007; Park and Qiu, 2014).

When there are multiple treatment groups, although the log-rank test and the Gehan-Wilcoxon test can be extended and applied, they may have little or no power if the hazard rate functions cross each other. Robust methods with good properties as those of the two-stage approach are highly desirable. However, there is no simple extension of the aforementioned two-stage approach for comparing multiple hazard rate functions. In this paper, we propose a new approach based on a series of asymptotically independent tests. An overall p-value for each test is calculated using a robust p-value combining method.

The rest parts of the paper are organized as follows. In Section 2, we describe the proposed method; in Section 3, through a simulation study, we study the numerical performance of the new test; in Section 4, we illustrate the new test using a real-data application; the paper is concluded with some concluding remarks.

2. Proposed Method

Suppose we have K treatment groups; for group k $(k = 1,2, \dots, K)$, the survival function of the surviving time, T_{ik} $(i = 1,2, \dots, n_k)$, is $S_k(t) = 1 - F_k(t)$, and the survival function of the censoring time, C_{ik}, is $L_k(t) = 1 - G_k(t)$, where $F_k(t)$ and $G_k(t)$ are the related cumulative distribution functions. Let $\{t_1, t_2, \cdots, t_D\}$ be the set of D distinct ordered event times in the pooled sample. Define d_{ik} as the observed number of events out of Y_{ik} individuals at risk in the

k^{th} group at time t_i $(i = 1,2,\cdots,D)$. Denote $d_i = (d_{i1}, d_{i2}, \cdots, d_{iK})^T$, and $Y_i = (Y_{i1}, Y_{i2}, \cdots, Y_{iK})^T$. Let $X_{ik} = \min(T_{ik}, C_{ik})$, $\delta_{ik} = \mathrm{I}_{\{T_{ik} < C_{ik}\}}$, and $\pi_k(t) = \mathrm{P}(X_{ik} > \mathrm{t}) = S_k(t)L_k(t)$ $(k = 1,2,\ldots,K)$, where $I_{(\cdot)}$ is the indicator function. In this paper, we make the conventional assumption that the surviving times T_{ik} and the censoring times C_{ik} are independent.

To test for the homogeneity of the hazard rate functions among the K groups, we consider the following hypotheses:

H_0: $S_1(t) = S_2(t) = \cdots = S_K(t)$, and

H_1: at least one S_k is different than others at some time t.

In the two-sample situation (i.e., $K = 2$), the un-weighted log-rank test is constructed as follows:

$$U = h\sum_{i=1}^{D} w_{1i}\left(d_{i1} - Y_{i1}\frac{d_{i1}+d_{i2}}{Y_{i1}+Y_{i2}}\right)\Big/\sqrt{\sum_{i=1}^{D} w_{1i}^2 \frac{Y_{i1}}{Y_{i1}+Y_{i2}}\frac{Y_{i2}}{Y_{i1}+Y_{i2}}\frac{Y_{i1}+Y_{i2}-(d_{i1}+d_{i2})}{(Y_{i1}+Y_{i2})-1}(d_{i1}+d_{i2})}, \quad (1)$$

where $h = (\frac{n_1+n_2}{n_1 n_2})^{1/2}$, and $w_{1i} = 1$ for all $i = 1,2,\ldots,D$. Under the null hypothesis, U in (1) has an asymptotic standard normal distribution.

The weighted log-rank test statistic considered by (Qiu and Sheng, 2008) is defined as follows.

$$V_m = h\sum_{i=1}^{D} w_{2i}^{(m)}\left(d_{i1} - Y_{i1}\frac{d_{i1}+d_{i2}}{Y_{i1}+Y_{i2}}\right)\Big/\sqrt{\sum_{i=1}^{D}\left(w_{2i}^{(m)}\right)^2 \frac{Y_{i1}}{Y_{i1}+Y_{i2}}\frac{Y_{i2}}{Y_{i1}+Y_{i2}}\frac{Y_{i1}+Y_{i2}-(d_{i1}+d_{i2})}{(Y_{i1}+Y_{i2})-1}(d_{i1}+d_{i2})}, \quad (2)$$

where $w_{2i}^{(m)} = \begin{cases} -1 & if\ i = 1,2,\cdots,m \\ \hat{c}_m & otherwise \end{cases}$, $m = [Dr]$, the integer part of Dr for any $r \in [\varepsilon, 1-\varepsilon]$, $0 < \varepsilon < 0.5$, and $\hat{c}_m = \sum_{i=1}^{m} \frac{\hat{L}_1(t_i)\hat{L}_2(t_i)}{\left(\frac{n_1}{n}\right)\hat{L}_1(t_i)+\left(\frac{n_2}{n}\right)\hat{L}_2(t_i)}\Delta\hat{S}(t_i) \Big/ \sum_{i=m+1}^{D} \frac{\hat{L}_1(t_i)\hat{L}_2(t_i)}{\left(\frac{n_1}{n}\right)\hat{L}_1(t_i)+\left(\frac{n_2}{n}\right)\hat{L}_2(t_i)}\Delta\hat{S}(t_i)$, which is an estimate of the quantity

$$c_r = \int_0^{F^{-1}(r)} \frac{L_1(s)L_2(s)}{p_1L_1(s)+p_2L_2(s)} dF(s) / \int_{F^{-1}(r)}^{u} \frac{L_1(s)L_2(s)}{p_1L_1(s)+p_2L_2(s)} dF(s), \quad (3)$$

where $\hat{L}_l(t)$ is the estimate of $L_l(t)$ $(l = 1,2)$, $\hat{S}(t)$ is the estimate of the survival distribution $S(t)$ using the pooled samples, $p_1 = \lim_{n\to\infty} \frac{n_1}{n_1+n_2}$, $p_{k2} = \lim_{n\to\infty} \frac{n_2}{n_1+n_2}$; $m = [Dr]$, and $u = \inf [s: \min\{\pi_1(s), \pi_2(s)\} = 0]$.

Under the null hypothesis, each V_m defined in (2) has an asymptotic standard normal distribution (Qiu and Sheng, 2008). Furthermore, let

$$V = \sup_{D_\varepsilon \le m \le D - D_\varepsilon} (V_m),$$

(3)

which is designed specifically for detecting the difference between two crossing hazard rate functions, we have the following result (Qiu and Sheng, 2008):

Lemma 1. Under the null hypothesis, U and V_m for each m, and therefore, U and V, are asymptotically independent.

The sampling distribution of V in (3) under the null hypothesis is hard to derive explicitly since $V_m's$ are correlated. However, its p-value can be estimated using bootstrap (Qiu and Sheng, 2008).

Next, we want to generalize the statistics U and V to cases with multiple hazard rate functions (i.e., $K > 2$). To this end, we will obtain K-1 pairs of U and V through comparing two hazard rate functions K-1 times. Specifically, for each k (k=1, 2 ,…, K-1), we compare two groups, one from the pooled original treatment groups of 1, 2, …, k, and the other is the original group of k+1, and obtain two respective statistics U_k and V_k. For those U_k and V_k, we have the following result.

Theorem 1. Under the null hypothesis, the 2(K-1) statistics, U_k, V_k (k=1, 2, …, K-1) are asymptotically independent .

The proof of Theorem 1 is given in the Appendix. It should be pointed out that when K=2, the two statistics U_1 and V_1 are the same statistics as in (1) and (3) obtained by the Qiu and Sheng method. Based on the properties of those statistics U_k and V_k, we define some test statistics.

2.1 *The U test*

Define the test statistic

$$U_O = \sum_{i=1}^{K-1} (\chi_1^2)^{-1}\left(1 - P_{U_i}\right), \tag{4}$$

where P_{U_i} is the p-value from the test U_i $(i = 1,2,\cdots,K-1)$, $(\chi_1^2)^{-1}(.)$ is the inverse function of the cumulative chi-square distribution with degree of freedom (df) 1.

We will call the test U_O in (4) the U test. Under the null hypothesis P_{U_i} are asymptotically and independently distributed uniformly between 0 and 1. Therefore, under the null hypothesis, U_O has an asymptotic chi-square distribution with $K-1$ df. Its p-value can be calculated as $P_U = \text{pr}\,[\chi_{K-1}^2 > U_O]$. The U test is an extension of the two-sample log-rank test and it has similar performance as the multiple-sample log-rank test. When the hazard rate functions are parallel, this test should be powerful.

2.2 *The V test*

Define the test statistic

$$V_O = \sum_{i=1}^{K-1} (\chi_1^2)^{-1}\left(1 - P_{V_i}\right), \tag{13}$$

where P_{V_i} is the p-value from the test V_i $(i = 1,2,\cdots,K-1)$.

We will call the test V_O in (13) the V test. From theorem 1, under the null hypothesis V_i are asymptotically independent. Therefore, under the null hypothesis, V_O has an asymptotic chi-square distribution with df K-1. Its p-value can be calculated as $P_V = \text{pr}\,[\chi_{K-1}^2 > V_O]$. The V

test is an extension of the weighted log-rank test which is powerful when the hazard rate functions are crossing.

2.3 *The UV test*

From Theorem 1, under the null hypothesis, $U_k, V_{k'}$ $(k, k' = 1,2, \cdots, K-1)$ are asymptotically independent. Therefore, we have the following result.

Theorem 2. Under the null hypothesis, the U test and the V test statistics are asymptotically independent.

The proof of Theorem 2 are given in the Appendix.

Based on this result, a test statistic can be constructed as $U_O + V_O$. We will call this test the UV test. Under the null hypothesis, the UV test has an asymptotic chi-square distribution with df equals 2(K-1). Therefore, its p-value can be calculated as

$$P_{UV} = \text{pr}\{\chi^2_{2(K-1)} > U_O + V_O\} = \text{pr}\{\chi^2_{2(K-1)} > \sum_{i=1}^{K-1}(\chi_1^2)^{-1}\left(1 - P_{U_i}\right) + \sum_{i=1}^{K-1}(\chi_1^2)^{-1}\left(1 - P_{V_i}\right)\}.$$

It should be pointed out that when K=2, the above UV test is constructed based on the two statistics U and V as obtained by the Qiu and Sheng method. However, the Qiu and Sheng approach obtains the overall p-value from U and V in a different way. Our simulation study (data not shown) indicates that our proposed UV test is usually more powerful than the Qiu and Sheng test when K=2.

3. A Simulation Study

In this section, we conduct a simulation study to demonstrate the performance of the U test, the V test and the UV test. We will compare these methods with the commonly used log-rank (LR) test, and the Gehan-Wilcoxon (GW) test for multiple samples. In this simulation study, we assume there are three treatment groups with hazard rate functions, $h_1(t), h_2(t)$, and $h_3(t)$,

respectively. We assume the censoring times are uniformly distributed between 0 and 2. Two different sample sizes are considered: 50 and 100 for each group. The empirical type I error rate and power are estimated based on 1000 replicates using the significance level of 0.05. When the null hypothesis is true, we set $h_1(t) = h_2(t) = h_3(t) = 1$ to estimate the type I error rate. To estimate the power, we consider the following four cases, as shown in Figure 1: (a) $h_1(t) = 1, h_2(t) = 1.2,\ h_3(t) = 1.5$; (b) $h_1(t) = 1, h_2(t) = 1.5,\ h_3(t) = 1 + 0.5t$; (c) $h_1(t) = 1, h_2(t) = 1.2,\ h_3(t) = 2t$; and (d) $h_1(t) = 1, h_2(t) = 0.5 + t,\ h_3(t) = 2t$. Note that, in (a), the three hazard rate functions are parallel; in (b), the three hazard rate functions cross once (beyond time 0); in (c), the three hazard rate functions cross twice; and in (d), the three hazard rate functions cross three times. From (a) to (d), the degree of crossing among the three risk rate functions is in an increasing order.

Table 1 reports the empirical type I error rate and power for each method considered in the simulation. From Table 1, we can see that the U test, the V test and the proposed UV test all control the type I error rate quite well. We also observe that the U test has similar performance to that of the LR test, because they are both extensions of the two-sample LR test in different ways. When there are no or very few crossings among the hazard rate functions (e.g., cases (a) and (b)), the V test has no or little power; therefore, the proposed UV test has slightly lower power than the U test. However, when there are more crossings (e.g., cases (c) and (d)), the V test can be very powerful while the LR and GW tests lose power dramatically. From this simulation example, it can be seen that the U, LR and GW tests are powerful when there are no or few crossings, the V test is powerful when there are more crossings, and the UV test is always close to the best test in all cases considered. Therefore, the UV test is robust to different patterns of the hazard rate functions. It should be pointed out that in the simulation study, we simply use the

number of crossing to indicate the degree of crossing. However, the performance of the V test, and the UV test may also depend on when and how the crossing take place.

4. A Real Data Example

In this section, we apply the proposed test to a real data set from the randomized, double-blinded Digoxin Intervention Trial (The Digitalis Investigation Group, 1997). In the trial, patients with left ventricular ejection fractions of 0.45 or less were randomly assigned to digoxin (3397 patients) or placebo (3403 patients) groups. A primary outcome was the mortality due to worsening heart failure. Figure 2 (a) plots the Kaplan-Meier curves for the two treatment groups.

In the original study, the authors used the LR test and obtained a p-value of 0.06, indicating that the evidence of the effectiveness of digoxin, in terms of reducing the mortality due to worsening heart failure, is at most marginal. However, based on the above Kaplan-Meier curves, the proportional hazard rates assumption made in the LR test may not be valid, and tests other than the LR should be considered. In fact, we obtain p-values from other methods as shown in Table 2 (a). The p-value from the UV test is 0.021; there is a relatively stronger evidence to support the effectiveness of the drug.

We then consider the possible interaction between the treatment and gender. Table 3 summarizes the data for the four groups. Figure 2 (b) plots the Kaplan-Meier curves for the four groups, which clearly shows that the LR test is not optimal as the proportional hazard rates assumption is violated. Table 2 (b) reports the p-values from different tests. Both the V test and the UV test obtain very small p-values. It is then interesting to compare the two treatments among male patients: M-P vs. M-D. Table 2 (c) gives the p-values from various tests. The p-value from the UV test is 0.0054; there is strong evidence that the drug is effective among male patients. Finally, we compare the treatments among female patients; the p-values from all the methods considered are shown in Table 2 (d). All tests obtained large p-values; there is no

evidence from the data to support the effectiveness of the drug for female patients, although it is noticeable that there were much fewer female patients in the study.

5. Discussion and Conclusion

In this paper we extend the two-sample LR test to multiple samples and call this extension the U test. The U test has similar performance as the traditional LR test, and therefore, it is powerful when the hazard rate functions are parallel. We also proposed another test, called the V test, which is designed specifically for the situation when the hazard rate functions cross each other and when the U test fails. The proposed UV test is a robust approach that combines information from the U test and the V test in an efficient way. If we have information about the hazard rate functions prior to seeing the data, we can use the U test, or the V test only when it is appropriate. For example, if the alternative hypothesis is directional, $S_1(t) \geq S_2(t) \geq \cdots \geq S_K(t)$ (or $S_1(t) \leq S_2(t) \leq \cdots \leq S_K(t)$), the one-sided p-values obtained based on the test statistics U_k $(k = 1,2,\cdots,K-1)$ can be used to obtain an overall p-value. However, if this kind of information is unavailable, the proposed UV test would be a good choice.

In each of the U, V and UV tests, we choose to combine the related p-values using the chi-square distribution with df 1 as this method is robust (Chen, 2011; Chen and Nadarajah, 2014). Although other robust methods of combing p-values, such as the Fisher test (Fisher, 1932), are also possible, the weighted z tests are not recommended here since they are not robust and may lose power dramatically in some situations. In addition, for the UV test, we can combine the two p–values from the U and V tests, namely, P_U and P_V, with different df values df_U and df_V, i.e., $P_{UV} = \text{pr}\{\chi^2_{df_U+df_V} > \left(\chi^2_{df_U}\right)^{-1}(1-P_U) + \left(\chi^2_{df_V}\right)^{-1}(1-P_V)\}$. If we know that the hazard rate functions are mainly parallel, we can assign a large number for df_U and a relatively small number for df_V. On the other hand, if the hazard rate functions are dominated by the crossing, we can

assign a small number for df_U and a large number for df_V. When $df_U = df_V = K-1$, it is the proposed UV test.

In summary, the UV test proposed in this paper is a flexible and robust approach, which has been confirmed by a simulation study and a real data application. More specifically, we have shown that this approach has good performance in terms of controlling the type I error rate and the detecting power. In some situations, the gain in power is substantial compared with other commonly used methods, such as the LR test and the GW test.

ACKNOWLEDGEMENTS

The authors are grateful to editor Professor Yi-Hau Chen for his helpful comments and suggestions that lead to improvement of the paper. The first author also acknowledges the support from the internal research funds awarded by Indiana University School of Public Health-Bloomington.

REFERENCES

Buyske, S., Fagerstrom, R., and Ying, Z. (2000). A class of weighted log-rank tests for survival data when the event is rare. *Journal of the American Statistical Association* **95**, 249-258.

Chen, Z. (2011). Is the weighted z-test the best method for combining probabilities from independent tests? *Journal of Evolutionary Biology* **24**, 926-930.

Chen, Z., and Nadarajah, S. (2014). On the optimally weighted z-test for combining probabilities from independent studies. *Computational statistics & data analysis* **70**, 387–394.

Fisher, R. A. (ed) (1932). *Statistical Methods for Research Workers.* Edinburgh: Oliver and Boyd.

Fleming, T. R., and Harrington, D. P. (1991). *Counting processes and survival analysis*. New York: Wiley

Gehan, E. A. (1965). A generalized Wilcoxon test for comparing arbitrarily singly-censored samples. *Biometrika* **52**, 203-223.

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 modelling. *Statistics in Medicine* **26**, 375-391.

Mantel, N., and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies. *J natl cancer inst* **22**, 719-748.

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

Park, K. Y., and Qiu, P. (2014). Model selection and diagnostics for joint modeling of survival and longitudinal data with crossing hazard rate functions,. *Statistics in Medicine* **33**, 4532–4546.

Qiu, P., and Sheng, J. (2008). A two-stage procedure for comparing hazard rate functions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70**, 191-208.

The Digitalis Investigation Group (1997). The effect of digoxin on mortality and morbidity in patients with heart failure. *N Engl J Med* **336**, 525-533.

Van der Vaart, A. W. (1998). *Asymptotic statistics*: Cambridge university press.

Yang, S., and Prentice, R. (2010). Improved Logrank-Type Tests for Survival Data Using Adaptive Weights. *Biometrics* **66**, 30-38.

APPENDIX

For the proofs, first, we define the following statistics:

$$Z_{1k}\{w_{k:1}^{(1)} = Z_{1k}(w_{k:1}) = h_k \sum_{i=1}^{D_k} w_{k:1i}\{d_{i(k+1)} - Y_{i(k+1)} \sum_{j=1}^{k+1} d_{ij} / \sum_{j=1}^{k+1} Y_{ij}\}, \quad \text{(A.1)}$$

$$Z_{2k}\{w_{k:2}^{(r_k)}\} = h_k \sum_{i=1}^{D_k} w_{k:2i}^{(r)}\{d_{i(k+1)} - Y_{i(k+1)} \sum_{j=1}^{k+1} d_{ij} / \sum_{j=1}^{k+1} Y_{ij}\}, \quad \text{(A.2)}$$

where $k = 1,2, \ldots, K-1$, $h_k = \sqrt{\sum_{j=1}^{k+1} n_j / \{n_{k+1}(\sum_{j=1}^{k} n_j)\}}$, D_k is the total number of distinct failure times from groups $1, 2, \ldots, k+1$, and $w_{k:l}^{(r)} = \{w_{k:l1}^{(r)}, w_{k:l2}^{(r)}, \cdots, w_{k:lD_k}^{(r)}\}'$ $(l = 1,2)$ are suitable weight functions defined as follows.

$w_{k:1i}^{(1)} = 1$, for i=1,.2,…, D_k, and

$$w_{k:2i}^{(r_k)} = \begin{cases} -1, & if\ i = 1,2, \cdots, m_k \\ c_{r_k}, & otherwise \end{cases}, \quad \text{(A.3)}$$

where

$$c_{r_k} \triangleq \int_0^{F_{k+}^{-1}(r_k)} \frac{L_{k1}(s)L_{k2}(s)}{p_{k1}L_{k1}(s)+p_{k2}L_{k2}(s)} dF_{k+}(s) \Big/ \int_{F_{k+}^{-1}(r_k)}^{u_k} \frac{L_{k1}(s)L_{k2}(s)}{p_{k1}L_{k1}(s)+p_{k2}L_{k2}(s)} dF_{k+}(s), \quad \text{(A.4)}$$

$F_{k+}(t) = 1 - S_{k+}(t)$ is the cumulative distribution function of the survival time using the pooled data from groups $1, 2, \dots, k+1$; $p_{k1} = \lim_{n\to\infty} \frac{n_{k1}}{n_{k+}}$, $p_{k2} = \lim_{n\to\infty} \frac{n_{k2}}{n_{k+}}$; $m_k = [D_k r_k]$, and $u_k = \inf [s: \min\{\pi_{k1}(s), \pi_{k2}(s)\} = 0]$. Note that the weights $w_{k:2i}^{(r_k)}$ defined in (A.3) and (A.4) are analogous to those defined in (2) and (3).

For the covariance between $Z_{lk}\{w_{k:l}^{((r_k)^{l-1})}\}$ and $Z_{l'k'}\{w_{k':l'}^{\left((r_k)^{l'-1}\right)}\}$ $(k \neq k')$, we have the following major result.

Lemma 2. If $k \neq k'$, $cov[Z_{lk}\{w_{k:l}^{((r_k)^{l-1})}\}, Z_{l'k'}\{w_{k':l'}^{\left((r_k)^{l'-1}\right)}\}] = 0$ for $l, l' = 1,2$, and $k, k' = 1,2, \dots, K-1$.

Proof of Lemma 2.

First we define the following counting processes:

$Y_{ij}(t) = I(X_{ij} > t), i = 1,2, \cdots, K, j = 1,2, \cdots, n_i$,

$N_{ij}(t) = I(X_{ij} < t, \delta_{ij} = 1), i = 1,2, \cdots, K, j = 1,2, \cdots, n_i$,

$\bar{Y}_{k1}(t) = \sum_{j=1}^{n_{k+1}} Y_{(k+1)j}(t)$, $k = 1,2, \cdots, K-1$,

$\bar{Y}_{k2}(t) = \sum_{i=1}^{k} \sum_{j=1}^{n_i} Y_{ij}(t)$, $k = 1,2, \cdots, K-1$,

$\bar{Y}_k(t) = \bar{Y}_{k1}(t) + \bar{Y}_{k2}(t) = \sum_{i=1}^{k+1} \sum_{j=1}^{n_i} Y_{ij}(t)$, $k = 1,2, \cdots, K-1$.

To simplify the notation, for given r_k, we use $w_{k:l}$ to denote $w_{k:l}^{\{(r_k)^{l-1}\}}$.

For the statistic $Z_{lk}(w_{k:l}; t) = \sum_{i=1}^{D_k} w_{k:li}\{d_{i(k+1)} - Y_{i(k+1)} \sum_{j=1}^{k+1} d_{ij} / \sum_{j=1}^{k+1} Y_{ij}\}$, where $l = 1,2, k = 1,2, \dots, K-1$, it can be shown that

$Z_{lk}(w_{k:l}; t) = \sum_{i=1}^{k+1} \sum_{j=1}^{n_i} \int_0^t H_{ijlk}(s) dM_{ij}(s)$,

where $H_{ijlk}(t) = (-1)^{I_{\{i\in(1,2,\ldots,k)\}}} w_{k:l}(t)\bar{Y}_{k1}(t)\bar{Y}_{k2}(t)(\bar{Y}_k(t)(\bar{Y}_{k1}(t))^{I_{\{i\in(1,2,\ldots,k)\}}}(\bar{Y}_{k2}(t))^{I_{(i=k+1)}})^{-1}$,

and martingale $M_{ij}(t) = N_{ij}(t) - \int_0^t I_{\{X_{ij}\geq u\}} d\Lambda(u)$.

Therefore, $Z_{l'k'}(w_{k':l'i};t) = \sum_{i=1}^{k'+1}\sum_{j=1}^{n_i}\int_0^t H_{ijl'k'}(s)dM_{ij}(s)$, where

$H_{ijl'k'}(t) = (-1)^{I_{\{i\in(1,2,\ldots,k')\}}} w_{k':l'}(t)\bar{Y}_{k'1}(t)\bar{Y}_{k'2}(t)(\bar{Y}_{k'}(t)(\bar{Y}_{k'1}(t))^{I_{\{i\in(1,2,\ldots,k')\}}}(\bar{Y}_{k'2}(t))^{I_{\{i=k'+1\}}})^{-1}$.

Without loss of generality, in this proof we assume $k < k'$. With the weight functions $w_{k:l}(t)$ and $w_{k':l'}(t)$ defined as in (A.3), it is easy to check that the required conditions in theorem 2.6.2 of Fleming and Harrington (Fleming and Harrington, 1991) are met; therefore, based on that theorem, the covariance between $Z_{lk}(w_{k:l};t)$ and $Z_{l'k'}(w_{k':l'};t)$ can be expressed as:

$cov\{Z_{lk}(w_{k:l};t), Z_{l'k'}(w_{k':l'};t)\} = \sum_{i=1}^{k+1}\sum_{j=1}^{n_i}\int_0^t E\left\{H_{ijlk}(s)H_{ijl'k'}(s)I_{\{X_{ij}\geq s\}}\right\}\{1-\Lambda(s)\}d\Lambda(s)$.

Therefore, $cov\{Z_{lk}(w_{k:l};t), Z_{l'k'}(w_{k':l'};t)\} = \int_0^t E\{\sum_{i=1}^{k+1}\sum_{j=1}^{n_i} H_{ijlk}(s)H_{ijl'k'}(s)I_{(X_{ij}\geq s)}\}\{1-\Lambda(s)\}d\Lambda(s)$.

But,

$\sum_{i=1}^{k+1}\sum_{j=1}^{n_i} H_{ijlk}(s)H_{ijl'k'}(s) =$

$\sum_{i=1}^{k}\sum_{j=1}^{n_i} H_{ijlk}(s)H_{ijl'k'}(s) + \sum_{j=1}^{n_{k+1}} H_{(k+1)jlk}(s)H_{(k+1)jl'k'}(s) =$

$w_{k:l}(s)w_{k':l'}(s)\left[-\sum_{i=1}^{k}\sum_{j=1}^{n_i} I_{\{X_{ij}\geq s\}}\frac{\bar{Y}_{k1}(s)}{\bar{Y}_k(s)}\frac{\bar{Y}_{k'1}(s)}{\bar{Y}_{k'}(s)} + \sum_{j=1}^{n_{k+1}} I_{\{X_{(k+1)j}\geq s\}}\frac{\bar{Y}_{k2}(s)}{\bar{Y}_k(s)}\frac{\bar{Y}_{k'1}(s)}{\bar{Y}_{k'}(s)}\right] =$

$c\left\{-\sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}(s)\sum_{j=1}^{n_{k+1}} Y_{(k+1)j}(s) + \sum_{j=1}^{n_{k+1}} Y_{(k+1)j}(s)\sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}(s)\right\} = 0$, where

$c = w_{k:l}(s)w_{k':l'}(s)\frac{1}{\bar{Y}_k(s)}\frac{\bar{Y}_{k'1}(s)}{\bar{Y}_{k'}(s)}$.

Finally, let $u_{k,k'} = \inf[s: \min\{\pi_{k1}(s), \pi_{k2}(s), \pi_{k'1}(s), \pi_{k'2}(s)\} = 0]$, covariance $cov(Z_{lk}, Z_{l'k'}) = cov\{Z_{lk}(u_{k,k'}), Z_{l'k'}(u_{k,k'})\} = 0$ when $k \neq k'$.

Proof of theorem 1.

Note that

$$U_k = Z_{1k}(w_{k:1})/h_k\hat{\sigma}\{Z_{1k}(w_{k:1})\}\ (k = 1,2,\dots,K-1), \tag{A.5}$$

where $\hat{\sigma}\{Z_{1k}(w_{k:1})\} = \sqrt{\sum_{i=1}^{D_k} w_{k:1i}^2 \frac{Y_{k:i1}}{Y_{k:i}} \frac{Y_{k:i2}}{Y_{k:i}} \frac{Y_{k:i}-d_{k:i}}{Y_{k:i}-1} d_{k:i}}$, $Y_{k:i1} = Y_{i(k+1)}$, $Y_{k:i2} = \sum_{j=1}^{k} Y_{ij}$,

$Y_{k:i} = Y_{k:i1} + Y_{k:i2}$, $d_{k:i} = \sum_{j=1}^{k+1} d_{ij}$.

And

$$V_k = \sup_{D_{k\varepsilon} \le m_k \le D_k - D_{k\varepsilon}} (V_{k:m_k}), \tag{A.6}$$

where $V_{k:m_k} = Z_{2k}\{\widehat{w}_{k:2}^{(m_k)}\}/h_k\hat{\sigma}[Z_{2k}\{\widehat{w}_{k:2}^{(m_k)}\}]$,

$\hat{\sigma}[Z_{2k}\{\widehat{w}_{k:2}^{(m_k)}\}] = \sqrt{\sum_{i=1}^{D_k}\{\widehat{w}_{k:2i}^{(m_k)}\}^2 \frac{Y_{k:i1}}{Y_{k:i}} \frac{Y_{k:i2}}{Y_{k:i}} \frac{Y_{k:i}-d_{k:i}}{Y_{k:i}-1} d_{k:i}}$,

$\widehat{w}_{k:2}^{(m_k)} = \{\widehat{w}_{k:21}^{(m_k)}, \widehat{w}_{k:22}^{(m_k)}, \cdots, \widehat{w}_{k:2D_k}^{(m_k)}\}^T$,

$\widehat{w}_{k:2i}^{(m_k)} = \begin{cases} -1 & if\ i = 1,2,\cdots,m_k \\ \hat{c}_{m_k} & otherwise \end{cases}$,

$$\hat{c}_{m_k} = \sum_{i=1}^{m_k} \frac{\hat{L}_{k1}(t_i)\hat{L}_{k2}(t_i)}{\left(\frac{n_{k1}}{n_{k+}}\right)\hat{L}_{k1}(t_i)+\left(\frac{n_{k2}}{n_{k+}}\right)\hat{L}_{k2}(t_i)} \Delta\hat{S}_k(t_i) \Big/ \sum_{i=m_k+1}^{D_k} \frac{\hat{L}_{k1}(t_i)\hat{L}_{k2}(t_i)}{\left(\frac{n_{k1}}{n_{k+}}\right)\hat{L}_{k1}(t_i)+\left(\frac{n_{k2}}{n_{k+}}\right)\hat{L}_{k2}(t_i)} \Delta\hat{S}_k(t_i), \tag{A.7}$$

$n_{k1} = n_{k+1}$, $n_{k2} = \sum_{i=1}^{k} n_i$, $n_{k+} = n_{k1} + n_{k2} = \sum_{i=1}^{k+1} n_i$,

and $\hat{L}_{k1}(t)$ is the K-M estimate of the survival function for censoring time using the (k+1)th sample, and $\hat{L}_{k2}(t)$ is the K-M estimate of the survival function for censoring time using pooled sample from groups $1,2,\dots,k$, and $\hat{S}_k(t)$ is the K-M estimate of the survival function for survival time using pooled sample from groups $1,2,\dots,k+1$.

We have the following result for $\hat{c}_{m_k}$ in (A.7) (Qiu and Sheng, 2008): it convergences in probability to

$$c_{r_k} = \int_0^{F_{k+}^{-1}(r_k)} \frac{L_{k1}(s)L_{k2}(s)}{p_{k1}L_{k1}(s)+p_{k2}L_{k2}(s)} dF_{k+}(s) \Big/ \int_{F_{k+}^{-1}(r_k)}^{u_k} \frac{L_{k1}(s)L_{k2}(s)}{p_{k1}L_{k1}(s)+p_{k2}L_{k2}(s)} dF_{k+}(s). \tag{A.8}$$

The quantity defined in (A.8) is analogous to that in (3).

(i) This part can be proved in the same way as the proof of lemma 1.

(ii) From Corollary 7.2.1 in Fleming and Harrington (Fleming and Harrington, 1991), see also Appendix A in Qiu and Sheng (Qiu and Sheng, 2008), we know that under the null hypothesis $Z_{1k}(w_{k:1})/\hat{\sigma}\{Z_{1k}(w_{k:1})\}$ has an asymptotic standard normal distribution, and $\sigma^2\{Z_{1k}(w_{k:1})\}/\hat{\sigma}^2\{Z_{1k}(w_{k:1})\}$ convergences to 1 in probability, where

$\sigma^2\{Z_{1k}(w_{k:1})\} = \int_0^{u_k} w_{k:1}^2(s)\frac{L_{k1}(s)L_{k2}(s)}{p_{k1}L_{k1}(s)+p_{k2}L_{k2}(s)}dF_{k+}(s)$.

From lemma 2, we also have $cov(Z_{1k}, Z_{1k'}) = 0$ for $k \neq k'$. Therefore, $cov(U_k, U_{k'}) \rightarrow \frac{1}{\sigma(Z_{1k}(w_{k:1}))\sigma\left(Z_{1k'}(w_{k':1})\right)}cov(Z_{1k}, Z_{1k'}) = 0$, and U_k and $U_{k'}$ $(k, k' = 1,2, \dots, K-1, k \neq k')$ are asymptotically independent.

(iii) For any $k \neq k'$, as in (ii), it can be shown that under the null hypothesis $V_{k:m_k}$ and $V_{k':m_{k'}}$ are asymptotically independent. Notice that $V_k = \sup_{D_{k_\varepsilon} \leq m_k \leq D_k - D_{k_\varepsilon}}(V_{k:m_k})$, based on theorems 18.10, and 18.11 in van der Vaart (Van der Vaart, 1998),we have under the null hypothesis V_k and each $V_{k':m_{k'}}$ are asymptotically independent. Similarly, under the null hypothesis $V_{k'}$ and each $V_{k:m_k}$ are asymptotically independent. Therefore, under the null hypothesis V_k and $V_{k'}$ are asymptotically independent.

(iv) For any $k \neq k'$, as in (ii), it can be shown that under the null hypothesis U_k and each $V_{k':m_{k'}}$ are asymptotically independent. Therefore, again by theorems 18.10, and 18.11 in van der Vaart (Van der Vaart, 1998), under the null hypothesis U_k and $V_{k'}$ are asymptotically independent.

Proof of theorem 2.

Under the null hypothesis, the U test is the sum of $K-1$ asymptotically independent chi-square distributions each with df 1; and the V test is the sum of another $K-1$ asymptotically independent chi-square distributions each with df 1. The two sets of chi-square distributions are mutually asymptotically independent, therefore, under the null hypothesis, the U test and the V test statistics are asymptotically independent.

TABLES AND FIGURES

Table 1. Empirical type I error rate and power of each method obtained from 1000 replicates with nominal significance level of 0.05.

Sample size	Test	Type I error rate	Power			
			(a)	(b)	(c)	(d)
$n_1 = n_2 = n_3 = 50$	U	0.048	0.281	0.289	0.132	0.044
	V	0.043	0.036	0.079	0.845	0.721
	UV	0.047	0.187	0.244	0.832	0.628
	LR	0.048	0.273	0.310	0.185	0.066
	GW	0.052	0.235	0.268	0.449	0.215
$n_1 = n_2 = n_3 = 100$	U	0.047	0.545	0.495	0.301	0.072
	V	0.036	0.049	0.137	0.995	0.965
	UV	0.054	0.431	0.464	0.995	0.935
	LR	0.051	0.533	0.521	0.349	0.092
	GW	0.053	0.469	0.479	0.777	0.356

Table 2. P-values from various tests when compare (a) two groups (placebo, drug), (b) compare four groups (male placebo, male drug, female placebo, and female drug); (c) two groups (male placebo, male drug), and (d) two groups (female placebo, female drug).

Data used	U	V	UV	LR	GW
(a) P vs. D	0.061	0.040	0.021	0.061	0.050
(b) M-P, M-D, F-P, F-D	0.11	0.0083	0.0070	0.11	0.092
(c) M-P vs. M-D	0.019	0.026	0.0054	0.019	0.014
(d) F-P vs. F-D	0.66	0.69	0.84	0.66	0.66

Table 3. Death due to worsening heart failure by treatment and gender.

Group	Yes (%)	No (%)	Total
male-placebo (M-P)	358 (13.6)	2281 (86.4)	2639
male-drug (M-D)	300 (11.4)	2342 (88.6)	2642
female-placebo (F-P)	91 (11.9)	673 (88.1)	764
female-drug (F-D)	94 (12.5)	661 (87.5)	755

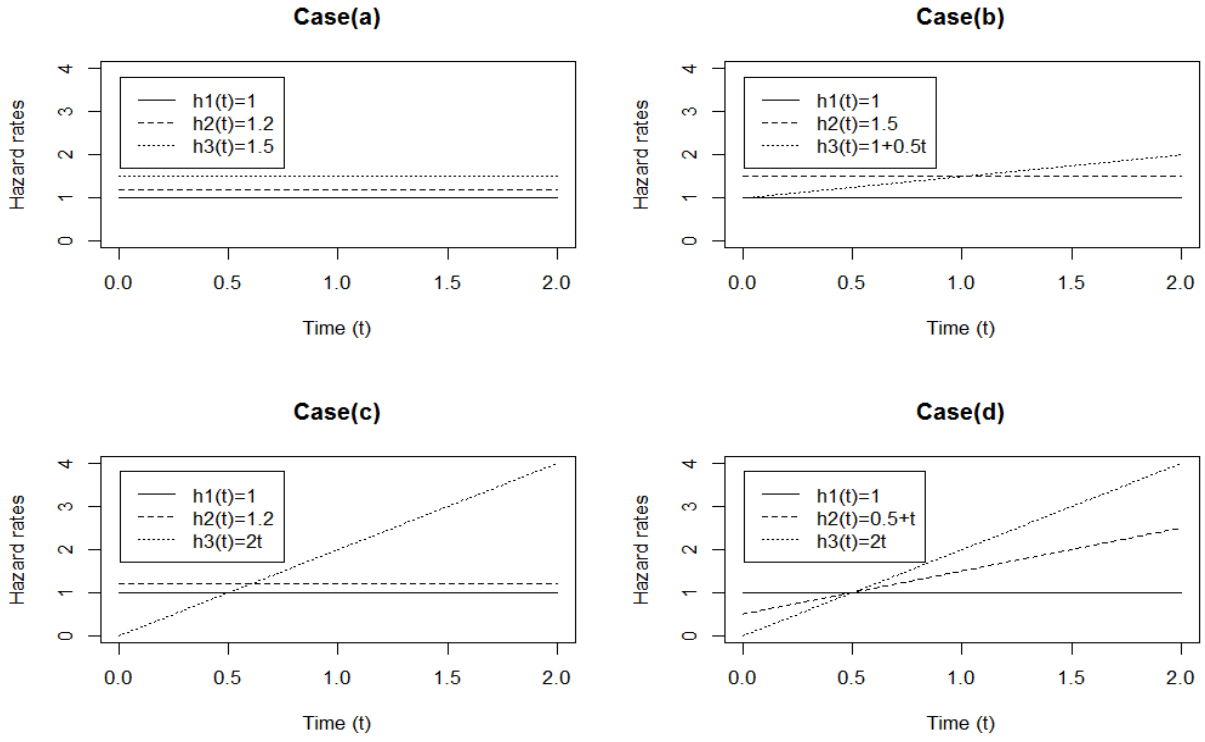

Figure 1. Four sets of hazard rate functions used in the simulation study for power estimation.

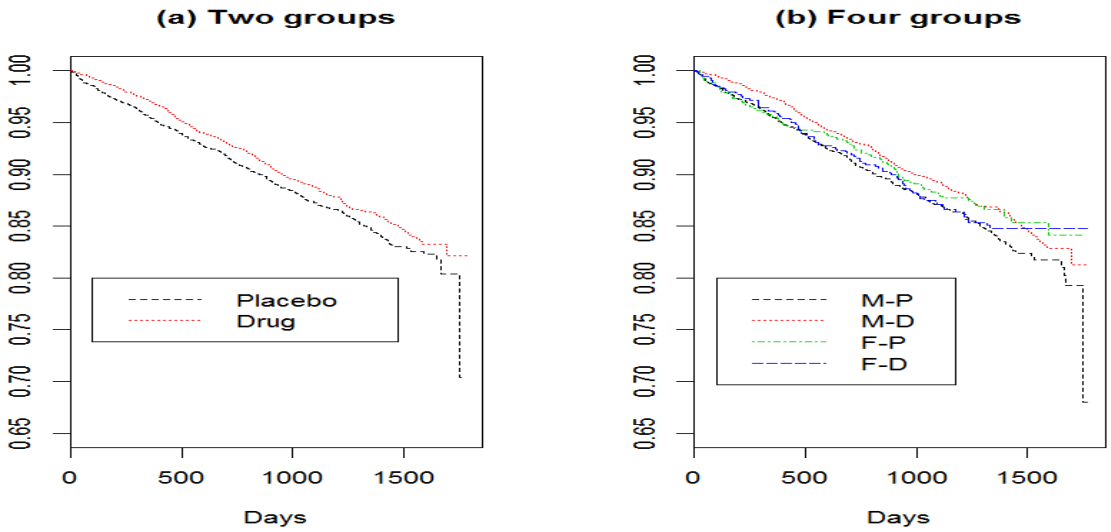

Figure 2. Kaplan-Meier curves for the two treatment groups (plot (a)) and the four treatment by gender groups (plot (b)).