Combined Multiple Testing
by Censored Empirical Likelihood

Arne Bathke [*], Mi-Ok Kim [†], and Mai Zhou [‡]

Abstract

We propose a new procedure for combining multiple tests in samples of right-censored observations. The new method is based on multiple constrained censored empirical likelihood where the constraints are formulated as linear functionals of the cumulative hazard functions. We prove a version of Wilks' theorem for the multiple constrained censored empirical likelihood ratio, which provides a simple reference distribution for the test statistic of our proposed method. A useful application of the proposed method is found in examining the survival experience of one or more populations by combining different weighted log-rank tests. A real data example is given using the log-rank and Gehan-Wilcoxon tests. In a simulation study, we compare the new method to different weighted log-rank statistics, Renyi-type suprema, and maximin efficiency robust tests. The empirical results demonstrate that, in addition to its computational simplicity, the proposed combined testing method can also be more powerful than previously developed procedures. Statistical software is available in an R package 'emplik'.

Key words: Multiple constraints; Survival analysis; Weighted log-rank test; Wilks theorem.

Mathematics Subject Classification 2000: Primary 62G10, 62G20, 62N03 secondary 62G05, 62N02, 62P10.

[*] Arne Bathke, Assistant Professor, Department of Statistics, University of Kentucky, email: arne@ms.uky.edu

[†] Mi-Ok Kim, Assistant Professor, Department of Statistics and College of Public Health, University of Kentucky, email: miokkim@ms.uky.edu

[‡] Mai Zhou (Corresponding Author), Associate Professor, Department of Statistics, University of Kentucky, email: mai@ms.uky.edu

1 Introduction

In studies examining the survival experience of one or more populations, one has a choice among many different tests that are versions of weighted log-rank statistic, differing only in the choice of the weight function. If the shape of the hazard ratio under the alternative hypothesis is known, a test with an optimal weight function can be selected. For example, the log-rank test is most powerful when the true hazard curve is proportional to the hypothesized one. However, in general the shape is unknown and the selection of the weights is problematic as the power of the weighted tests varies depending on where and how the curves depart from the hypothesized one (see, e.g., Fleming and Harrington, 1991, Chapter 7; Lin and Kosorok, 1999; Letón and Zuluaga, 2005). A wrong choice may result in a great loss of power. Kosorok and Lin (1999) observe in the β-Blocker Heart Attack Trial (BHAT) that the beneficial effect of propranolol hydrochloride in patients with acute myocardial infarction can be detected with an optimally weighted test at a much earlier calendar time than with the log-rank statistic that was originally chosen by the investigators (β-Blocker Heart Attack Trial Research Group, 1982). Naturally, the differing powers of the tests can lead to disagreement. Klein and Moeschberger (1997, p. 197) use a kidney dialysis data set to illustrate the situation. We re-analyze the kidney dialysis data, using our newly proposed method (see Section 3).

Several versatile procedures sensitive to a range of alternatives have been developed. Among others, these include the maximin efficiency robust tests (MERT; Gastwirth, 1985), the supremum version tests (Fleming, Harrington, and O'Sullivan, 1987), the maximum of a finite cluster of statistics (Tarone, 1981; Fleming and Harrington, 1984; Lee, 1996), and a test with function indexed scheme of the weights and time (Kosorok and Lin, 1999). These methods do not yield asymptotically efficient tests. Lai and Ying (1991) proposed an asymptotically efficient test by estimating efficient weights. However, their method uses kernel estimates. Therefore, it requires large sample sizes to perform well, and it is inappropriate for small or moderate samples.

In this paper we take a different approach; we propose increasing the power by combining different tests. While different versions of weighted log-rank tests are available, practitioners are most familiar with the log-rank and Mann-Whitney-Wilcoxon test or its modified versions. The aforementioned versatile procedures that are sensitive to a range of alternatives have the disadvantage of being complex and computationally intensive as the null distribution of the test statistics often needs to be simulated. Hence, in practice, a choice is often made between versions of the log-rank and the Mann-Whitney-Wilcoxon tests. We suggest, instead of making a choice, to combine the two tests. If there is a simple way of combining the tests, the combined test will be reasonably sensitive to a broad range of alternatives without being computationally burdensome. However, standard ways of combining the tests are not trivial. A direct way of combining the tests requires estimating the covariance matrix of the test statistics of the individual tests. A simple alternative is adjusting the significance levels of the individual tests by multiple testing procedures. However, in the latter case the simplicity is achieved at the expense of the power of the combined test.

We propose a simple and powerful alternative method of combining tests that is based on censored empirical likelihood (EL) with multiple constraints. The test statistics of the individual tests are formulated as linear functionals of the cumulative hazard functions and serve as the constraints for the censored EL. We show that Wilks' theorem holds for the censored EL with multiple constraints similarly as in an uncensored case. This provides a simple reference distribution for the test statistic of our proposed method. Clearly, the new method utilizes the likelihood and hence avoids directly estimating the covariance matrix of the test statistics. The proposed combined test can be much more powerful than each of its member tests (see Example 2 in Section 3 below), while it can be less powerful if one member is optimal. We discuss the relative loss of power of the combined test at the end of Section 3.

We note that the multiple testing procedure is only one possible application of the more general EL approach. It could also be used, e.g., to obtain confidence regions for a collection of the population quantiles. More specifically, the EL results of this paper are readily applicable to the two sample quantile testing problems of Kosorok (1999) and to one sample quantile problems by using the test statistics of the quantile tests as constraints for the censored EL. The proposed EL approach is simpler than the existing quantile tests as it does not require estimating the covariance matrix of the test statistics.

The theoretical interest of this paper is filling a gap in the literature for censored empirical likelihood with multiple parameters. Although empirical likelihood has appeared as a useful nonparametric statistical inference method since Owen (1988), there are less available results for censored data and most, if not all, are concerned with just one parameter. Pan and Zhou (2002) studied the right censored data empirical likelihood with a parameter that is a general functional of the cumulative hazard. For functionals that are simple indicator functions, results are in Murphy (1995) and Thomas and Grunkemeier (1975). For the case where the parameter is a general functional of the distribution function, results can be found in Murphy and Van der Vaart (1997) and Pan and Zhou (1999). However, no results for the censored empirical likelihood with multiple parameters have been available.

The rest of the paper is organized as follows. Section 2 describes theoretical aspects of the proposed method. We present the general results in Sections 2.1 and 2.2, and apply the empirical likelihood approach to the multiple testing situation in Section 2.3. Section 3 provides empirical results to confirm Wilks' theorem for the multiple constrained censored empirical likelihood ratio, and a simulation study comparing our proposed method to different weighted log-rank statistics, Renyi-type suprema, and maximin efficiency robust tests (MERT, Gastwirth, 1985). Application of the proposed method is also illustrated on real data with the log-rank and Gehan-Wilcoxon tests. The empirical results are obtained by implementing the proposed method with functions in an R package 'emplik'. All proofs are deferred to the appendix.

2 Censored Empirical Likelihood with k ($k > 1$) Constraints

We will first explain the underlying theory of the proposed method in the one sample case. The results extend straightforwardly to the two sample situation.

2.1 One Sample Censored Empirical Likelihood

For n independent, identically distributed observations, $X_1, \cdots, X_n$, assume the distribution of the X_i is $F_x(t)$ and the cumulative hazard function of X_i is $\Lambda_x(t)$. With right censoring, we only observe

$$T_i = \min(X_i, C_i) \quad \text{and} \quad \delta_i = I_{[X_i \le C_i]} \tag{1}$$

where the C_i's are the censoring times, assumed to be independent, identically distributed, and independent of the X_i's. Based on the censored observations, the log empirical likelihood pertaining to the distribution F_x is

$$\log EL(F_x) = \sum [\delta_i \log \Delta F_x(T_i) + (1-\delta_i) \log\{1 - F_x(T_i)\}] \ . \tag{2}$$

As shown in Pan and Zhou (2002), computations are much easier with the empirical likelihood reformulated in terms of the (cumulative) hazard function. The equivalent hazard formulation of (2), denoted by $\log EL(\Lambda_x)$, is given as follows:

$$\log EL(\Lambda_x) = \sum_i \{d_i \log v_i + (R_i - d_i)\log(1-v_i)\} \tag{3}$$

where $d_i = \sum_{j=1}^n I_{[T_j = t_i]}\delta_j$, $R_i = \sum_{j=1}^n I_{[T_j \geq t_i]}$, and t_i are the ordered, distinct values of T_i. See, for example, Thomas and Grunkemeier (1975) and Li (1995) for similar notation. Here, $0 < v_i \leq 1$ are the discrete hazards at t_i. The maximization of (3) with respect to v_i is known to be attained at the jumps of the Nelson-Aalen estimator: $v_i = d_i/R_i$.

Let us consider a hypothesis testing problem for a k dimensional parameter $\theta = (\theta_1, \cdots, \theta_k)^T$ with respect to the cumulative hazard function such that

$$H_0 : \theta = \mu \quad \text{vs.} \quad H_A : \theta \neq \mu \qquad \text{for } \ \theta_r = \int g_r(t)\log(1 - d\Lambda_x(t)), \ r = 1, \cdots, k$$

where the $g_r(t)$ are some nonnegative functions and $\mu = (\mu_1, \cdots, \mu_k)^T$ is a vector of k constants. We note that the θ_r are functionals of the cumulative hazard function. The constraints we shall impose on the hazards v_i are: for given functions $g_1(\cdot), \cdots, g_k(\cdot)$ and constants $\mu_1, \cdots, \mu_k$, we have

$$\sum_i^{N-1} g_1(t_i)\log(1-v_i) = \mu_1 \ , \quad \cdots \quad , \ \sum_i^{N-1} g_k(t_i)\log(1-v_i) = \mu_k \ , \tag{4}$$

where N is the total number of distinct observation values. We need to exclude the last value as we always have $v_N = 1$ for discrete hazards. Let us abbreviate the maximum likelihood estimators of $\Delta\Lambda_x(t_i)$ under constraints (4) as v_i. Application of the Lagrange multiplier method shows

$$v_i(\lambda) = \frac{d_i}{R_i + n\lambda^T G(t_i)} \ ,$$

where $G(t_i) = \{g_1(t_i), \cdots, g_k(t_i)\}^T$ and λ is the solution to the maximization of (3) under the constraints in (4) (Lemma 1 in the appendix). Then, the test statistic in terms of hazards is given by

$$W_2 = -2\{\log \max EL(\Lambda_x)(\text{with constraint (4)}) - \log \max EL(\Lambda_x)(\text{without constraint})\} \ .$$

We have the following result that proves a version of Wilks' theorem for W_2 under some regularity conditions which include the standard conditions on censoring that allow the Nelson-Aalen estimators to have an asymptotic normal distribution (see Andersen *et al.*, 1993, for details).

Theorem 1. *Suppose that the null hypothesis H_0 holds, i.e. $\mu_r = \int g_r(t) \log\{1 - d\Lambda_x(t)\},\ r = 1, \ldots, k$. Then, the test statistic W_2 has asymptotically a chi-square distribution with k degrees of freedom.*

Remark 1 The integration constraints are originally given as $\theta_r = \int g_r(t) d \log\{1 - F_x(t)\}$, $r = 1, \cdots, k$. The above formulations are found by using the suggestive notation $d \log\{F_x(t)\} = \log\{d\Lambda_x(t)\}$. These two formulations are identical for both continuous and discrete $F_x(t)$.

Remark 2: If the functions $g_r(t)$ are random but predictable with respect to the filtration $\mathcal{F}_t$ (see Gill, 1980), Theorem 1 is still valid.

2.2 Two Sample Censored Empirical Likelihood

Suppose in addition to the censored sample of X-observations, we have a second sample $Y_1, \cdots, Y_m$ coming from a distribution function $F_y(t)$ with a cumulative hazard function $\Lambda_y(t)$. Assume that the Y_j's are independent of the X_i's. With censoring, we can only observe

$$U_j = \min(Y_j, S_j) \qquad \text{and} \qquad \tau_j = I_{[Y_j \le S_j]} \tag{5}$$

where S_j are the censoring variables for the second sample. Denote the ordered, distinct values of the U_j by s_j.

Similar to (3), the log empirical likelihood function based on the two censored samples pertaining to the cumulative hazard functions Λ_x and Λ_y is simply $EL(\Lambda_x, \Lambda_y) = L_1 + L_2$ where

$$L_1 = \sum_i d_{1i} \log v_i + \sum_i (R_{1i} - d_{1i}) \log(1 - v_i) \quad \text{and}$$

$$L_2 = \sum_j d_{2j} \log w_j + \sum_j (R_{2j} - d_{2j}) \log(1 - w_j), \tag{6}$$

with d_{1i}, R_{1i}, d_{2j} and R_{2j} defined analogous to the one sample situation (see p.5). Accordingly, let us consider a hypothesis testing problem for a k dimensional parameter $\theta = (\theta_1, \cdots, \theta_k)^T$ with respect to the cumulative hazard functions Λ_1 and Λ_2 such that

$$H_0 : \theta = \mu \quad \text{vs.} \quad H_A : \theta \neq \mu,$$

where $\theta_r = \int g_{1r}(t) \log\{1 - d\Lambda_x(t)\} - \int g_{2r}(t) \log\{1 - d\Lambda_y(t)\}$, $r = 1, \cdots, k$, for some predictable functions $g_{1r}(t)$ and $g_{2r}(t)$. Then, the constraints imposed on v_i and w_j are

$$\mu_r = \sum_{i=1}^{N-1} g_{1r}(t_i) \log(1 - v_i) - \sum_{j=1}^{M-1} g_{2r}(s_j) \log(1 - w_j), \quad r = 1, \ldots, k, \tag{7}$$

where N and M are the total number of distinct observation values from the two samples. As in the one sample case, we need to exclude the last values.

Let us abbreviate the maximum likelihood estimators of $\Delta\Lambda_x(t_i)$ and $\Delta\Lambda_y(s_j)$ under the constraints (7) as v_i and w_j, respectively, where $i = 1, \cdots, N$ and $j = 1, \cdots, M$. Application of the Lagrange multiplier method shows

$$v_i(\lambda) = \frac{d_{1i}}{R_{1i} + \min(n, m)\lambda^T G_1(t_i)} \ , \qquad w_j(\lambda) = \frac{d_{2j}}{R_{2j} - \min(n, m)\lambda^T G_2(s_j)} \ ,$$

where $G_1(t_i) = \{g_{11}(t_i), \cdots, g_{1k}(t_i)\}^T$, $G_2(s_j) = \{g_{21}(s_j), \cdots, g_{2k}(s_j)\}^T$, and λ is the solution to maximizing $EL(\Lambda_x, \Lambda_y) = L_1 + L_2$ under the constraints in (7). Then, the two-sample test statistic is given as follows:

$$W_2^* = -2\{\log \max EL(\Lambda_x, \Lambda_y)(\text{with constraint (7)}) - \log \max EL(\Lambda_x, \Lambda_y)(\text{without constraint})\}$$

analogous to the one-sample case. The following theorem provides the asymptotic distribution result for W_2^*.

Theorem 2. *Suppose that the null hypothesis $H_0 : \theta_r = \mu_r$ holds. i.e. $\mu_r = \int g_{1r}(t) \log\{1 - d\Lambda_x(t)\} - \int g_{2r}(t) \log\{1 - d\Lambda_y(t)\}$, $r = 1, \ldots, k$. Then, as $\min(n, m) \to \infty$, W_2^* has asymptotically a chi-square distribution with k degrees of freedom.*

2.3 Combined Multiple Testing Based on Censored Empirical Likelihood

The basic idea of combining a family of tests by the multiple constrained censored EL is to formulate the test statistics of the individual tests as linear functionals of the cumulative hazard functions and using them as the constraints of the multiple constrained empirical likelihood. To be more specific, let us consider a hypothesis $H_0 : \Lambda_x = \Lambda_0$ vs. $H_1 : \Lambda_x \neq \Lambda_0$ and consider combining the log rank test and one of its weighted versions for the one sample problem. We can formulate the test statistics of the $G^{\rho,\gamma}$ family of Harrington and Fleming (1982) with respect to the hazard as $\sum_i^{N-1} h(t_i, \rho, \gamma) \log(1 - v_i)$, where

$$h(t, \rho, \gamma) = R(t)\hat{S}(t^-)^\rho (1 - \hat{S}(t^-))^\gamma \qquad \text{for } \rho, \gamma \geq 0, \tag{8}$$

where $R(t) = \sum I_{[T_i \geq t]}$ and $\hat{S}(t)$ denotes the Kaplan-Meier estimator. The test statistics of the log rank and Wilcoxon tests correspond to (8) with $(\rho, \gamma) = (0, 0)$ and $(\rho, \gamma) = (1, 0)$ respectively. Note that the function $h(t, \rho, \gamma)$ is a nonnegative, random yet predictable function. In the combined test, the null hypothesis in Theorem 1 becomes $\mu_r = \int g_r(t) \log(1 - d\Lambda_0(t)),\ r = 1, \ldots, k$, where different functions g_r correspond to $h(t, \rho, \gamma)$ with different choices of ρ and γ in display (8). Then, the test statistic W_2 is obtained under the constraints in (4) with $h(t, \rho, \gamma)$ with appropriate choices of ρ and γ serving as g_r.

In a two-sample problem, the test statistics of individual tests can be formulated as $\sum_{i=1}^{N-1} h^*(t_i, \rho, \gamma) \log(1 - v_i) - \sum_{j=1}^{M-1} h^*(s_j, \rho, \gamma) \log(1 - w_j)$, where

$$h^*(u, \rho, \gamma) = (\frac{n+m}{nm})^{1/2} W(u)^\rho (1 - W(u))^\gamma \frac{R_1(u)R_2(u)}{R_1(u) + R_2(u)} \qquad \text{for } \rho \geq 0, \tag{9}$$

and where $R_1(u) = \sum I_{[T_i \geq u]}$ and $R_2(u) = \sum I_{[U_j \geq u]}$. If $W(u) = \hat{S}(u^-)$ and $\hat{S}$ is the pooled sample Kaplan-Meier estimator, then the test corresponds to the $G^{\rho,\gamma}$ family of Harrington and Fleming (1982). If $W(u) = \{R_1(t) + R_2(t)\}/(n + m)$ and $\gamma = 0$, then it corresponds to the Tarone-Ware (1977) class of statistics. The value $(\rho, \gamma) = (0, 0)$ corresponds to the log-rank statistic in both cases, while $(\rho, \gamma) = (1, 0)$ corresponds to the Prentice-Wilcoxon statistic in the Harrington and Fleming (1982) class and the Gehan-Wilcoxon statistics (Gehan, 1965) in the Tarone-Ware (1977) class. Also note that $h^*(u, \rho, \gamma)$ are predictable functions. In the combined test, we choose g_{1r} and g_{2r} to be the function $h^*(u, \rho, \gamma)$ with some appropriate ρ and γ. Hence, the null hypothesis in Theorem 2 becomes $\mu_r = \int g_r(t)\log\{1 - d\Lambda_x(t)\} - \int g_r(t)\log\{1 - d\Lambda_y(t)\},\ r = 1,\ldots,k$, where different functions g_r correspond to $h^*(u,\rho,\gamma)$ with different choices of ρ and γ in display (9). If we are concerned with testing whether $\Lambda_x(t)$ is equal to $\Lambda_y(t)$, then $\mu_r = 0$ for $r = 1,\ldots,k$. The test statistic W_2^* is obtained under the constraints in (7) where $h^*(u,\rho,\gamma)$ with appropriate choices of ρ and γ serves as $g_r = g_{1r} = g_{2r}$.

3 Examples and Simulations

We provide Monte Carlo simulation results for one- and two-sample cases to empirically confirm the chi-square limit distribution of the -2 log empirical likelihood ratio with multiple constraints. The proposed methods are illustrated with real data sets where we combine the log-rank and Wilcoxon tests for one sample and the log-rank and Gehan-Wilcoxon tests for two samples. Furthermore, we show results from an extensive comparative simulation study including members of the $G^{\rho,\gamma}$ family of weighted log-rank statistics, the associated Renyi-type suprema ($GS^{\rho,\gamma}$), and their maximin efficiency robust test (MERT, Gastwirth, 1985) counterparts. All the computations have been carried out using version 0.9-1 of the emplik package in R.

Simulation 1

This simulation study examines the distribution of the -2 log empirical likelihood ratio in the one sample case with multiple constraints where the constraints are the non-random functions $g_1(t) = \exp(-t)$, $g_2(t) = \frac{1}{2}t \cdot I_{[t<1]}$, and $g_3(t) = I_{[t<0.9]}$. We use the following distributions to generate the random variables.

$$X \sim \exp(1)\ , \qquad C \sim \exp(0.5), \tag{10}$$

and the censored observations are created via (1). The Q-Q plot (Figure 1) is based on 5,000 runs. It agrees well with the theoretically derived $\chi^2_{(3)}$ distribution.

Simulation 2

This simulation study examines the distribution of the -2 log empirical likelihood ratio with multiple constraints where the constraints are random functions. We choose random functions to correspond to the test statistics of the log rank and Gehan-Wilcoxon tests in order to confirm the chi-square limit distribution of the proposed test statistic and the level of the combined test.

The test statistics of the log rank and Gehan-Wilcoxon tests belong to the Tarone-Ware (1977) class of statistics and correspond to (9) with $W(u) = \{R_1(u)+R_2(u)\}/(n+m)$, $(\rho,\gamma) = (0,0)$ and $(\rho,\gamma) = (1,0)$ respectively. In each of 10,000 runs, two identically distributed equal sized random samples are generated from the simulation setup in (10), and the test statistic W_2^* is calculated under the constraints in (7) where $h^*(u,\rho,\gamma)$ with the prescribed $W(u)$ serve as g_{1r} and g_{2r}. Table 1 shows that the proposed combined test attains the type I error at the nominal levels. Figure 2 shows that the distribution of W_2^* agrees well with $\chi^2_{(2)}$. The distribution deviates in the tail area with the sample size $n = 30$, but the deviation is in the extreme end of the tail.

Example 1. Iowa Psychiatric Patient Data

We apply the combined test of the log-rank and Wilcoxon tests to a sample of survival times of 26 psychiatric inpatients to compare with the survival time distribution of the general population in Iowa. The data is part of a larger study of psychiatric inpatients admitted to the University of Iowa hospital during the years 1935-1948 (for more information on the data, see Tsuang and Woolson, 1977). Klein and Moeschberger (1997, p. 189) use the data to illustrate the one-sample log-rank test. The test statistics of the log rank and Wilcoxon tests are $h(t,\rho,\gamma)$ in (8) with $(\rho,\gamma) = (0,0)$ and $(\rho,\gamma) = (1,0)$, respectively. The $h(t,\rho,\gamma)$ are adjusted to accommodate the delayed entries. We use them as $g_r(t)$, $r = 1,2$. When applied individually, the log-rank and Gehan tests both reject the null with p-values < 0.001 and 0.0432. The combined test statistic reaches the same conclusion with the p-value 0.00088.

Example 2. Kidney Dialysis Patient Data

We apply the combined test of the log-rank and Gehan-Wilcoxon test to re-analyze the kidney dialysis data of Klein and Moeschberger (1997, p. 197). The test statistics of the log-rank and Gehan-Wilcoxon tests correspond to (9) where $W(u) = \{R_1(t) + R_2(t)\}/(n+m)$ and $(\rho,\gamma) = (0,0)$ and $(\rho,\gamma) = (1,0)$, respectively. Out of a total of 119 patients, 43 had a catheter surgically placed and 76 percutaneously (for a detailed description of the data, see Nahman *et al.*, 1992). The plot of the estimated survival functions (Figure 3) shows that the curves cross each other at about 6 months and suggests that the survival experience of the two groups is different. However, as indicated in the introduction, the log-rank test and its weighted versions make different decisions. Both the log-rank and Gehan-Wilcoxon tests, two of the most popular ones, fail to reject the null hypothesis with p-values 0.112 and 0.964 respectively, while tests of the $G^{\rho,\gamma}$ family with emphasis on the later time period reject the null. Electing to apply such $G^{\rho,\gamma}$ family class test, though, is usually a post hoc decision. When our proposed method of combining the tests is applied, it rejects the null with a p-value of 0.001. This indicates that the combined test can be much more powerful than either one of the individual tests.

Simulation 3

We compare the small and moderate sample size behaviors of the proposed test with the $G^{\rho,\gamma}$ family of weighted log-rank statistics, the associated Renyi-type suprema ($GS^{\rho,\gamma}$), and their maximin efficiency robust test (MERT, Gastwirth, 1985) counterparts. Kosorok and Lin (1999) conducted extensive Monte Carlo simulation studies to compare their function-indexed weighted log-rank test with the $G^{\rho,\gamma}$ family of weighted log-rank statistics, the associated Renyi-type suprema and suprema plus infimum, and their MERT counterparts. We have replicated Kosorok and Lin's (1999) simulation study design with our proposed method and compared the results (Tables 2 and 3). $M^{[0,\rho_0]\times[0,\gamma_0]}$ denotes the MERT test with the statistic taken for the $G^{\rho,\gamma}$ family for $\rho \in [0, \rho_0]$ and $\gamma \in [0, \gamma_0]$. When only 0 is considered as a value for γ, it is reduced to $M^{[0,\rho_0]\times\{0\}}$. Kosorok and Lin (1999) implemented the MERT by taking the test statistics of $M^{[0,\rho_0]\times\{0\}}$ and $M^{[0,\rho_0]\times[0,\gamma_0]}$ as linear combinations of $G^{0,0}$, $G^{\rho_0,0}$ and $G^{\rho_0,0}$, G^{0,γ_0}, and G^{ρ_0,γ_0} respectively. Therefore, the proposed EL counterparts are $E^{[0,\rho_0]\times\{0\}}$ and $E^{[0,\rho_0]\times[0,\gamma_0]}$ with the test statistics of the corresponding $G^{\rho,\gamma}$ family as constraints. We use Theorem 2 to find critical values for our test from a regular χ^2-distribution, while Kosorok and Lin (1999) conducted 1,000 Monte Carlo replications to construct the critical regions for each simulated data. As Kosorok and Lin's (1999) function-indexed weighted log-rank test is more computationally intensive than the MERT with slightly better performance, we only present the results for the $G^{\rho,\gamma}$ family of weighted log-rank statistics, the associated $GS^{\rho,\gamma}$, and their MERT counterparts in Tables 2 and 3. The column labels "N" and "A"-"E" stand for a "null" and 5 different alternative models indexed similarly in Kosorok and Lin (1999). The number of simulations is 10,000 for the null distributions and 1,000 for the alternatives (2,000 for the EL alternatives).

The simulation results show that the proposed EL test performs comparably to the MERT where the MERT performs well, while it is more reliable where the MERT performs poorly. For example, with the alternative model A where the hazards are proportional and the log-rank test ($G^{0,0}$) is optimal, both the MERT and EL have comparable powers to the power of $G^{0,0}$: the ranges of loss of power are (0.003, 0.033) and (0.0095, 0.0285) for the MERT and EL in a moderate sample, respectively, and (0.013, 0.087) and (0.125, 0.3) in a small sample. However, with the alternative model D where the hazards differ at the beginning and their difference disappears later, $GS^{4,0}$ has the highest power among those considered in Kosorok and Lin (1999)'s original simulation and the MERT performs poorly, while the EL performs reasonably: the ranges of loss of power are (0.373, 0.658) and (0.0145, 0.2445) for the MERT and EL in a moderate sample and (0.291, 0.431) and (0.0575, 0.266) in a small sample. Similar results are observed for alternative model E where the hazard functions cross and $G^{0,1}$ has the highest power: the ranges of loss of power are (0.466, 0.8577) and (0.1815, 0.685) in a moderate sample and (0.266, 0.700) and (0.1275, 0.699) in a small sample. The reliable performance of the proposed EL test is even more distinct in the censored case.

In addition, we would like to point out the computational advantage of our proposed test whose critical values can easily be obtained from a χ^2-distribution.

Appendix

Mathematical Derivations and Proofs

Recall the column vectors $G(t) = \{g_1(t), \cdots, g_k(t)\}^T$ and $\lambda = \{\lambda_1, \cdots, \lambda_k\}^T$.

Lemma 1. *The hazards that maximize the log likelihood function (3) under the constraints (4) are given by*

$$v_i(\lambda) = \frac{d_i}{R_i + n\lambda^T G(t_i)} \, . \tag{11}$$

where the λ value is obtained as the solution of the following k equations

$$\sum_i^{N-1} g_1(t_i) \log\{1 - v_i(\lambda)\} = \mu_1 \, , \quad \cdots \quad , \sum_i^{N-1} g_k(t_i) \log\{1 - v_i(\lambda)\} = \mu_k \, . \tag{12}$$

PROOF OF LEMMA 1. The result follows from a standard Lagrange multiplier argument applied to (3) and (4). See Fang and Zhou (2000) for some similar calculations. $\diamondsuit$.

We denote the solution of (12) by λ_x.

Lemma 2. *Assume the data are such that the Nelson-Aalen estimator is asymptotically normal and the variance-covariance matrix Σ defined in the appendix (p. 14) is invertible. Then, for the solution λ_x of the constrained problem (12), corresponding to the null hypothesis $H_0 : \ \mu_r = \int g_r(t) \log\{1 - d\Lambda_x(t)\}, \ r = 1, \ldots, k$, we have that $n^{1/2}\lambda_x$ converges in distribution to $N(0, \Sigma)$.*

PREPARATION FOR THE PROOFS OF LEMMA 2 AND THEOREM 1.

Let

$$f(\lambda) = \sum \left[d_i \log v_i(\lambda) + (R_i - d_i) \log\{1 - v_i(\lambda)\} \right] \ . \tag{13}$$

In order to show that $f'(0) = 0$, we compute

$$\frac{\partial}{\partial \lambda_r} f(\lambda) = \sum_i \frac{d_i}{v_i(\lambda)} \frac{\partial v_i(\lambda)}{\partial \lambda_r} - \frac{(R_i - d_i)}{v_i(\lambda)} \frac{\partial(1 - v_i(\lambda))}{\partial \lambda_r}, \ r = 1, \ldots, k.$$

Letting $\lambda = 0$ and after some simplification we have

$$\frac{\partial}{\partial \lambda_r} f(\lambda)|_{\lambda=0} = -\sum_i (R_i - R_i) \frac{d_i n g_r(t_i)}{R_i^2} \equiv 0 \ .$$

We now compute $f''(0) = \sum$. The rl^{th} element of the $k \times k$ matrix $\sum$ is

$$D_{rl} = \frac{\partial^2}{\partial \lambda_r \partial \lambda_l} f(\lambda)|_{\lambda=0} \ .$$

After straightforward but tedious calculations, we obtain

$$D_{rl} = -\left\{ \sum_i \frac{n^2 g_r g_l}{R_i} \frac{d_i}{R_i - d_i} \right\} \ .$$

By a now standard counting process martingale argument, we see

$$-\frac{D_{rl}}{n} \to D^*_{rl} \ .$$

PROOF OF LEMMA 2. We derive the asymptotic distribution of λ. The argument is similar to, for example, Owen (1990) and Pan and Zhou (2002). Define a vector function $h(s) = \{h_1(s), \cdots, h_k(s)\}$ by

$$h_1(s) = \sum_i g_1(t_i) \log\{1 - v_i(s)\} - \mu_1 \ , \cdots , \ h_k(s) = \sum_i g_k(t_i) \log\{1 - v_i(s)\} - \mu_k \ . \tag{14}$$

Then, λ is the solution of $h(s) = 0$. Thus we have

$$0 = h(\lambda) = h(0) + h'(0)\lambda + o_p(n^{-1/2}) \ , \tag{15}$$

where $h'(0)$ is a $k \times k$ matrix.

Indeed,

$$\begin{aligned} 0 = |h(\lambda)| &= |\sum_i G(t_i) \log\{1 - v_i(s)\} - \mu| = |\sum_i G(t_i) \log\{1 - \frac{d_i}{R_i + n\lambda^T G(t_i)}\} - \mu| \\ &\geq |\sum_i G(t_i) \log(1 - \frac{d_i}{R_i}) - \mu| - |\sum_i G(t_i) \log\Big[\frac{1 - d_i/\{R_i + n\lambda^T G(t_i)\}}{1 - d_i/R_i}\Big]| \\ &= A - B \ , \end{aligned}$$

where the first expression A is of the order $O_p(n^{-1/2})$. Considering the second expression,

$$\begin{aligned} B &= |\sum_i G(t_i) \log\Big[\frac{1 - d_i/\{R_i + n\lambda^T G(t_i)\}}{1 - d_i/R_i}\Big]| \\ &\geq |\sum_i G(t_i) \frac{nG(t_i)^T \lambda d_i}{(R_i - d_i)(R_i + n\lambda^T G(t_i))}| \\ &\geq |\frac{|\lambda|}{1 + n|\lambda^T| \max_i G(t_i)/R_i} \sum_i \frac{nG(t_i)G(t_i)^T d_i}{(R_i - d_i)R_i}| \end{aligned}$$

The sum converges to $|D^*|$ and is therefore of order $O_p(1)$, so it follows that $|\lambda|$ is of order $O_p(n^{-1/2})$, and hence the expansion (15) is valid.

Therefore,

$$n^{1/2}\lambda = \{h'(0)\}^{-1}\{-n^{1/2}h(0)\} + o_p(1) \ .$$

The elements of $h'(0)$ are easily computed:

$$h'_{rl} = \sum_i \frac{n g_r g_l d_i}{R_i(R_i - d_i)} \ .$$

Notice we have verified $nh'_{rl} = -D_{rl}$. By the counting process martingale central limit theorem (see, for example, Gill, 1980; Andersen *et al.*, 1993; or Fang and Zhou, 2000), we can show that $n^{1/2}h(0)$ converges in distribution to $N(0, \Sigma_h)$ with $\Sigma_h = \lim h'(0)$.

Finally, putting it together, we have that $n^{1/2}\lambda(0) = \{h'(0)\}^{-1}\{-n^{1/2}h(0)\} + o_p(1)$ converges in distribution to $N(0, \Sigma)$ with $\Sigma = \lim\{h'(0)\}^{-1}$. Recalling $nh'_{rl} = -D_{rl}$, we see that $\Sigma^{-1} = D^*$. $\diamondsuit$

PROOF OF THEOREM 1. Let $f(\lambda)$ be defined as in (13). Then, we have $W_2 = -2\{f(\lambda_x) - f(0)\}$. By Taylor expansion, we obtain

$$W_2 = 2\{f(0) - f(0) - f'(0)\lambda_x - \frac{1}{2}\lambda_x^T D\lambda_x + o_p(1)\}, \tag{16}$$

where we use D to denote the matrix of second derivatives of $f(\cdot)$ with respect to λ. The expansion is valid in view of Lemma 2 (λ_x is close to zero).

Notice that we have $f'(0) = 0$ (see above), so the expression above is reduced to

$$W_2 = -\lambda_x^T D\lambda_x + o_p(1) \ . \tag{17}$$

Notice that $-D$ is symmetric and positive definite for large enough n because $-D/n$ converges to a positive definite matrix, see below. Therefore, we may write

$$W_2 = \lambda_x^T(-D)^{1/2}(-D)^{1/2}\lambda_x + o_p(1) \ . \tag{18}$$

Recalling the distributional result for λ_x in Lemma 2 and noticing that

$$-\frac{D}{n} \to D^* \ ,$$

and $D^* = \Sigma^{-1}$ (see above in the proof of Lemma 2), it is not hard to show that $n^{1/2}\lambda_x^T(D^{1/2}n^{-1/2})$ converges in distribution to $N(0, I)$. This together with (17) implies that W_2 converges in distribution to χ_k^2 . ♢

The proof of Theorem 2 is analogous to the one for the one-sample situation and is therefore omitted.

REFERENCES

Andersen, P. K., Borgan, O., Gill, R., and Keiding, N., 1993. *Statistical Models Based on Counting Processes.* Springer, New York.

β-Blocker Heart Attack Trial Research Group, 1982. A randomized trial of propranolol in patients with acute myocardial infarction. *J. Amer. Medical Assoc.* 247, 1707–1714.

Fang, H. and Zhou, M., 2000. On empirical likelihood ratio method with k-sample problems. Technical Report, Department of Statistics, University of Kentucky.

Fleming, T. R. and Harrington, D. P., 1984. Nonparametric estimation of the survival distribution in censored data. *Comm. in Stat. Theory and Methods* 13, 20, 2469–2486.

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

Fleming, T. R., Harrington, D. P., and O'Sullivan, M., 1987. Supremum versions of the log-rank and generalized Wilcoxon statistics. *J. Amer. Statist. Assoc.* 82, 312–320.

Gastwirth, J. L., 1985. The use of maximin efficiency robust tests in combining contingency tables and survival analysis. *J. Amer. Statist. Assoc.* 80, 380–384.

Gill, R., 1980. *Censoring and Stochastic Integrals.* Mathematical Centre Tracts 124. Mathematisch Centrum, Amsterdam.

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

Klein, J. P. and Moeschberger, M. L., 1997. *Survival Analysis – Techniques for Censored and Truncated Data.* Springer, New York.

Kosorok, M. R., 1999. Two-sample quantile tests under general conditions. *Biometrika* 86, 909–921.

Kosorok, M. R. and Lin, C.-Y., 1999. The versatility of function-indexed weighted log-rank statistics. *J. Amer. Statist. Assoc.* 94 (445), 320–332.

Lai, T.L. and Z. Ying, Z., 1991. Rank regression methods for left-truncated and right-censored data. *Ann. Statist.* 19, 531–556.

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

Letón, E. and Zuluaga, P., 2005. Relationships among tests for censored data. *Biom. J.* 47 (3), 377–387.

Li, G., 1995. On nonparametric likelihood ratio estimation of survival probabilities for censored data. *Statist. Probab. Lett.* 25, 95–104.

Lin, C.-Y. and Kosorok, M. R., 1999. A general class of function-indexed nonparametric tests for survival analysis. *Ann. Statist.* 27 (5), 1722–1744.

Murphy, S. A., 1995. Likelihood ratio based confidence intervals in survival analysis. *J. Amer. Statist. Assoc.* 90, 1399–1405.

Murphy, S. and Van der Vaart, 1997. Semiparametric likelihood ratio inference. *Ann. Statist.* 25, 1471–1509.

Nahman, N.S., Middendorf, D.F., Bay, W.H., McElligott, R., Powell, S., and Anderson, J., 1992. Modification of the percutaneous approach to peritoneal dialysis catheter placement under peritoneoscopic visualization: Clinical results in 78 patients. *J. Amer. Society of Nephrology* 3, 103–107

Owen, A., 1988. Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* 75, 237–249.

Owen, A., 1990. Empirical likelihood ratio confidence regions. *Ann. Statist.* 18, 90–120.

Owen, A., 2001. *Empirical Likelihood*. Chapman & Hall, London.

Pan, X.R. and Zhou, M., 1999. Using 1-parameter sub-family of distributions in empirical likelihood ratio with censored data. *J. Statist. Plann. Inference* 75 (2), 379–392.

Pan, X.R. and Zhou, M., 2002. Empirical likelihood in terms of cumulative hazard function for censored data. *J. Multivariate Analysis* 80 (1), 166–188.

Tarone, R. E., 1981. On the distribution of the maximum of the log-rank statistic and the modified Wilcoxon statistic. *Biometrics* 37, 79–85.

Tarone, R. E. and Ware, J., 1977. Distribution-free tests for equality of survival distributions. *Biometrika* 64, 1, 156–160.

Thomas, D. R. and Grunkemeier, G. L., 1975. Confidence interval estimation of survival probabilities for censored data. *J. Amer. Statist. Assoc.* 70, 865–871.

Tsuang, M. T. and Woolson, R. F., 1977. Mortality in patients with schizophrenia, mania and depression. *British J. Psychiatry* 130, 162–166

A Tables and Figures

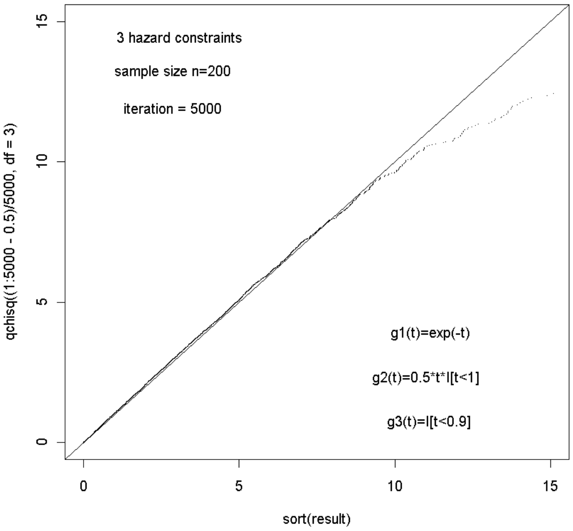

Figure 1: Q-Q plot of -2log-lik Ratios vs. $\chi^2_{(3)}$ percentiles for sample size $= 200$ (one sample).

n	α				
	0.01	0.05	0.1	0.15	0.2
30	0.0144	0.0550	0.1032	0.1511	0.1981
50	0.0103	0.0524	0.1057	0.156	0.2052
100	0.0102	0.0476	0.1	0.1508	0.19993

Table 1: Estimated type I errors at various significance levels α

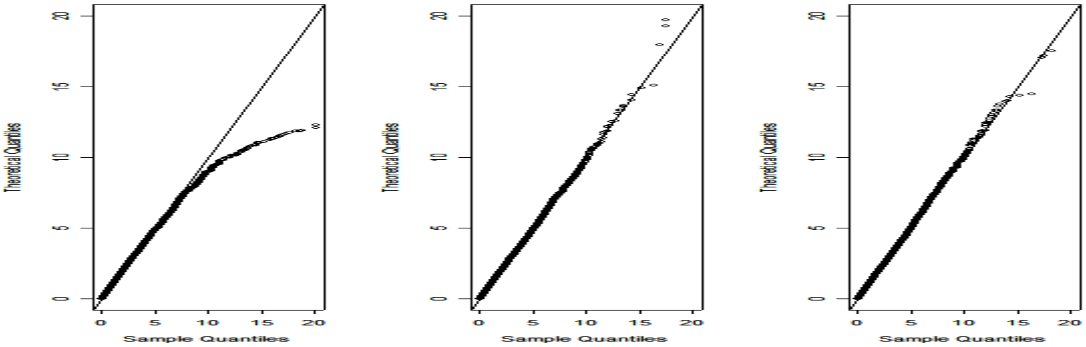

Figure 2: Q-Q plot of -2log-lik Ratios vs. $\chi^2_{(2)}$ percentiles for sample size $n = 30, 50, 100$

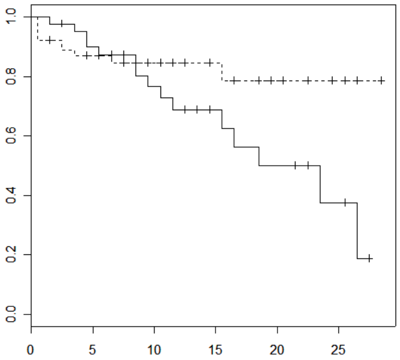

Figure 3: Estimated survival functions for kidney dialysis patients with percutaneously (dashed line) and surgically (solid line) placed catheters.

Table 2: Moderate sample results with $n = 100$.

	Uncensored						Censored					
	N	A	B	C	D	E	N	A	B	C	D	E
$G^{0,0}$	.0525	.996	.809	.654	.105	.711	.0512	.965	.846	.299	.155	.131
$GS^{0,0}$	.0506	.994	.935	.567	.392	.649	.0518	.955	.939	.240	.592	.166
$G^{1,0}$	.0518	.974	.979	.130	.324	.130	.0524	.935	.965	.091	.386	.049
$GS^{1,0}$	.0520	.955	.992	.105	.810	.243	.0488	.900	.980	.084	.820	.202
$G^{4,0}$	.0470	.780	.995	.056	.756	.197	.0474	.738	.990	.056	.759	.215
$GS^{4,0}$	.0482	.706	.995	.054	.855	.324	.0474	.668	.988	.059	.849	.304
$G^{0,1}$	.0566	.979	.180	.959	.079	.980	.0520	.890	.223	.677	.086	.572
$GS^{0,1}$	.0512	.983	.202	.940	.087	.974	.0574	.900	.280	.622	.097	.555
$G^{4,1}$	.0502	.917	.942	.053	.280	.079	.0496	.878	.893	.057	.320	.055
$GS^{4,1}$	.0516	.889	.960	.054	.784	.144	.0498	.830	.921	.059	.778	.125
$M^{[0,1]\times\{0\}}$	.0526	.993	.943	.349	.197	.400	.0514	.959	.959	.164	.253	.067
$E^{[0,1]\times\{0\}}$	.0386	.9840	.9750	.8325	.6105	.2950	.0544	.9275	.9610	.5975	.7810	.2915
$M^{[0,4]\times\{0\}}$	.0492	.979	.986	.239	.442	.123	.0524	.932	.932	.126	.476	.055
$E^{[0,4]\times\{0\}}$	.0440	.9865	.9870	.6590	.8405	.5470	.0518	.9255	.9730	.3810	.8345	.3905
$M^{[0,4]\times[0,1]}$	.0536	.993	.939	.608	.286	.446	.0504	.957	.957	.329	.273	.106
$E^{[0,4]\times[0,1]}$	.0452	.9645	.9845	.7930	.7830	.4585	.0558	.9005	.9560	.5505	.8205	.3320

Table 3: Small sample results with $n = 20$ for a null and $n = 50$ for the alternatives.

	Uncensored						Censored					
	N	A	B	C	D	E	N	A	B	C	D	E
$G^{0,0}$	.0632	.902	.532	.384	.091	.422	.0612	.759	.601	.176	.102	.094
$GS^{0,0}$	.0564	.881	.670	.308	.150	.349	.0568	.728	.710	.128	.279	.114
$G^{1,0}$	.0522	.801	.801	.099	.193	.101	.0494	.687	.777	.061	.218	.060
$GS^{1,0}$	.0476	.740	.860	.074	.463	.142	.0500	.625	.813	.059	.481	.129
$G^{4,0}$	.0456	.497	.889	.049	.477	.122	.0456	.447	.837	.047	.464	.138
$GS^{4,0}$	.0470	.409	.881	.047	.564	.181	.0464	.376	.822	.044	.539	.183
$G^{0,1}$	.0958	.832	.141	.745	.090	.789	.0818	.643	.170	.407	.091	.349
$GS^{0,1}$	.0814	.827	.147	.688	.083	.769	.0826	.653	.199	.364	.097	.335
$G^{4,1}$	.0500	.674	.706	.057	.168	.098	.0500	.607	.629	.053	.194	.068
$GS^{4,1}$	.0538	.603	.740	.049	.436	.109	.0528	.549	.665	.054	.436	.103
$M^{[0,1]\times\{0\}}$	.0546	.877	.702	.211	.133	.218	.0540	.738	.704	.108	.160	.067
$E^{[0,1]\times\{0\}}$	.0222	.7180	.7585	.2435	.2980	.0900	.0496	.6270	.7165	.2185	.4590	.1725
$M^{[0,4]\times\{0\}}$	.0538	.819	.823	.148	.259	.089	.0496	.672	.784	.084	.280	.060
$E^{[0,4]\times\{0\}}$	.0352	.7395	.8120	.2400	.4885	.2350	.0592	.6340	.7465	.1680	.5065	.2215
$M^{[0,4]\times[0,1]}$	.0662	.887	.694	.371	.173	.262	.0626	.746	.660	.194	.161	.083
$E^{[0,4]\times[0,1]}$	.0614	.6020	.7650	.2275	.4320	.1985	.1062	.5680	.7070	.2170	.4810	.2025