東海大學統計研究所

碩士論文

指導教授：張玉媚博士

Data-driven Based Versatile Test for Two-sample

Problems with Right-censored Data

NATIONAL CENTRAL LIBRARY 國家圖書館 ROC

研究生：黃柏傑

中華民國九十九年七月

Data-driven Based Versatile Test for Two-sample Problems with Right-censored Data

Po-Chieh, Huang

Dept. of Statistics, Tunghai University,

Taichung, 40704, Taiwan, R.O.C.

致謝

首先誠摯的感謝指導教授張玉媚博士，承蒙張老師的悉心教導，且不時指點我學習的方向及態度，使我得以拓展統計學領域的視野，並建立正確的處事態度，讓我獲益匪淺。另外，本論文蒙陳春樹博士與俞一唐博士於百忙之中審核，斧正謬誤，並給予精闢的建議，使本文更臻完善，心中不勝感激。

在研究所修業期間，感謝統計系諸多老師傳授課業知識。此外，感謝佳玲、雅文、淑姿、如君及小津等學姊在研究生事務的幫助與行政事務的協助，使我能夠專心於課業之上，而免於諸多雜務的紛擾。

在校兩年，多蒙學長姊的督促指點，並於基礎觀念的疏通及程式撰寫給予指正，在此特表感謝。感謝同窗夥伴們，在面對困難時共同砥礪、在閒暇時的閒話家常、抑或是趕作業的革命情感，都使苦悶的研究生活多加了些光彩。

感謝姿婷、惠雅、煒婷等總是能夠為我打氣、聽我吐苦水。感謝哲論和德綱，你們真的是我進步的動力，也是我的支柱。感謝姣嬋、佳吟、慧蘋、文慈、美玲等，謝謝你們在忙碌之餘支持、關心我，也總是花時間聽我訴苦、安撫我。

感謝家人的支持，使我無後顧之憂的在外地求學，謝謝你們。

Abstract

For the two-sample censored data problem, the weighted log-rank (WLR) tests and weighted Kaplan-Meier (WKM) test are commonly used for testing the equality of two survival distributions. Since each test has different advantages against various alternatives, it's hard to decide in advance which of the tests can be used to gain more power when the alternative is unknown. Hence, in order to combine the advantages of these two classes of tests, a versatile test based on WLR test and WKM test is then proposed. We develop a cross-validation versatile test to select appropriate weights in combining WLR and WKM which differs from Chi and Tsai who suggested the equal weights. Some numerical experiments are performed for illustrating the superiority of the proposed method and then the proposed testing procedure is applied to two real data sets.

Key Words: weighted log-rank tests; weighted Kaplan-Meier test; linear combination test; versatile test

摘要

針對雙樣本右設限資料，加權對數秩 (weighted log-rank) 檢定與加權 Kaplan-Meier (weighted Kaplan-Meier) 檢定是最常被使用來檢定兩個存活機率的分配是否相等的問題。因為這兩個檢定在針對不同的情況下各有優點，故很難在未知的情況下，先行挑選檢定，使其能夠擁有較大的檢定力。因此，為了要將此兩種檢定的優點結合，本論文中將著重於同時使用此兩種檢定的機動檢定。

我們延續 Chi 與 Tsai (2001) 的想法，針對加權對數秩和加權 Kaplan-Meier 此兩種檢定的線性組合，使用交叉驗證挑選此線性組合的權重，並與 Chi 和 Tsai 提出的建立在相同權重的線性組合來做比較。透過模擬研究說明我們所提出方法的優越性，並且將此檢定方法應用於實際資料。

關鍵字：加權對數秩檢定，加權 Kaplan-Meier 檢定，線性組合的檢定，

機動檢定

Contents

List of Tables

List of Figures

1. Introduction

In clinical trials, the primary objective is to evaluate the effect of an experiment agent by comparing the survival durations among some groups. Most situations are to test the equality of two survival distributions under randomly right censorship. The most commonly used test statistics for testing the equality of two survival distributions are the log-rank statistic (Mantel, 1966) and the Peto-Prentice-Wilcoxon (PPW) statistic (Gehan, 1965; Peto and Peto, 1972; Prentice, 1978), where the log-rank test is the locally most power test against proportional hazards alternatives, while the PPW test benefits the difference of hazards at early times. In addition, the WLR statistics is based on the integrated weighted differences between two estimated hazard functions and is often used to test the related issues. Unfortunately, the WLR statistics would be insensitive against the stochastic ordering alternatives particularly when the hazard functions of two groups are crossing.

Hence, Pepe and Fleming (1989) proposed a class of test statistics based on the integrated weighted differences in Kaplan-Meier (1958) estimates, and showed that these statistics is competitive with the log-rank test and PPW test under the proportional hazard and early hazard differences, respectively, and may perform better than WLR test under crossing hazards alternatives.

Furthermore, the weighted log-rank tests have various advantages against

different types of alternatives with different weights. However, for a given data, it's hard to know what type of alternatives to expect and hence the choice of weight function is unclear in practice. In order to maintaining better power across a wide range of alternatives, Lee (1996) proposed some compromise strategies based on a linear combination of or the maximum of selected members from the family of weighted log-rank tests. Also, Shen and Cai (2001) studied the versatile tests based on the maximum of selected members from the class of weighted Kaplan-Meier for randomized controlled screening trials. Recently, Lee (2007) suggested the tests based on a linear combination of or the maximum of the absolute value of selected members from the class of weighted log-rank tests to reduce the correlation among selected tests to gain power.

Since weighted log-rank and weighted Kaplan-Meier tests have different advantages against various alternatives, Chi and Tsai (2001) proposed a class of versatile tests based on a linear combination of or the maximum of these two types of tests for two independent samples of right-censored data. The result in Chi and Tsai (2001) showed that these versatile tests were more robust in detecting different alternatives than the linear combination tests which proposed by Lee (1996). However, Chi and Tsai (2001) combined WLR and WKM statistics by using equal weights, say 0.5 and 0.5, resulting in the test that may not maintain the highest power across a

broad range of alternatives. In this paper, we will continued the idea of Chi and Tsai's linear combination test and propose a cross-validation based method to select the appropriate weights for combining WLR and WKM statistics.

The rest of this paper is organized as follows: Section2 reviews the weighted log-rank test and weighted Kaplan-Meier test, for testing the equality of two survival distributions in the presence of independent right censorship. Section 3 introduces the proposed data-driven versatile test based on a cross-validation approach. Some comparative results in terms of the error rates and powers based on a simulation study are shown in Section 4. The proposed method is illustrated two data sets in Section5. Finally, concluding remarks are presented in the last section.

2. Weighted log-rank and weighted Kaplan-Meier tests

Let T_{ij} and C_{ij} denote the survival time and censoring time, respectively, for the j^{th} patient in the i^{th} group, $i = 1, 2,\ j = 1, \cdots, n_i$. We assume T_{ij} and C_{ij} are independent. When data are subject to random right censorship, we can only observe the random variables $X_{ij} = min\{T_{ij},\ C_{ij}\}$ and $\delta_{ij} = I\{T_{ij} \leq C_{ij}\}$, where $I\{E\}$ is an indicator function, taking value 1 if the event E occurs and 0 otherwise. Let $S_i(t) = P(T_{ij} > t)$ and $G_i(t) = P(C_{ij} > t)$ be the survival functions of failure and censoring times for the i^{th} population, respectively. Let $N_i(t) = \sum_{j=1}^{n_i} I\{X_{ij} \leq t,\ \delta_{ij} = 1\}$ being the counting process of the number of failure before specified time t and let $Y_i(t) = \sum_{j=1}^{n_i} I\{X_{ij} \geq t\}$ being the risk process for $i = 1, 2$.

2.1 Weighted Log-rank Test

For testing the equality of two survival distributions, the null hypothesis is set to be the equality of two survival functions generally, that is $H_0 : S_1(t) = S_2(t)$ for all t, and the alternative hypothesis may be the omnibus alternatives $H_1 : S_1(t) \neq S_2(t)$ for some t or the stochastic ordering alternatives $H_1 : S_1(t) \leq S_2(t)$ for all t with $S_1(t) < S_2(t)$ for some t. The commonly used test statistic, called weighted log-rank (WLR) statistic, was proposed by Fleming and Harrington (1991) and can be expressed as

$$K_1 = \sqrt{\frac{n_1+n_2}{n_1 n_2}} \int_0^T \widehat{W}(t) \frac{Y_1(t)Y_2(t)}{Y_1(t)+Y_2(t)} \left\{ \frac{dN_1(t)}{Y_1(t)} - \frac{dN_2(t)}{Y_2(t)} \right\},$$

where $T = sup\{t : Y_1(t)Y_2(t) > 0\}$, $\widehat{W}(t)$ is the predictable weight function of the form $\{\hat{S}(t-)\}^{\tau}\{1-\hat{S}(t-)\}^{\gamma}$ with $\tau \geq 0$, $\gamma \geq 0$, and $\hat{S}(t)$ is the estimated probability of not fail before time t based on Kaplan-Meier (1958) estimator for combined samples. As suggested in Fleming and Harrington (1991), $\widehat{W}(t)$ is referred to the family of censored data rank tests $\{G^{\tau,\gamma} : \tau \geq 0, \gamma \geq 0\}$. Note that $G^{0,0}$ and $G^{1,0}$ corresponds to the log-rank test statistic which benefits to the proportional hazards model and the PPW test statistic which is appropriate for testing the early difference of hazards, respectively. Since the null asymptotic distribution of K_1 is normal with mean zero and variance σ_{11}^2 which can be consistently estimated by

$$\hat{\sigma}_{11}^2 = \frac{n_1+n_2}{n_1 n_2} \int_0^T \widehat{W}^2(t) \frac{Y_1(t)Y_2(t)}{Y_1(t)+Y_2(t)} \left(1 - \frac{\Delta N_1(t) + \Delta N_2(t) - 1}{Y_1(t)+Y_2(t)-1} \right) \frac{d\big(N_1(t)+N_2(t)\big)}{Y_1(t)+Y_2(t)}, \quad (1)$$

where $\Delta N_i(t) = N_i(t) - N_i(t-)$, $i = 1, 2$. Hence, a two-sample weighted log-rank test rejects H_0 and concludes that the survival probability is better for group 2 if $K_1^* = K_1/\sqrt{\hat{\sigma}_{11}^2} \geq z_\alpha$, where z_α is the upper α^{th} percentile of a standard normal distribution.

2.2 Weighted Kaplan-Meier Test

Since weighted log-rank tests are based on ranks, they might not be sensitive to the magnitude of the difference in survival times against a specific alternative. Therefore, a weighted Kaplan-Meier statistic (Pepe and Fleming, 1989) based on the integrated weighted difference in Kaplan-Meier (1958) estimators for censored data is defined as

$$K_2 = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \int_0^{T_c} \widehat{\omega}(t) \left\{ \hat{S}_2(t) - \hat{S}_1(t) \right\} dt$$

where $T_c = sup\left\{ t : min\left(\hat{G}_1(t), \hat{G}_2(t), \hat{S}_1(t), \hat{S}_2(t) \right) > 0 \right\}$ with $\hat{S}_i(t)$ and $\hat{G}_i(t)$ being Kaplan-Meier (1958) estimators of the true and censoring survival functions of group i, respectively, and $\widehat{\omega}(t)$ is a random weight function given by

$$\widehat{\omega}(t) = \frac{\hat{G}_1(t-)\hat{G}_2(t-)}{p_1 \hat{G}_1(t-) + p_2 \hat{G}_2(t-)},$$

where $p_i = n_i/(n_1 + n_2)$. Note that $\widehat{\omega}(t)$ satisfies the stability condition in Pepe and Fleming (1989), that downweights the difference $\hat{S}_2(t) - \hat{S}_1(t)$ in the integrand over later time periods when there is heavy censoring. Similar to the WLR test statistic, the asymptotic distribution of K_2 under the null hypothesis follows a normal with mean zero and variance σ_{22}^2 can be consistently estimated by

$$\hat{\sigma}_{22}^2 = -\int_0^{T_c} \left\{ \int_t^{T_c} \widehat{\omega}(u) \hat{S}(u) du \right\}^2 \frac{p_1 \hat{G}_1(t-) + p_2 \hat{G}_2(t-)}{\hat{G}_1(t-)\hat{G}_2(t-)} \frac{d\hat{S}(t-)}{\hat{S}(t)\hat{S}(t-)}, \qquad (2)$$

where $\hat{S}(t-)$ is the estimated probability of not fail before time t based on

Kaplan-Meier (1958) estimator for combined samples and $d\hat{S}(t-) = \hat{S}(t) - \hat{S}(t-)$. Again, a two-sample weighted Kaplan-Meier test rejects H_0 and concludes that the survival probability is better for group 2 if $K_2^* = K_2/\sqrt{\hat{\sigma}_{22}^2} \geq z_\alpha$, where z_α is the upper α^{th} percentile of a standard normal distribution. Pepe and Fleming (1989) also showed that the resulting test statistic is a competitor to the log-rank test for the proportional hazards alternative, and may perform better under early and crossing hazards difference alternatives. Nevertheless, it is not sensitive to the late survival difference since the weight function is chosen to put less weight over later time period if censoring rate is heavy.

Although weighted log-rank and weighted Kaplan-Meier tests have different advantages against various alternatives, each of the tests cannot satisfactorily perform under various situations. Hence, a data-driven based test will be illustrated in the next section.

3. The proposed method

A natural idea is to combine the WLR and WKM tests. In this paper, we focus on testing the equality of two survival distributions for right censorship data and perform a linear combination between WLR and WKM tests. As suggested in Chi and Tsai (2001), they used equal weights, say 0.5 and 0.5, for combining to detect the test. However, it may not maintain the highest power across a broad range of alternatives and is not data-adaptive. In order to combine the advantages of WLR and WKM tests but mitigate their weaknesses for various alternatives, we are trying to propose a new testing procedure by controlling the weight from data. Therefore, a cross-validation based idea for selecting an appropriate weight between WLR and WKM tests is introduced in the following.

We consider the following class of test statistics as suggested in Chi and Tsai (2001):

$$\{K(\beta) = \beta K_1^* + (1-\beta)K_2^* : 0 \le \beta \le 1\}.$$

Under the null hypothesis $H_0 : S_1(t) = S_2(t)$, the asymptotic distribution of $K(\beta)$ for a given β follows a normal distribution with mean zero and variance can be consistently estimated by

$$\hat{\sigma}_\beta^2 = \beta^2 + (1-\beta)^2 + \beta(1-\beta)\hat{\rho},$$

where $\hat{\rho} = \hat{\sigma}_{12}/\sqrt{\hat{\sigma}_{11}^2\hat{\sigma}_{22}^2}$,

$$\hat{\sigma}_{12} = -\int_0^{T_c} \widehat{W}^2(t) \left\{ \int_t^{T_c} \widehat{\omega}(u)\hat{S}(u)du \right\} \frac{d\hat{S}(t)}{\hat{S}(t)},$$

and $\hat{\sigma}_{11}^2$ and $\hat{\sigma}_{22}^2$ are given in (1) and (2). Thus, one can conclude that the survival rate is better for group 2 at the α-level if $K^*(\beta) = K(\beta)/\sqrt{\hat{\sigma}_\beta^2} \geq z_\alpha$. Then, the procedure is to select an appropriate weight $\hat{\beta} \in [0,1]$ from data to use in the test statistic $K^*(\beta)$ that is sensitive to detect the survival differences between the two groups. The idea of cross-validation method, which can be reviewed, for example, in Shao (1993) and Arlot and Celisse (2009), is usually used for model selection. Therefore, based on a cross-validation approach, we proposed a criterion for selecting weight which is data-adaptive:

$$\hat{\beta} \equiv arg \min_{\beta \in [0,1]} \sum_i^{n_1} \sum_j^{n_2} \left(K_{-i,-j}^*(\beta) - K^*(\beta) \right)^2,$$

where $K_{-i,-j}^*(\beta)$ is a $K^*(\beta)$ calculated from the data with deleting the i^{th} and j^{th} sample points of group 1 and 2, respectively. Then, the proposed data-driven versatile statistic is obtained as follows:

$$K^*(\hat{\beta}) = \hat{\beta} K_1^* + (1 - \hat{\beta}) K_2^*. \quad (3)$$

Hence, the proposed versatile test can be performed for testing the equality of two survivals distributions regardless of various alternatives.

4. Simulations

To understand the error rates and powers of the proposed data-driven versatile test, we conduct a simulation study to examine the performance under various situations. A null case and four distinct alternative cases are considered in the simulation study. In the error rate study, the common survival distribution for group 1 and 2 is generated from exponential distribution with scale parameter 1. The alternatives investigated include Weibull proportional hazard models, early, late, and crossing hazard differences alternatives. Figure 1 presents the survival functions of these alternatives. Note that configurations II, III, and IV are all generated from piecewise exponential distributions with various hazard functions. Each of the cases is carried out under moderate sample sizes (30 and 50 sample sizes per group) and uniform (0,2) censorship. Note that the censoring proportion is 0.43 for both 30 and 50 sample sizes under the null hypothesis. While in the configurations I to IV, the two groups suffer different probabilities of censorship ranging from 0.24 to 0.50. Estimated error rates and powers of these test statistics at the 0.05 level of significance are based on 3,000 replications. Thus, the standard error of the error rate estimator is about 0.004 ($\approx \sqrt{(0.05 \times 0.95)/3000}$).

Table 1 and 2 present the results on the error rates and powers under 30 and 50 sample sizes for various test statistics, which include WKM, four individual member

in $G^{\tau,\gamma}$ class as suggested in Fleming and Harrington (1991), and four proposed data-driven versatile statistics $K^*_{\tau,\gamma}(\hat{\beta})$. Let $K^*_{\tau,\gamma}(\hat{\beta})$ denote the combination of WKM and $G^{\tau,\gamma}$. We also perform and compare the corresponding versatile statistics with fixed weight $\hat{\beta} \equiv 0.5$ in (3) as suggested by Chi and Tsai (2001), and denoted as $K^*_{\tau,\gamma}(0.5)$. The average of the estimated weight $\hat{\beta}$ in (3) based on 3000 replications for 30 and 50 sample sizes for each configuration are also reported in Table 3 and 4, respectively.

In Table 1, the type I error rate of $K^*_{0,1}(\hat{\beta})$ and $K^*_{1,1}(\hat{\beta})$ are much higher than the pre-specified error rate 0.05, but when the observation raise to 50 in Table 2, the type I error rate of these two tests is under controlled. The rest of the tests considered here maintain their error rates well. On the other hand, the performance of the power of the proposed data-driven versatile test is nearly as sensitive as the most powerful individual statistic for detecting a specific alternative. Moreover, it performs better than Chi and Tsai's (2001) test for almost all cases. Table 3 and 4 showed that it's not appropriate to combine WKM and WLR tests with a fixed weight 0.5.

5. Examples

In this section, the data-driven versatile tests developed in Section 3 are illustrated through two real examples. The first data set (data set 1) was conducted by Ichida et al. (1993) to evaluate a protocol change in disinfectant practices in a large Midwestern university medical center. Infection of a burn wound is a common complication resulting in extended hospital stays and in the death of severely burned patients. Control of infection remains a prominent component of burn management. The purpose of the burn wound infections study is to compare a routine bathing care method, which initial surface decontamination with 10% povidone-iodine followed with regular bathing with Dial soap, with a body-cleansing method using 4% chlorhexidine gluconate. In data set 1, 154 patient records and charts were reviewed, including 70 patients in the body-cleansing method group with 44% (31 patients) censored data and 84 patients in the routine bathing care method group with 29% (24 patients) censored data. The time to excision was recorded (in days). Figure 2 displays the estimated survival function and log cumulative hazard function and shows a difference between the two estimated survival functions over middle time period. The relevant summary statistics obtained from various tests and the associated one-sided p-values are shown in table 5. At the 0.05 significance level, the p-values for WKM and WLR statistics are less than 0.05, except $G^{0,1}$. Although the performance of

$K^*_{\tau,\gamma}(0.5)$ and $K^*_{\tau,\gamma}(\hat{\beta})$ better than $G^{0,1}$, $K^*_{\tau,\gamma}(\hat{\beta})$ still gain more power by selecting the appropriate weights.

Next, the second data set (data set 2) was obtained from Pantadosi (1997) originally designed from a randomized clinical trial for evaluating the benefit of cytoxan, doxorubicin, and platinum (CAP) as an adjuvant to radiotherapy (R) for treatment of locally advanced non small-cell lung cancer patients. A total of 164 patients were randomized to receive either the R only or the combined treatment of R and CAP group, denoted by R+CAP, between 1979 and 1985. In data set 2, there are 86 patients in the R only group with 18% (14 patients) censored data, while 78 patients in the R+CAP group produce 16% (14 patients) censored data. Figure 3 displays the estimated survival function and log cumulative hazard plot. However, it's hard to see which period reveals significant visual differences between the two estimated survival functions, resulting in the selection of weight function for WLR test is difficult. Therefore, our proposed data-driven versatile test seems usefully here. The relevant summary statistics and the associated one-sided p-values are shown in table 6. Obviously, the p-value of $G^{1,0}$ is the only one smaller than 0.05. Therefore, the corresponding versatile test $K^*_{1,0}(\hat{\beta})$ can still chose an appropriate weight to reject the null hypothesis. It indicates that the combined treatment R+CAP could prolong the lifetime of the patients with lung cancer. On the contrary, we find that the

all of versatile tests $K^*_{\tau,\gamma}(0.5)$ are fail to detect the survival difference between the two groups.

These examples denote the shortcomings of $K^*_{\tau,\gamma}(0.5)$ and confirm the behavior of these tests as demonstrated in the simulation study.

6. Concluding remarks

In this paper, we develop data-driven versatile tests statistic based on linear combination of WLR and WKM to preserve better power by using a cross-validation approach for selecting an appropriate weight, results the weight that is data-adaptive, across a broad range of alternatives for two independent right-censored data. The performance of the proposed data-driven versatile tests in this paper is superior to Chi and Tsai (2001) in terms of the power testing. As seen in two examples, before testing the equality of survival distributions, investigators must choose a weighted function according to their clinical knowledge or based on the survival plots, so that the used test statistic would be sensitive to a certain alternative. However, when lacking in the clinical knowledge or the visual difference is not obvious in the survival plots, our proposed method provide a convenience way to detect the difference between survival distributions. Because the proposed method can be performed for testing the equality of two survival distributions regardless of various alternatives, it becomes an attractive feature.

Reference

1. Arlot, S. and Celisse, A. Cross-validation methods for model selection. *Technical Report, arXiv*, 2009, 0907.4728.

2. Chi, Y. C. and Tsai, M. H. Some versatile tests based on the simultaneous use of weighted logrank and weighted Kaplan-Meier statistics. *Communications in Statistics, Part B—Simulation and Computation*, 2001, 30, 743-759.

3. Fleming, T. R.; Harrington, D. P. Counting Processes and Survival Analysis. *New York: Wiley*, 1991.

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

5. Ichida, J. M., Wassell, J. T., Keller, M. D., and Ayers, L. W. Evaluation of Protocol Change in Burn-Care Management Using the Cox Proportional Hazards Model with Time-Dependent Covariates. *Statistics in Medicine*, 1993, 12, 301-310.

6. Kaplan, E. L.; Meier, P. Nonparametric estimator from incomplete observations. *Journal of the American Statistical Association*, 1958, 53, 457-481.

7. Lad, T., Rubinstein, L., Sadeghi, A. et al. The benefit of adjuvant treatment for resected locally advanced non-small cell lung cancer. *Journal of Clinical Oncology*, 1988, 6,9-17.

8. Lee, J. W. Some versatile tests based on the simultaneous use of weighted log-rank statistics. *Biometrics*, 1996, 52, 721-725.

9. Lee, S. H. On the versatility of the combination of the weighted. *Computational Statistics and Data Analysis*, 2007, 51, 6557-6564.

10. Mantel, N. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports*, 1966, 50, 163-170.

11. Pantadosi, S. Clinical Trials: A Methodologic Perspective. *New York: Wiley*, 1997.

12. Pepe, M, S,; Fleming, T. R. Weighted Kaplan-Meier statistics: a class of distance tests for censored survival data. *Biometrics*, 1989, 45, 497-507.

13. Peto, R.; Peto, J. Asymptotically efficient rank invariant test procedures (with discussion). *Journal of the Royal Statistical Society, Series A*, 1972, 135, 185-206.

14. Prentice, R. L. Linear rank tests with right censored data. *Biometrika*, 1978, 65, 167-169.

15. Shen, Y.; Cai, J. Maximum of the weighted Kaplan-Meier tests with applications to cancer prevention and screening trials. *Biometrics*, 2001, 57, 837-843.

16. Shao, J. Linear model selection by cross-validation. *Journal of the American Statistical Association*,1993,88,486–494.

Appendix

Table 1. Estimated error rates, powers and censoring rates for each case with $n_1 = n_2 = 30$ and censoring distribution $U(0, 2)$.

case	I	II	III	IV	Null
WKM	0.653	0.642	0.247	0.228	0.057
$G^{0,0}$	0.700	0.535	0.356	0.231	0.054
$G^{1,0}$	0.654	0.577	0.251	0.197	0.052
$G^{0,1}$	0.602	0.291	0.457	0.269	0.057
$G^{1,1}$	0.635	0.335	0.441	0.300	0.057
$K^*_{0,0}(0.5)$	0.682	0.609	0.299	0.227	0.054
$K^*_{1,0}(0.5)$	0.654	0.620	0.250	0.211	0.055
$K^*_{0,1}(0.5)$	0.683	0.522	0.378	0.273	0.055
$K^*_{1,1}(0.5)$	0.672	0.537	0.360	0.281	0.056
$K^*_{0,0}(\hat{\beta})$	0.693	0.652	0.309	0.244	0.058
$K^*_{1,0}(\hat{\beta})$	0.653	0.650	0.249	0.215	0.054
$K^*_{0,1}(\hat{\beta})$	0.707	0.661	0.382	0.309	0.075
$K^*_{1,1}(\hat{\beta})$	0.687	0.669	0.361	0.316	0.073
Censoring rate for group 1	0.244	0.241	0.307	0.402	0.431
Censoring rate for group 2	0.433	0.380	0.431	0.496	

Table 2. Estimated error rates, powers and censoring rates for each case with $n_1 = n_2 = 50$ and censoring distribution $U(0, 2)$.

case	I	II	III	IV	Null
WKM	0.857	0.830	0.387	0.930	0.045
$G^{0,0}$	0.882	0.736	0.507	0.821	0.045
$G^{1,0}$	0.848	0.776	0.360	0.143	0.046
$G^{0,1}$	0.786	0.401	0.641	0.449	0.052
$G^{1,1}$	0.831	0.484	0.634	0.588	0.051
$K^*_{0,0}(0.5)$	0.871	0.801	0.454	0.892	0.046
$K^*_{1,0}(0.5)$	0.855	0.815	0.375	0.907	0.046
$K^*_{0,1}(0.5)$	0.866	0.713	0.533	0.800	0.048
$K^*_{1,1}(0.5)$	0.867	0.730	0.529	0.832	0.045
$K^*_{0,0}(\hat{\beta})$	0.873	0.826	0.455	0.916	0.047
$K^*_{1,0}(\hat{\beta})$	0.853	0.828	0.372	0.918	0.046
$K^*_{0,1}(\hat{\beta})$	0.879	0.828	0.578	0.919	0.066
$K^*_{1,1}(\hat{\beta})$	0.871	0.832	0.541	0.921	0.059
Censoring rate for group 1	0.244	0.242	0.307	0.403	0.431
Censoring rate for group 2	0.432	0.381	0.433	0.497	

Table 3. The averages of estimated weight $\widehat{\boldsymbol{\beta}}$ for each case with $n_1 = n_2 = 30$ and censoring distribution $U(0, 2)$.

case	I	II	III	IV	Null
$K^*_{0,0}(\hat{\beta})$	0.330	0.399	0.616	0.645	0.738
$K^*_{1,0}(\hat{\beta})$	0.395	0.424	0.633	0.674	0.780
$K^*_{0,1}(\hat{\beta})$	0.264	0.289	0.576	0.506	0.634
$K^*_{1,1}(\hat{\beta})$	0.288	0.291	0.524	0.521	0.625

Table 4. The averages of estimated weight $\widehat{\boldsymbol{\beta}}$ for each case with $n_1 = n_2 = 50$ and censoring distribution $U(0, 2)$.

case	I	II	III	IV	Null
$K^*_{0,0}(\hat{\beta})$	0.192	0.238	0.511	0.146	0.752
$K^*_{1,0}(\hat{\beta})$	0.240	0.263	0.517	0.184	0.774
$K^*_{0,1}(\hat{\beta})$	0.189	0.167	0.644	0.090	0.655
$K^*_{1,1}(\hat{\beta})$	0.206	0.160	0.581	0.090	0.639

Table 5. The test statistics of various tests and the associated one-sided p-values for data set 1.

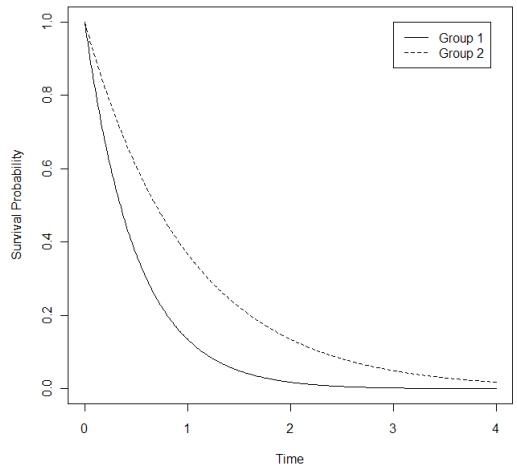
	z-value	p-value			z-value	p-value	z-value	p-value
WKM	3.028	0.001		$\widehat{\boldsymbol{\beta}}$			$\widehat{\boldsymbol{\beta}} = 0.5$	
$G^{0,0}$	2.691	0.004	$K^*_{0,0}(\hat{\beta})$	0.36	2.907	0.002	2.860	0.002
$G^{1,0}$	3.254	0.001	$K^*_{1,0}(\hat{\beta})$	0.80	3.209	0.001	3.141	0.001
$G^{0,1}$	0.936	0.175	$K^*_{0,1}(\hat{\beta})$	0.08	2.931	0.002	2.164	0.015
$G^{1,1}$	2.000	0.023	$K^*_{1,1}(\hat{\beta})$	0.22	2.875	0.002	2.611	0.005

Table 6. The test statistics of various tests and the associated one-sided p-values for data set 2.

	Statistic	p-value			Statistic	p-value	Statistic	p-value
WKM	1.063	0.144		$\widehat{\boldsymbol{\beta}}$			$\widehat{\boldsymbol{\beta}} = 0.5$	
$G^{0,0}$	1.136	0.128	$K^*_{0,0}(\hat{\beta})$	0.40	1.094	0.137	1.101	0.135
$G^{1,0}$	1.792	0.037	$K^*_{1,0}(\hat{\beta})$	1.00	1.792	0.037	1.460	0.072
$G^{0,1}$	-0.007	0.503	$K^*_{0,1}(\hat{\beta})$	0.00	1.063	0.144	0.546	0.292
$G^{1,1}$	0.588	0.278	$K^*_{1,1}(\hat{\beta})$	0.28	0.937	0.174	0.833	0.202

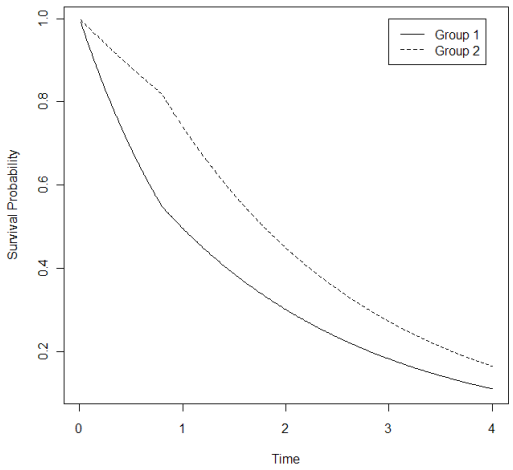
I. $\lambda_1(t) = 2$

$\lambda_2(t) = 1$



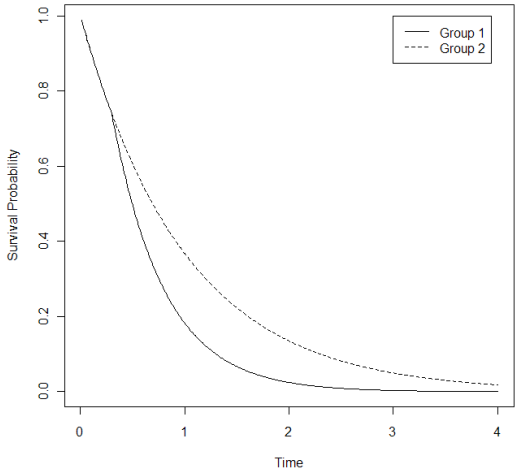
II. $\lambda_1(t) = 0.75\, I\{t < 0.8\} + 0.5\, I\{t \geq 0.8\}$

$\lambda_2(t) = 0.25\, I\{t < 0.8\} + 0.5\, I\{t \geq 0.8\}$



III. $\lambda_1(t) = 1\, I\{t < 0.3\} + 2\, I\{t \geq 0.3\}$

$\lambda_2(t) = 1$



IV. $\lambda_1(t) = 1\, I\{t < 0.4\} + 1.5\, I\{0.4 \leq t \leq 1\}$
$+0.5\, I\{1 \leq t \leq 1.8\} + 1.5\, I\{t \geq 1.8\}$

$\lambda_2(t) = 1\, I\{t < 0.4\} + 0.4\, I\{0.4 \leq t \leq 1\}$
$+1.2\, I\{1 \leq t \leq 1.8\} + 1.5\, I\{t \geq 1.8\}$

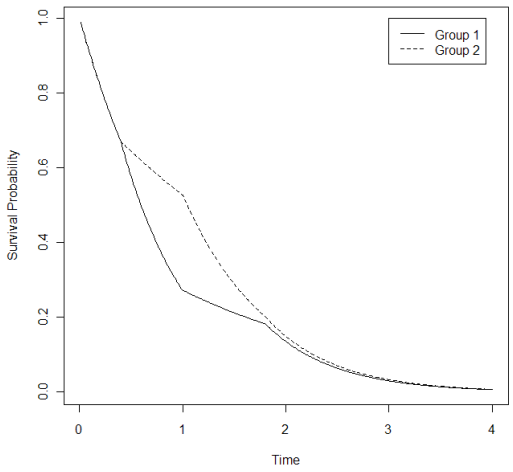


Figure 1. Survival functions for various alternative configurations.

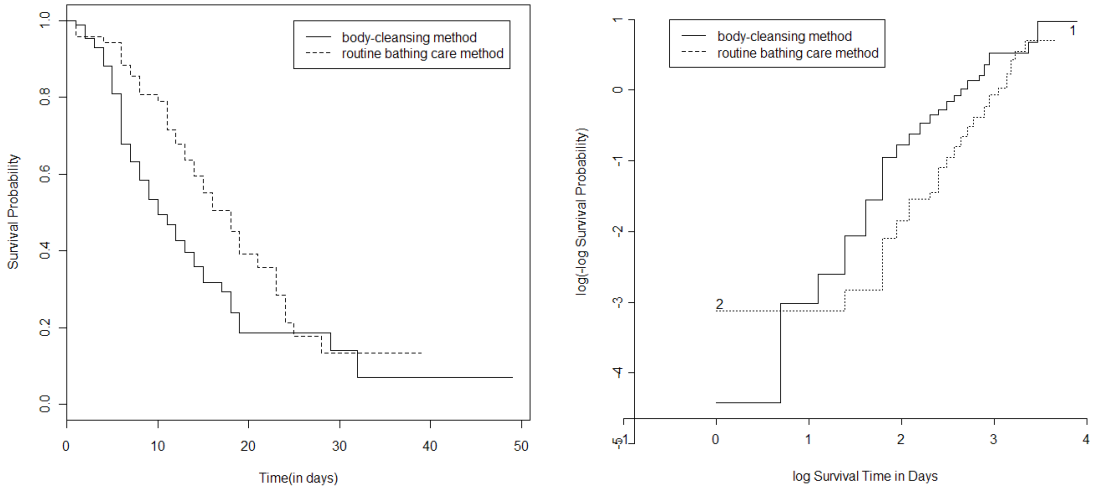


Figure 2. The estimated survival function and log cumulative hazard function for data set 1.

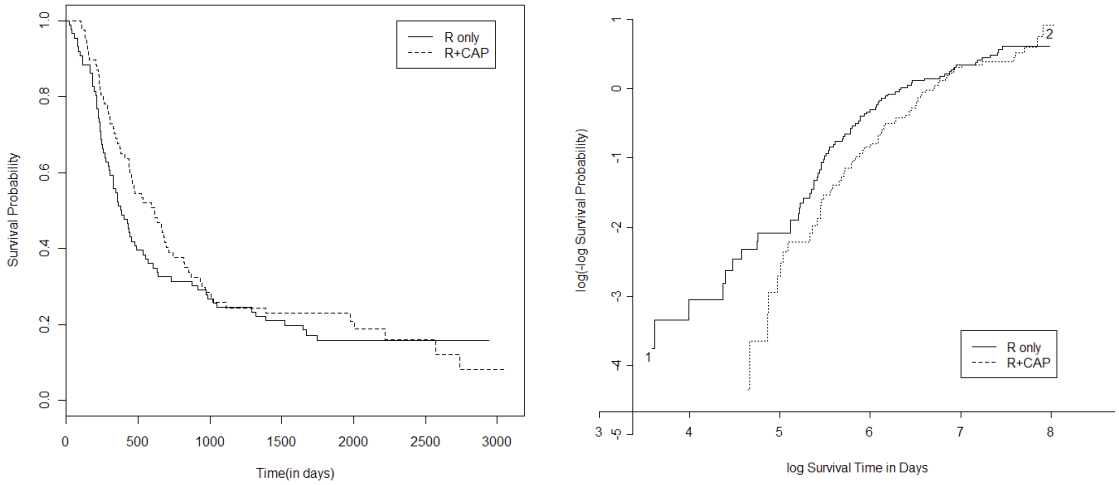


Figure 3. The estimated survival function and log cumulative hazard function for data set 2.