NONPARAMETRIC TESTS FOR COMPARING SURVIVAL DATA WITH NONPROPORTIONAL HAZARDS: EXPLORATION OF A NEW WEIGHT FUNCTION

by

Qing Xu

BMed, HeBei Medical College, China, 1994

Submitted to the Graduate Faculty of

The Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Master of Science

University of Pittsburgh

2005

UNIVERSITY OF PITTSBURGH

Graduate School of Public Health

This thesis was presented

by

Qing Xu

It was defended on

November 16 , 2004

and approved by

Thesis Advisor:
Jong Hyeon Jeong, Ph.D.
Assistant Professor
Department of Biostatistics
Graduate School of Public Health
University of Pittsburgh

Committee Member:
Howard Rochette, Ph.D.
Professor
Department of Biostatistics
Graduate School of Public Health
University of Pittsburgh

Committee Member:
Lawrence Kingsley, Dr. PH
Associate Professor
Department of IDM and Epidemiology
Graduate School of Public Health
University of Pittsburgh

NONPARAMETRIC TESTS FOR COMPARING SURVIVAL DATA WITH NONPROPORTIONAL HAZARDS: EXPLORATION OF A NEW WEIGHT FUNCTION

Qing Xu, MS

University of Pittsburgh, 2005

Abstract

For survival data with nonproportional hazards, the weighted log-rank tests with a proper weighting function are expected to be more sensitive than the simple log-rank statistics for comparing survival data with random effects. A series of simulations were carried out to investigate how much better the weighted log-rank test performs under these situations. The nonproportional hazards data were generated by changing the hazard ratios and piecewise exponential functions. Our Monte Carlo simulation study shows the test with a newly developed weight function has an overall better sensitivity (statistical power) than the simple log-rank test and Harrington-Fleming's weighted log-rank test in detecting the difference between two survival distributions when populations become more homogeneous as time progresses (early difference). For the datasets with middle difference, the test with the new weight function has better sensitivity than that of Harrington-Fleming's weighted log-rank test, similar to that of the simple-log rank test. For late difference, all three tests have similar sensitivity. The new weight function can be used in testing the survival data with nonproportional hazards in public health relevance applications.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

I would like to thank my thesis advisor, Dr. Jong Hyeon Jeong, for his guidance and patience. I learned a lot from him in the past two year, and his excellent view intrigued me in both my research and curriculum study. I would also like to thank the members of my thesis committee, Dr. Howard Rochette, and Dr. Larry Kingsley, for helpful discussion and advice. I would like to give my special thank to Dr. Ada Youk for her help on my study during the past two years at the Graduate School of Public Health.

1. INTRODUCTION

In this chapter, we review briefly important statistical tools often used in survival analysis. We also propose a new weighting function for the weighted log-rank test statistic which will be used in our simulations.

1.1. Survival Data

Survival data usually refers to data in the form of a time from a well-defined time origin until the occurrence of some particular event of interest. In medical research, the time origin will often correspond to the recruitment of an individual into an experimental study, such as a clinical trial to compare two or more treatments. The end point may correspond to the relief of pain, the recurrence of symptoms, or the death of a patient.

A survival model can be used when we want to relate potential prognostic factors or covariates to the length of time to a particular end point (survival time). Often we want to make inferences about the association between the survival time and certain covariates (explanatory variables) rather than estimate a one-sample survival function. Therefore, we often want to compare at least two groups of survival data adjusted for some covariates. For such comparison, the null hypothesis is that there is no difference among survival distributions from different selected comparison groups.

1.2. Proportional Hazards Model

In summarizing survival data, two functions of central interest are the survival function and the hazard function. The survival function, *S(t)*, is defined to be the probability that the survival time is greater than or equal to *t*,

$$S(t) = P(T \geq t), \tag{1}$$

and the hazard function is defined as

$$h(t) = \lim_{\Delta \to 0} \left\{ \frac{P(t \leq T < t + \Delta | T \geq t}{\Delta} \right\}, \tag{2}$$

which is the limiting conditional probability of experiencing an end point immediately after time *t* given the event has not occurred to the individual up to time *t*. (Collett, 2003) The most widely method of estimating the hazard function in the presenceof covariates is the proportional hazards model proposed by Cox (Cox, 1972). The Cox model assumes that the ratio of the hazards between two levels of a covariate (i.e treatment group) is constant over time. It is analytically expressed in the form

$$h_i(t) = e^{\beta x_i} h_0(t), \tag{3}$$

where $h_i(t)$ denotes the hazard function for the i^{th} patient, i=1, 2, ……, n. x_i is the value that the i^{th} patient takes for the explanatory variable $\boldsymbol{X}$. The term $h_0(t)$ is the baseline hazard function. Thus, the null hypothesis that there is no difference in survival distribution between groups corresponds to the null hypothesis β=0 in the model presented in Eq. (3) when i=1, 2, ……, n is a group indicator.

1.3. Simple Log-rank Test

The simple log-rank test (Savage, 1956; Mantel, 1966; Peto, 1972) is perhaps the most widely used method in two-sample comparisons of time-to-event data. It is simple to use, nonparametric in nature, and highly efficient under the proportional hazards assumptions. It incorporates the commonly encountered rightcensorship of survival data without adding complicated elements to the method itself. The log-rank test can be viewed as the score test from the partial likelihood under the Cox model (Cox, 1975)

$$L = \prod_{i \in D} \frac{e^{\beta x_{ii}}}{\sum_{k \in R_i} e^{\beta x_{ki}}},$$

where D represents the total number of failures and R represents the total number of individuals at risk at time of the i^{th} failure. The log rank test can also be derived from the ranks of the survival times in the two groups, with the resulting rank test statistics based on the logarithm of the Nelson-Aalen estimate (Altshuler, 1970; Nelson, 1972; and Aalen, 1978) of the survival function.

When the baseline hazard function $h_0(t)$ is totally unknown, the simple log-rank test is the optimal nonparametric test for testing the null hypothesis $\beta=0$ in the model presented in Eq. (3), for the influence of the explanatory variable x_i on the survival time of individual i. If there is no good reason to doubt the proportional hazards assumption of the survival data, the simple log-rank test should be used to test the hypothesis of equality of two survival distributions. However, if the data show some characteristics of nonproportionality, a weighted log-rank test may serve as a better testing scheme.

1.4. Frailty and Weighted Log-rank Test

In the analysis of survival data, we often encounter the situation where the survival times of a group of individuals are not independent. Such correlations among survival times may arise when different individuals share some feature in common. For example, the survival data from the same clinic may be more similar than those from another clinic. This could be due to different treat teams in different clinics. Such random effects that can cause dependence in survival data are often referred to as frailties.

Frailty in survival data may complicate survival analysis. The efficiency of a test statistic for survival data may decrease if the frailty factor is not considered mainly due to the nonproportionality caused by frailty. (Oak and Jeong, 1998) In addition, failure to include frailty in a test may result in the misspecification of the hazards model. (Oak and Jeong, 1998) Some methods have been proposed to attack this problem. (Aalen, 1998) One can include the random effect in survival modeling by introducing a corresponding term into the proportional hazards model. For example, if we denote an unobserved random effect by a covariate z_i, then Eq. (3) becomes

$$h_i(t) = e^{\beta x_i + \gamma z_i} b_0(t), \tag{4}$$

where $b_0(t)$ is an unknown baseline hazard function. Changing of baseline hazard function from $h_0(t)$ to $b_0(t)$ will not affect our testing results since it is a nonparametric test and the baseline hazard function will cancel when we form a ratio. Comparing Eq. (3) and Eq. (4), we can see that we actually introduced a weighting function $\exp(\gamma z_i)$ to the simple log-rank test in order to model the frailty. An optimal weighting function can be derived if a distribution is assumed for the frailty. (Oakes and Jeong, 1998) For example, if the frailty has a gamma distribution, then the

nonparametric test presented in Eq. (4) with a derived optimal weighting function is equivalent to the G-rho tests proposed by Harrington and Fleming (Harrington and Fleming 1982). When rho=0, the G-rho test reduces to the simple log-rank test. When rho=1, the G-rho test reduces to Wilcoxon test. (Collett, 2003)

Using a weighted log-rank test method is important in order to account for the possible frailty in the data. This is due to the fact that the loss of the efficiency of the test from omitting a covariate is generally more important than the additional loss of the efficiency due to the resulting misspecification of the proportional hazards model. (Jeong & Oakes, 1998)

1.5. Proposed New Weighting Function

For the proportional hazard data with some kinds of frailty, a weighted log-rank test is optimal (Jeong 1998) and is expected to be more sensitive than the simple log-rank test. For example, when frailty has a gamma distribution with an index κ, weighted log-rank test with a weighting function of

$$w(t) = S(t)^{\rho} \tag{5}$$

is still the optimal nonparametric test. These are equivalent to the "G-rho" tests of Harrington and Fleming (Harrington & Fleming, 1994) with rho=1/κ.

However, when the frailty distribution affects the proportionality of the hazard data, the simple log-rank test and weighted log-rank test of G-rho type is no longer the optimal test, a new weighting function must be used. For example, when the frailty follows an inverse Gaussian distribution, an optimal weighting function was derived as

$$w(t) = \frac{1}{2} + \frac{2\psi^2}{[2\psi - \log \hat{S}(t)]^2}, \tag{6}$$

where ψ is an arbitrary controlling parameter which can take the value of 0 to +∞, $\hat{S}(t)$ is the estimated common survival function based on the combined sample up to *t*. (Jeong & Oakes, 1998) This proposed weighting function is used in our simulations and the sensibility in detecting the difference in the simulated survival data is investigated. For observed survival data, the test statistic is given by

$$W = \sum_{i \in D} w(t_i) \left[d_{i1} - Y_{i1} \left(\frac{d_{i1} + d_{i2}}{Y_{i1} + Y_{i2}} \right) \right], \tag{7}$$

where $w(t_i)$ is a common weighting function shared by each group, Y_{i1} and Y_{i2} are the number of objects at risk in group 1 and 2 at time t_i, d_{i1} and d_{i2} are the number of events occurred in each group at time t_i, respectively. The summation is over *D*, which includes a subset of survival times that are observed as event of interest. The variance of *W* can be estimated by

$$V = \sum_{i \in D} w(t_i)^2 \frac{Y_{i1}}{Y_i} \left(1 - \frac{Y_{i1}}{Y_i} \right) \left(\frac{Y_i - d_i}{Y_i - 1} \right) d_i . \tag{8}$$

It was proved that $Z = \frac{W}{\sqrt{V}}$ has an asymptotic standard normal distribution if the dataset is big enough (Harrington & Fleming, 1982).

The common survival function estimator $\hat{S}(t)$ in Eq. (6) is given by

$$\hat{S}(t) = \prod_{t \le t_i} \left(1 - \frac{d_i}{Y_i} \right). \tag{9}$$

A frailty distribution is usually unobservable, thus we do not know if the frailty itself will affect proportionality of the survival data at hand. So we must test if the observed survival data still follows proportional hazards assumption before we decide what type of weighting function

should be used (Therneau and Grambsch, 2000). Testing the proportionality in survival data can be performed by using the cox.zph procedure provided in S-Plus. If the cox.zph test indicates proportionality in the data, the log-rank test statistics like simple log-rank test and Wilcoxon log-rank test can be chosen. However, if the cox.zph test shows that the dataset does not satisfy the proportional hazards assumption, we should use a log-rank test with a different type of weighting function, such as the one for the survival data with inverse Gaussian frailty or the Harrington-Fleming test.

2. SIMULATION METHODS

In this chapter, we describe our simulation procedure, i. e., how we generate the survival data and how we performed the simulation.

2.1. Methodology

Suppose there are two groups of survival data with corresponding hazard functions $h_2(t)$ and $h_1(t)$. Survival data can be generated by Monte Carlo method according to the characteristics of h_1 and h_2. Then we can use log-rank tests with different weight functions to determine the power of each test method in differentiating these two groups of data. An estimate of the statistical power of the test is provided by

$$power = \frac{m}{n_s}, \tag{10}$$

where m is the number of simulations in which the test can differentiate the data with significance, and n_s is the total number of simulations.

In our simulation, we take $h_1(t)$ as the baseline hazards function and set it to be a constant, ρ. Therefore h_2 becomes

$$h_2(t|z) = e^{\beta z} h_1(t) = e^{\beta z} \rho . \tag{11}$$

Here z is a covariate. The null hypothesis that h_1 and h_2 are identical corresponds to β=0. The survival functions become

$$S_1(t) = e^{-\rho t}, \tag{12}$$

and

$$S_2(t|z) = S_1(t)^{e^{\beta z}} . \tag{13}$$

Then for an object in group one, the probability that its survival time is less than value *t* is

$$F_1(t) = 1 - S_1(t) = 1 - e^{-\rho t} , \tag{14}$$

and likewise, for an object in group two,

$$F_2(t|z) = 1 - S_2(t) = 1 - e^{-(\rho t)e^{\beta z}} . \tag{15}$$

Since that $F_1(t)$ and $F_2(t|z)$ conform to a uniform distribution in the range of [0,1], we have

$$F_1(t) = 1 - e^{-\rho t} = u , \tag{16}$$

and

$$F_2(t) = 1 - e^{-(\rho t)e^{\beta}} = u , \tag{17}$$

as z=1 for group two data.

For group one survival data, we obtain

$$t_1 = -\frac{\ln(1-u)}{\rho} . \tag{18}$$

For group two survival data, we obtain

$$t_2 = -\frac{\ln(1-u)}{\rho e^{\beta}} = e^{-\beta} t_1 . \tag{19}$$

Therefore, we can generate two groups of survival data conforming to h_1 and h_2 in Eq. (11) by starting from a uniform distribution u, and using the relationships represented in Eqs. (18) and (19).

2.2. Procedure Details

The data generation procedures are as follows:

i. Generate N^* observations from uniform distribution $u(0,1)$, designate them as u_i, $i=1,2,\ldots N^*$.

ii. Generate the survival times for group one data, t_{1i}, $i=1,2,\ldots N^*$, base on Eq. (18). The parameter ρ in Eq. (18) is set arbitrarily; here we set it to be in [0.001, 0.1, 0.3]. The data are sorted ascendingly.

$$t_{1i} = -\frac{\ln(1-u_i)}{\rho}.$$

iii. Generate the survival times for group two data, t_{2i}. The random effect (frailty) of the survival times is substituted into the t_{2i} by multiplying the factor of $e^{-\beta}$ by t_{1i}, as shown in Eq. (19). In our simulations, we let the premultiplier $e^{-\beta}$ be in the range of (0, 1). The values of $e^{-\beta}$ were chosen according to the shapes of hazards ratios of interest. For example, if we are interested in two groups of data with early difference, we let the $e^{-\beta}$ take values of piece-wise proportionality reflecting early departures.

iv. Generate the censored data in two groups from a uniform distribution randomly. In this study we let the censoring occur randomly.

In testing the proportionality of the simulated data, we used the `cox.zph` function in the S-Plus package. The null hypothesis for this test is that the data obey the assumption of proportional hazards.

For the two sample test, we demonstrate how the test statistics can be evaluated step by step in Table 4, Appendix. We found it was very difficult to incorporate the new weighting function in Eq. (6) into the `survdiff` procedure in the S-Plus package. Therefore, we wrote our own program in S-Plus to evaluate Eq. (7) and (8). The survival data obtained in Section 2.1 were transformed accordingly in order to calculate the quantities in Eq. (7) and (8) numerically. The p-values correspond to the observed statistic, $Z = \frac{W}{\sqrt{V}}$, which follow the standard normal distribution. The test results based on the statistics in Eq. (7) and (8) with the new weight function Eq. (6) are compared with a simple log-rank test and Harrington-Fleming's weighted log-rank test.

3. SIMULATION RESULTS

In this chapter we present our simulation results. First we have used `cox.zph` function to test the proportional hazards assumption of the simulated data. Then we tested the hypothesis that two survival distributions are the same to evaluate the power of the three test methods, i.e., the simple log-rank test, the weighted log-rank test proposed by Harrington and Fleming, and the weighted log-rank test with the new weighting function shown in Eq. (6).

3.1. Test for Proportional Hazard Assumptions

The purpose of this work is to investigate how the simple log-rank test and the weighted log-rank test of Harrington-Fleming perform for the nonproportional hazards data, compared with the test with the new weight function in Eq. (6). Thus, first it is worthwhile to evaluate how significantly the simulated data violate the proportional assumption.

We used the `cox.zph` function in the S-Plus package to test the statistical significance of violation of the proportional hazards assumption in the simulated data. We tested simulated data with early, middle, and late departure. The datasets were generated by the procedures described in 2.2 with baseline hazard function ρ=0.3. The specific parameters for the tests can be found in Figure 1-3. Our results show that `cox.zph` tests do identify the nonproportionality existing in our simulated data. However, the level of nonproportionality detected by `cox.zph` varies from dataset to dataset. We examined simulated datasets with a data size of 1500 and found that a large fraction of datasets detected by the `cox.zph` procedure to be nonproportional at the

significance level of 0.05. For example, for late difference datasets, 32% of 100 simulations are identified as nonproportional using the `cox.zph` procedure. For early difference datasets, about 94% of the datasets were identified as nonproportional. For middle difference, 57% of datasets were shown to be nonproportional.

In Figure 1 to 3 we show some of the estimated patterns of change of the hazard ratios from the `cox.zph` function (smoothed scaled Schoenfeld residual plots), together with the corresponding Kaplan-Meier plots. These figures show that nonproportionality of the simulated data. For example, Figure 1 and 2 shows the `cox.zph` test for two survival datasets with a data size of 1500 that have later difference. The p-value of the `cox.zph` test is 0.0406. It can be seen from the residual plots that the drifting of the residual curve from zero when time progresses indicates significant late difference for this particular data group. Two similar example plots for early difference and middle difference datasets are shown in Figure 2 and 3. In Figure 4 to 6 we show some other examples of simulated nonproportional data with a much smaller data size of 100. These examples provide visual evidence that significant nonproportionality exists in the data.

3.2. The Parameter Choices

The simple log-rank and Harrington-Fleming's weighted log-rank test can be formulated by properly setting the weight function $w(t_i)$ in Eqs (7) and (8). For the simple log-rank test, $w(t_i)$ simply equals to unity for all t, which means all the failure times are treated with equal weight. For Harrington-Fleming's weighted log-rank test, the weight function $w(t_i)$ in Eqs. (7) and (8) is defined by

$$w_{p,q}(t_i) = \hat{S}(t_{i-1})^p [1 - \hat{S}(t_{i-1})]^q , \quad (20)$$

where $p \geq 0$, $q \geq 0$. (Klein and Moeschberger, 1997) Slightly different from the $\hat{S}(t_i)$ in Eq. (9), the $\hat{S}(t_{i-1})$ in Eq. (20) is the survival function at the previous failure time. When p=q=0 Eq. (20) reduces to the weigh function for the simple log-rank test. When p=1 and q=0 we have the G-rho test. By choosing the values of p and q properly, we assign different weights to the data points. For example, when q=0 and p>0, the tests with this weight function are more sensitive to early difference. When p=0 and q>0, the tests are more sensitive to late difference. In our simulation tests, we set the values of p and q in the Harrington-Fleming test as follows:

Early difference data: p=1, q=0;

Late difference data: p=0, q=1;

Middle difference data: p=q=1.

For the log-rank test with the new weight function, we need to set the parameter Ψ for the new weight function in Eq. (6). In this study we used three values for Ψ, 0.01, 1.0, and 5.0. For the baseline hazard function, we used the values of 0.001, 0.1 and 0.3.

3.3. Simulation Results

We chose the dataset size of the simulation to be 1500, with dataset 1 and dataset 2 having the same size. That is, $n_1=n_2=N^*$. The data points are sorted ascendingly according to t and divided into 10 subgroups. For each subgroup a factor is multiplied by t to create the early, middle, or late difference between data group one and two. For example, when we model early difference data, we set a factor array of (0.85, 0.80, 0.75, 0.80, 0.85, 0.90, 0.93, 0.95, 0.97, 0.99). The first subgroup of 150 survival times in group two equal to the product of 0.80 and the first 150 data from group one. The second subgroup of 150 survival times in group two equal to the product of 0.825 and the first 150 data from group one, and so on. The last subgroup of 150 survival time in group one and two are essentially the same. For each choice of factor array, n_s=1000 simulations with randomly generated survival times were performed, with level of test equal to 0.05. The power of each test was computed according to Eq. (10).

One set of results for comparing the simple log-rank, Harrington-Fleming's weighted log-rank, and the new weight function tests are shown in Table 1. This set of data has early difference. As can be seen from the simulation results, the Harrington-Fleming's weighted log-rank test fails to capture the difference in two survival data. The simple log-rank test has shown a much higher sensitivity than the Harrington-Fleming's test, giving an average power of about 0.22. In contrast, the test with new weight function has about twice the power of a simple log-rank when the parameter Ψ equals to the value of 1.0. We should point out that all three test methods show a low power (less than 0.5) in differentiating the data group mainly due to the very small difference we design in the simulated datasets themselves. The power of Harrington-

Fleming's test is negligibly low. However, this does not indicate that Harrington-Fleming fails completely. If we increase the data difference by changing the premultiplier factor array, the testing powers for all three tests increase rapidly, but the power of the new weight function test remains the highest before they reach unity. Also we found that variation in the constant baseline hazards function does change the relative sensitivity of these three test methods in differentiating the data groups with early departure.

For the datasets with middle difference, the test with the new weight function shows a much higher power than the Harrington-Fleming test. However, its powers are in the same range as that of the simple log-rank test. From Table 2, we can see the new test has a slightly higher power than simple log-rank test when we choose Ψ to be 1.0. This fact is similar to that we have seen for early difference test. It seems that Ψ = 1.0 is good choice for testing the early and middle difference survival data using the new weight function. Theoretically, the value of Ψ can be any arbitrary positive number between 0 and infinity. (Oakes and Jeong, 1998) As long as a positive Ψ is chosen, the weighting function will be always between 0 and 1. However, the change in Ψ value will change the distribution of the weighting function. Therefore, further careful work needs to be done before we can give a reasonable rule in choosing the optimal Ψ value. Again, the testing results have only a minor change when we vary the baseline function values. This indicates that the value of the constant baseline hazards function has negligible effect on the sensitivity of these testing methods in light of the simulation fluctuations.

Simulation results from the late difference data are shown in Table 3. We can see that the tests with all three methods have a power in the same range.

We note that the factor arrays for the simulated data with different departure pattern (early, middle, or late difference) are in the same magnitude, ranging from 0.80 to 0.99. However, the

significance of nonproportionality in these datasets identified by `cox.zph` varies. Early difference data was identified by `cox.zph` as having the most significant nonproportionality; in turn the new weigh function is more powerful than the simple log-rank and Harrington-Fleming's weighted log-rank tests in differentiating the early difference data.

Summarizing all the results shown in Table 1 to 3, we can see that the advantage of the new test is obvious. For early difference data, the new method shows better performance than either simple log-rank or Harrington-Fleming's method. For middle difference, it is better than the Harrington-Fleming's test. Even for the later difference, which has the least nonproportionality, the new weight function method has a similar sensitivity. This indicates that the new weighting function is successful in properly accounting for the nonproportional frailty effect of the simulated survival data.

Table 1. Monte Carlo estimate of the power of the simple log-rank, Harrington-Fleming's weighted log-rank test, and the new weighted logrank test. Hypothesis: β=0; 1000 simulations for each N^*; $n_1=n_2=N^*$, level of test=0.05. The datasets have early difference, the piecewise nonproportional hazard array is (0.85, 0.80, 0.75, 0.80, 0.85, 0.90, 0.93, 0.95, 0.97, 0.99).

baseline hazards function	**Ψ**	**Test with new weight function, P_1**	**Simple log-rank, P_2**	**Harrington-Fleming weighted log-rank, P_3**	**P_2/P_1**	**P_3/P_1**
0.001	5.0	0.285	0.226	0.05	0.793	0.007
	1.0	0.445			0.508	0.004
	0.01	0.229			0.987	0.009
0.1	5.0	0.297	0.237	0.063	0.798	0.003
	1.0	0.446			0.531	0.002
	0.01	0.239			0.992	0.004
0.3	5.0	0.312	0.228	0.057	0.731	0.003
	1.0	0.439			0.519	0.002
	0.01	0.214			1.065	0.004

Table 2. Monte Carlo estimate of the power of the the simple log-rank, Harrington-Fleming's weighted log-rank test, and the new weighted logrank test. Hypothesis: β=0; 1000 simulations for each N^*; $n_1=n_2=N^*$, level of test=0.05. The datasets have middle difference, the nonproportional hazard array is (0.99, 0.95, 0.90, 0.85, 0.80, 0.80, 0.85, 0.90, 0.95, 0.99).

baseline hazards function	Ψ	**Test with new weight function, P_1**	**Simple log-rank, P_2**	**Harrington-Fleming weighted log-rank, P_3**	P_2/P_1	P_3/P_1
0.001	5.0	0.890	0.834	0.064	0.937	0.072
	1.0	0.945			0.883	0.068
	0.01	0.849			0.982	0.075
0.1	5.0	0.915	0.832	0.057	0.909	0.062
	1.0	0.947			0.879	0.060
	0.01	0.818			1.017	0.070
0.3	5.0	0.887	0.834	0.055	0.940	0.062
	1.0	0.957			0.871	0.057
	0.01	0.835			0.999	0.066

Table 3. Monte Carlo estimate of the power of the simple log-rank, Harrington-Fleming's weighted log-rank test, and the new weighted logrank test. Hypothesis: β=0; 1000 simulations for each N^*; $n_1=n_2=N^*$, level of test=0.05. The datasets have late difference, the nonproportional hazard array is (1.00, 0.99, 0.98, 0.97, 0.95, 0.93, 0.91, 0.89, 0.87, 0.85).

baseline hazards function	Ψ	**Test with new weight function, P_1**	**Simple log-rank, P_2**	**Harrington-Fleming weighted log-rank, P_3**	P_2/P_1	P_3/P_1
0.001	5.0	0.977	0.986	0.988	1.009	1.011
	1.0	0.977			1.009	1.011
	0.01	0.981			1.005	1.007
0.1	5.0	0.978	0.985	0.984	1.007	1.006
	1.0	0.967			1.019	1.018
	0.01	0.987			0.998	0.997
0.3	5.0	0.983	0.986	0.981	1.003	0.998
	1.0	0.972			1.014	1.009
	0.01	0.983			1.003	0.998

4. CONCLUSION

We studied the sensitivity of a newly developed weighted log-rank test, and compared it with the simple log-rank test and Harrington-Fleming's weighted log-rank test, in testing treatment with nonproportional survival data using Monte Carlo simulations. We found that the new test shows a better sensitivity in capturing the difference between the data group when the survival data has significant nonproportionality (here the data with early difference). For the datasets with less nonproportionality (here the data with middle difference), the test with the new weight function has better sensitivity than that of Harrington-Fleming's weighted log-rank test, similar to that of the simple-log rank test. For late difference which has least nonproportionality, all three tests have similar sensitivity.

5. FUTURE WORK

In the present study, we only chose three values arbitrarily for the parameter in the new weight function and tested its sensitivity. In future work, a more systematic study will be performed so that we can provide a rule of thumb in selecting the proper value for the Ψ parameter according to the survival data pattern.

We only did simulation with a sample size of 1500. Such a large data size may disguise some of the problem in the test model. For example, the abundance of data points may compensate the inaccuracy in the specification of the survival model and gave an incorrect conclusion that a particular model is effective in capturing the data difference. In the next step, we may continue the simulation to determine the impact of sample size on the performance of the new weight function by conducting tests with various sample size.

The purpose of this study is to use the new weight function in analyzing the survival data in public health applications. We will apply the method developed in this project to study the real data collected in cancer survival studies.

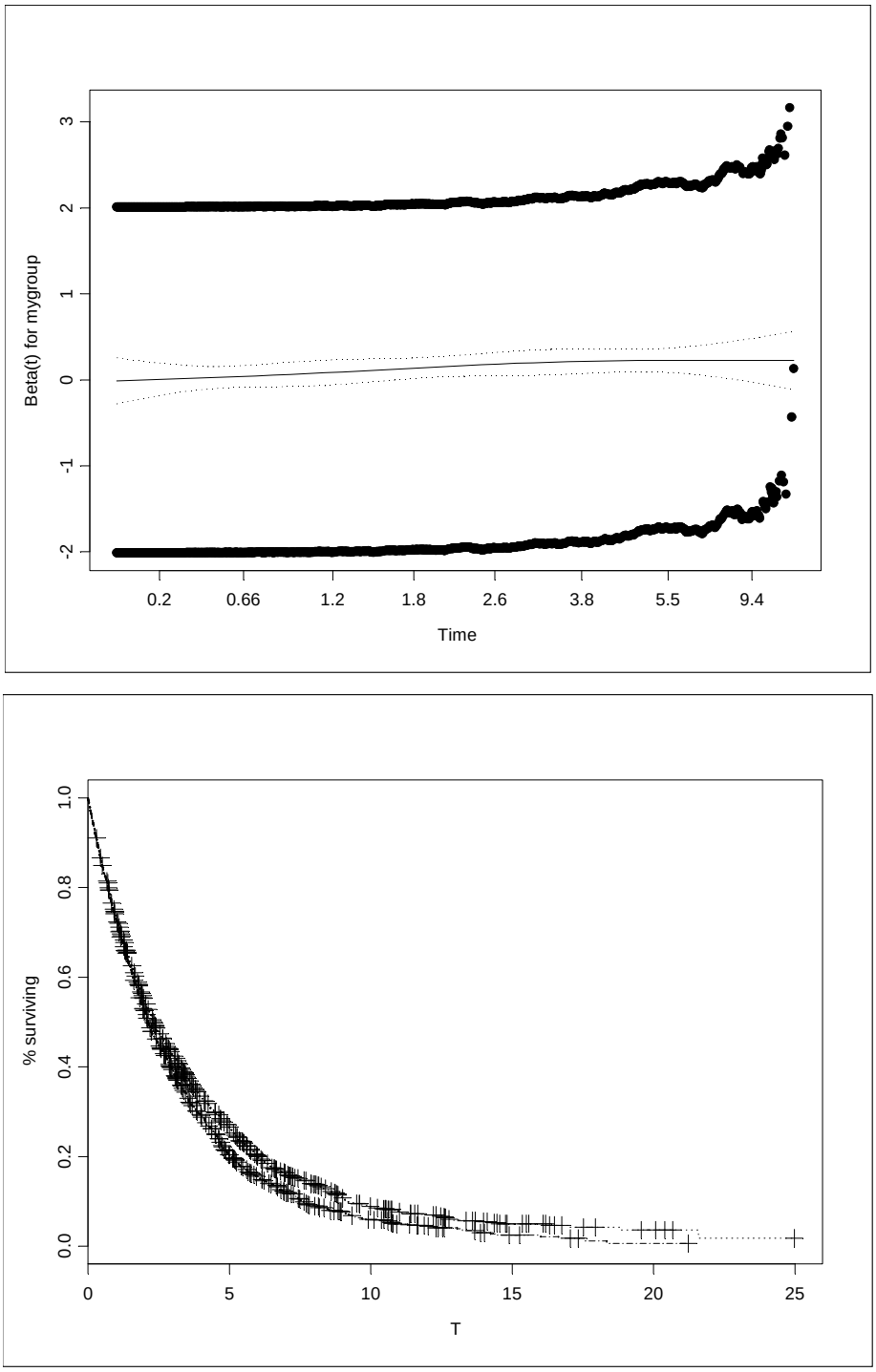

Figure 1. Test of nonproportional hazards in a simulated dataset. The size of datasets $n_1=n_2=N^*=1500$. Late difference is implanted in the data, with the nonproportional hazard array (1.00, 0.99, 0.98, 0.97, 0.95, 0.93, 0.91, 0.89, 0.87, 0.85). The cox.zph test of the data shows that the data have nonproportional hazards with a p-value of 0.0406.

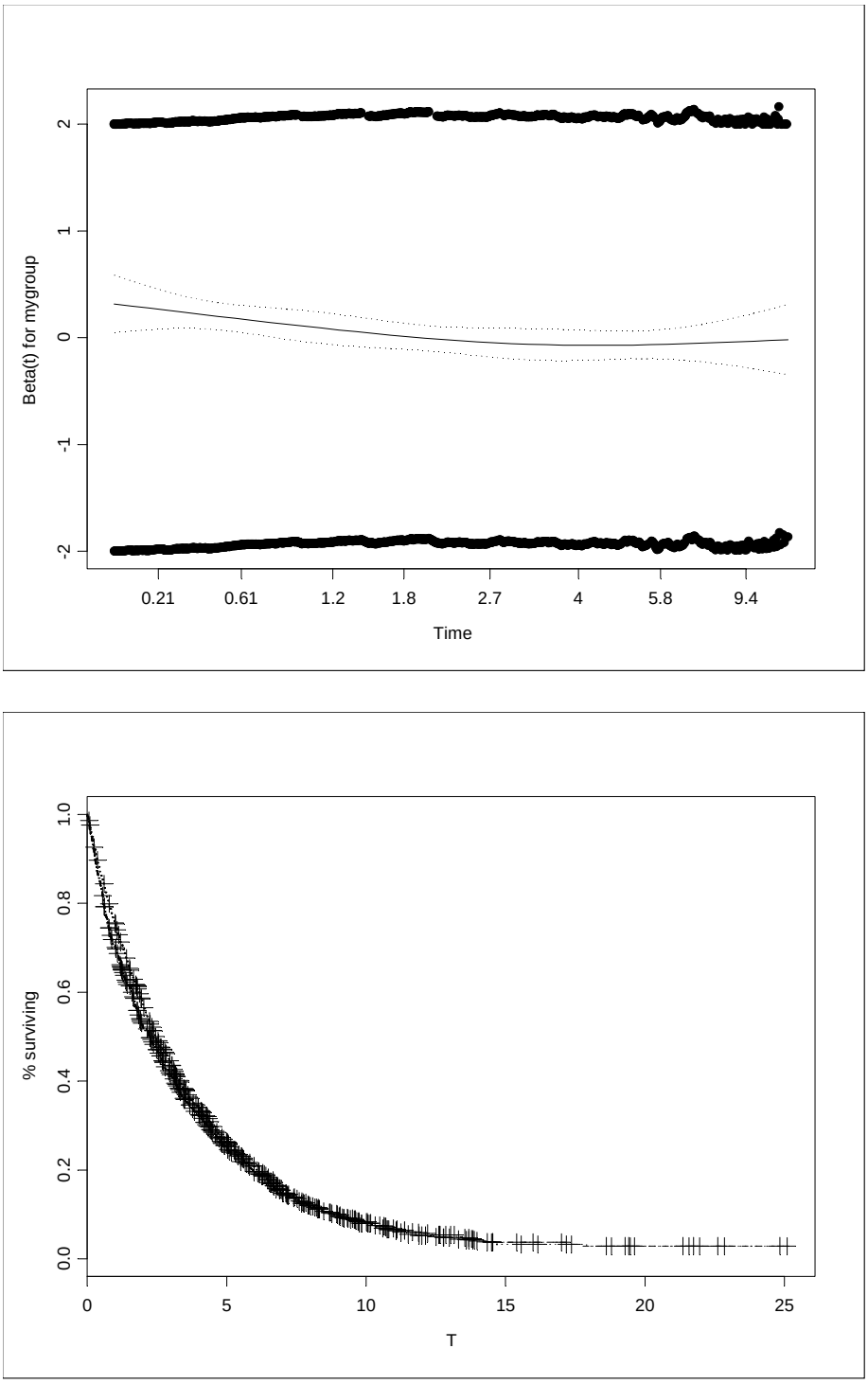

Figure 2. Test of nonproportional hazards in a simulated dataset. The size of datasets $n_1=n_2=N^*=1500$. Early difference is implanted in the data, with the nonproportional hazard array (0.85, 0.80, 0.75, 0.80, 0.85, 0.90, 0.93, 0.95, 0.97, 0.99). The cox.zph test of the data shows that the data have nonproportional hazards with a p-value of 0.00466.

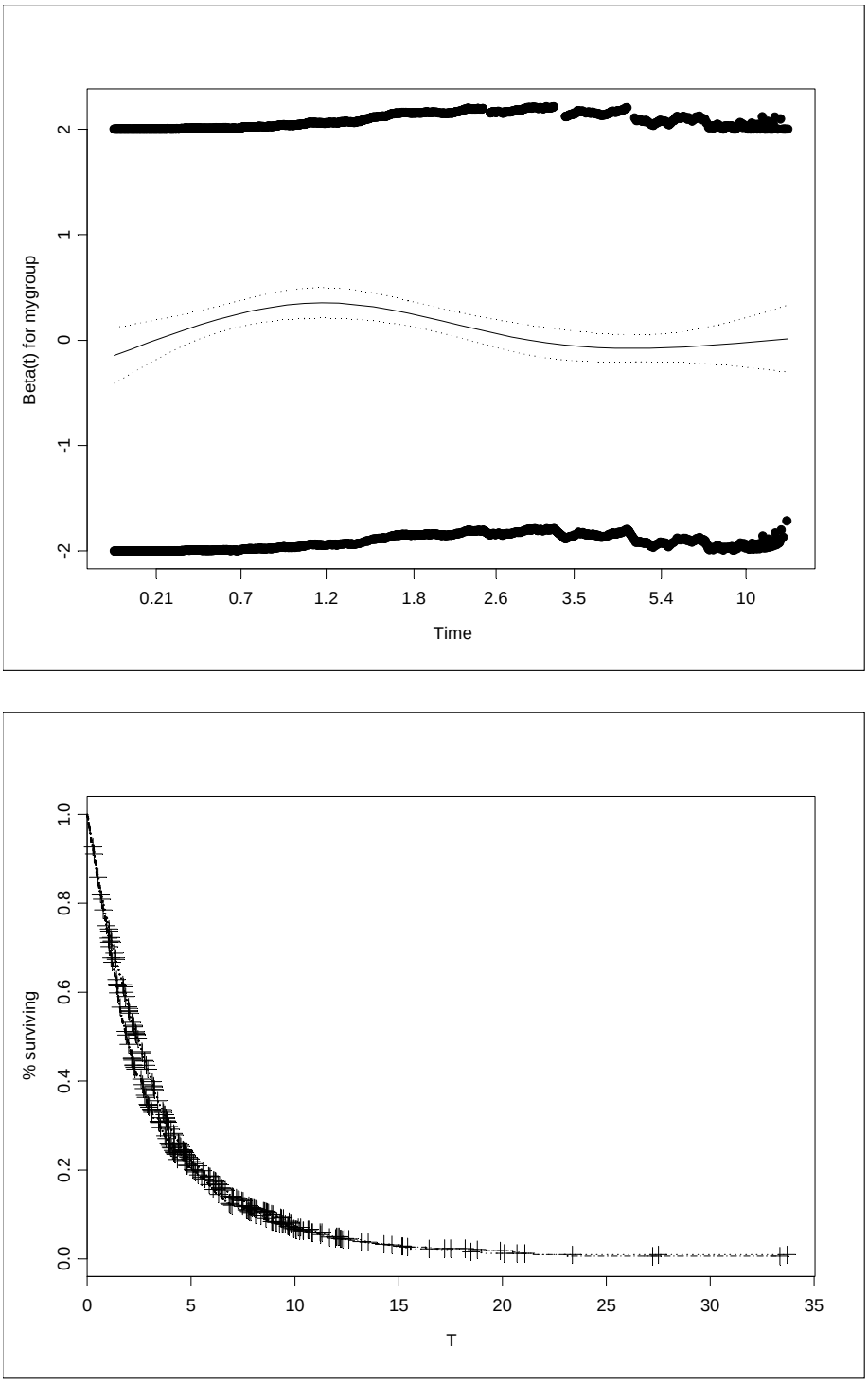

Figure 3. Test of nonproportional hazards in a simulated dataset. The size of datasets $n_1=n_2=N^*=1500$. Middle difference is implanted in the data, with the nonproportional hazard array (0.99, 0.95, 0.90, 0.85, 0.80, 0.80, 0.85, 0.90, 0.95, 0.99). The cox.zph test of the data shows that the data have nonproportional hazards with a p-value of 0.0374.

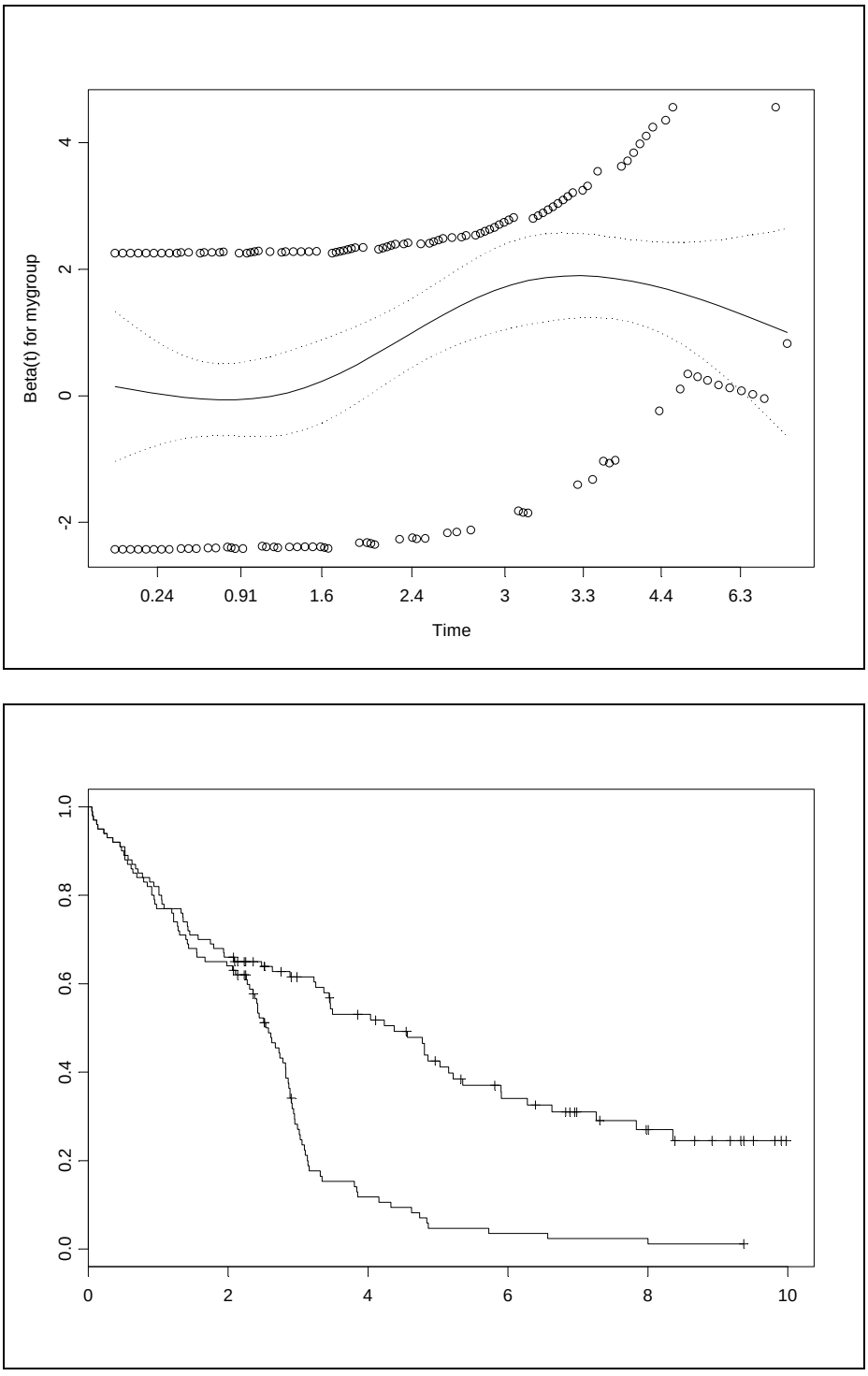

Figure 4. Example of nonproportional hazards simulated dataset with a small data size. The size of datasets $n_1=n_2=N^*=100$. Late difference is implanted in the data.

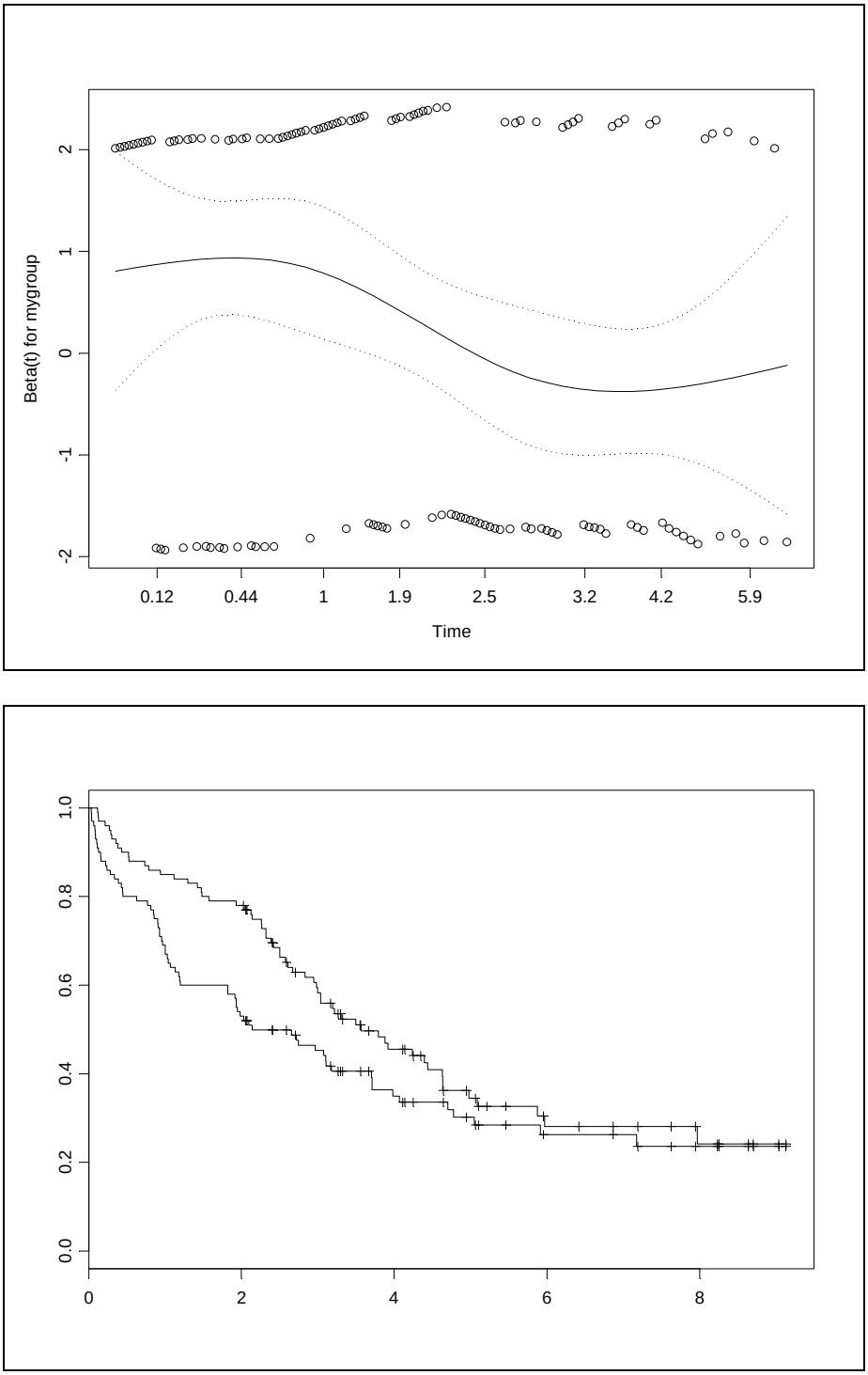

Figure 5. Example of nonproportional hazards simulated dataset with a small data size. The size of datasets $n_1=n_2=N^*=100$. Early difference is implanted in the data.

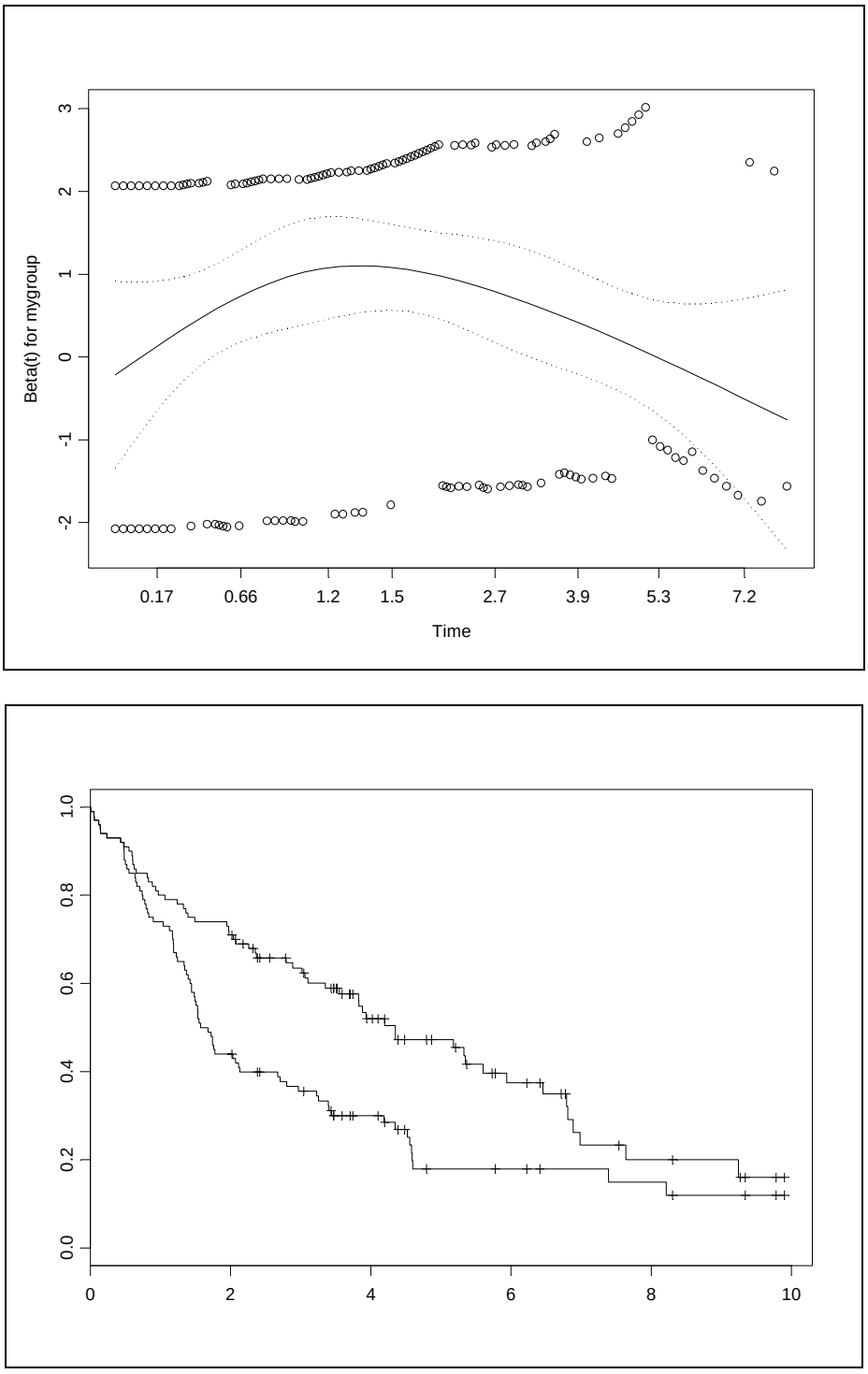

Figure 6. Example of nonproportional hazards simulated dataset with a small data size. The size of datasets $n_1=n_2=N^*=100$. Middle difference is implanted in the data.

APPENDIX: A Sample Simulation Table

Table 4. An example of two sample test table. The table is constructed based on two groups of sorted survival data, t_{i1} and t_{i2} with early difference, using Eqs. (7) and (8).

t_i	Y_{i1}	d_{i1}	Y_{i2}	d_{i2}	Y_i	d_i	$Y_{i1}\left(\frac{d_i}{Y_i}\right)$	$d_{i1}-Y_{i1}\left(\frac{d_i}{Y_i}\right)$	$\frac{Y_{i1}}{Y_i}\left(1-\frac{Y_{i1}}{Y_i}\right)\left(\frac{Y_i-d_i}{Y_i-1}\right)d_i$	$\hat{S}(t_i)$
0.0	100	14	100	21	200	35	17.500	-3.500	7.255	1.000
1.0	86	18	79	13	165	31	16.158	1.842	6.321	0.813
2.0	68	7	66	12	134	19	9.641	-2.642	4.106	0.699
3.0	61	8	54	5	115	13	6.896	1.104	2.897	0.620
4.0	51	9	47	6	98	15	7.806	1.194	3.203	0.526
5.0	40	6	40	8	80	14	7.000	-1.000	2.924	0.435
6.0	34	4	32	0	66	4	2.060	1.939	0.953	0.409
7.0	30	3	30	2	60	5	2.500	0.500	1.165	0.376
8.0	27	2	27	5	54	7	3.500	-1.500	1.552	0.328
9.0	22	3	21	3	43	6	3.070	-0.070	1.321	0.283
10.0	15	2	15	1	30	3	1.500	0.500	0.698	0.256
11.0	13	1	13	1	26	2	1.000	0.000	0.480	0.237
12.0	10	0	10	1	20	1	0.500	-0.500	0.250	0.226
13.0	9	1	7	0	16	1	0.562	0.438	0.246	0.212
14.0	7	0	6	0	13	0	0.000	0.000	0.000	0.212
15.0	6	0	6	0	12	0	0.000	0.000	0.000	0.212
16.0	6	0	6	1	12	1	0.500	-0.500	0.250	0.196
17.0	4	0	4	0	8	0	0.000	0.000	0.000	0.196
18.0	4	0	4	0	8	0	0.000	0.000	0.000	0.196
19.0	4	0	4	1	8	1	0.500	-0.500	0.250	0.174
20.0	4	1	3	0	7	1	0.571	0.429	0.245	0.152
21.0	3	0	3	0	6	0	0.000	0.000	0.000	0.152
22.0	3	0	2	0	5	0	0.000	0.000	0.000	0.152
23.0	1	0	1	0	2	0	0.000	0.000	0.000	0.152
24.0	1	0	1	0	2	0	0.000	0.000	0.000	0.152
sum		79		80		159	81.264	-2.266	34.116	

BIBLIOGRAPHY

Aalen, O. O. (1978). Nonparametric inference for a family of counting processes. *Annals of Statistics*, **6**, 534-545.

Aalen, O. O. (1998). Frailty models. In *Statistical Analysis of Medical Data: New Developments.* (eds. B. S. Everitt & G. Dunn). Arnold: London.

Altshuler, B. (1970). Theory for the measurement of competing risks in animal experiments. *Mathematical Biosciences*, **6**, 67-77.

Collett, D. (2003). *Modelling Survival Data in Medical Research*. Second Edition. Chapman & Hall/CRC: Boca Raton.

Cox, D. R. (1972). Regression models and life tables. *Journal of the Royal Statistical Society*, *B*, **34**, 187-220.

Cox, D. R. (1975). Partial likelihood. *Biometrika*, **62**, 269-276.

DiRienzo, A. G. (2003). Nonparametric comparison of two survival-time distributions in the presence of dependent censoring. *Biometrics*, **59**, 497-504.

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

Harrington, D. P. & Fleming T. R. (1982). A class of rank test procedures for censored survival data. *Biometrika*, **69**, 553-566.

Hougaard, P. (1995). Frailty models for survival data. *Lifetime Data Analysis*, **1**, 255-273.

Hougaard, P. (2000). *Analysis of Multivariate Survival Data.* Springer, New York.

Kaplan, E. L. & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, **53**, 457-481.

Klein, J. P. & Moeschberger, M. L. (1997) *Survival Analysis*. Springer.

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

Mantel, N. & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, **22**, 719-748.

Nelson, W. (1972). Theory and applications of hazard plotting for censored failure data. *Technometrics*, **14**, 945-965.

Oakes, D. & Jeong, J. H. (1998). Frailty model and rank tests. *Lifetime Data Analysis*, **4**, 209-228.

Oakes, D. (1977). The asymptotic information in survival data. *Biometrika*, **64**, 441-448.

Peto, R. (1972). Contribution to the discussion of a paper by D. R. Cox. *Journal of the Royal Statistical Society, B,* **34**, 205-207.

Peto, R. & Peto, J. (1972). Asymptotically efficient rank invariant procedures. *Journal of the Royal Statistical Society, A,* **35**, 185-207.

Therneau, T. M. & Grambsch P. M. (2000). *Modelling Survival Data: Extending the Cox Model*. Springer: New York.