Electronic Journal of Statistics
Vol. 10 (2016) 2511–2536
ISSN: 1935-7524
DOI: 10.1214/16-EJS1180

Empirical likelihood based tests for stochastic ordering under right censorship

Hsin-wen Chang

Institute of Statistical Science, Academia Sinica,
128 Academia Road, Section 2, Nankang, Taipei 11529, Taiwan
e-mail: hwchang@stat.sinica.edu.tw

and

Ian W. McKeague

Department of Biostatistics, Columbia University,
722 West 168th Street, New York, NY 10032, U.S.A.
e-mail: im2131@columbia.edu

Abstract: This paper develops an empirical likelihood (EL) approach to testing for stochastic ordering between two univariate distributions under right censorship. The proposed test is based on a maximally selected local EL statistic. The asymptotic null distribution is expressed in terms of a Brownian bridge. The new procedure is shown via a simulation study to have superior power to the log-rank and weighted Kaplan–Meier tests under crossing hazard alternatives. The approach is illustrated using data from a randomized clinical trial involving the treatment of severe alcoholic hepatitis.

MSC 2010 subject classifications: 62G30, 62G10, 62N03, 62G20.
Keywords and phrases: Crossing survival/hazard functions, order restricted inference, survival analysis, two-sample problem.

Received August 2015.

Contents

1. Introduction

When comparing survival patterns between two treatment groups in a randomized clinical trial (RCT), it is often of interest to examine whether there is a uniformly higher survival function in one of the groups. For example, in a recent RCT involving patients with severe alcoholic hepatitis, the objective is to compare a combination therapy of prednisolone plus *N*-acetylcysteine with prednisolone alone. Testing whether the combination therapy has a consistently higher/lower survival probability (throughout the follow-up period) addresses the issue directly, as opposed to the practice of using an omnibus alternative (i.e., *any* difference between the survival functions), or using the log-rank test which detects ordered hazards instead of ordered survival functions. This paper develops such a testing procedure that allows us to establish an ordering between two survival curves uniformly over time.

Our approach is framed in terms of the classical notion of *stochastic ordering*. A survival function S_1 is said to be *stochastically larger* than another survival function S_2 if $S_1(t) \geq S_2(t)$ for all $t \geq 0$, and denoted $S_1 \succeq S_2$. When in addition there is strict inequality for some t, it is denoted $S_1 \succ S_2$. Here and in the sequel we implicitly assume that $t \geq 0$ is further restricted to a given follow-up period $[t_1, t_2]$, a common practice in simultaneous inference of survival functions in censored data [see, e.g., 3, 33].

Notice that the parameter space for (S_1, S_2) can be split into four separate hypotheses: $H_0 \colon S_1 = S_2$, $H_1 \colon S_1 \succ S_2$, $H_1' \colon S_2 \succ S_1$, and $H_c \colon S_1(u) > S_2(u)$ for some $u \geq 0$, $S_1(v) < S_2(v)$ for some $v \geq 0$, i.e., *crossing survival functions*. The problem then is to test the null hypothesis $H_0 \cup H_c$ versus the alternative $H_1 \cup H_1'$ of stochastic ordering based on independent right-censored random samples from S_1 and S_2. Our proposed approach is to combine a test that the survival functions do not cross,

$$H_c \text{ versus } H_0 \cup H_1 \cup H_1', \tag{1}$$

with a test of stochastic ordering,

$$H_0 \text{ versus } H_1 \cup H_1' \tag{2}$$

under the assumption of no crossing. If the first test does not reject H_c, we terminate the procedure and conclude no evidence for $H_0 \cup H_1 \cup H_1'$, the cases of equal and stochastically ordered survival functions. If the first test rejects H_c, we proceed to the second test of stochastic ordering. The composite procedure concludes the alternative of stochastic ordering if both of these tests reject. We will show in Section 2.3 that the family-wise error rate of the combined procedure can be controlled at the same alpha level as the individual tests. Also, if prior information precludes the possibility of a crossing (e.g., when comparing survival probability of early- and late-stage cancer patients), one can skip testing (1).

A feature of our proposed composite procedure is that it can be adapted easily to utilize prior information on the direction of stochastic ordering. This can result in more powerful testing than simply reading a simultaneous (two-sided) confidence band for the difference $S_1 - S_2$ on $[t_1, t_2]$. For example, when comparing survival curves between patients with early-stage (S_1) and late-stage (S_2) cancer diagnoses, it is not biologically plausible to consider $S_1 \prec S_2$. We could utilize such prior information by testing only a one-sided alternative (see (3)) as H_1' is not reasonable, providing greater power. In contrast, the usual confidence bands correspond to two-sided tests and can have lower power in testing one-sided alternatives. We could construct a one-sided confidence band for this purpose. But then the band alone cannot test for absence of a crossing, and a two-step method as our composite procedure is still needed.

Our test for the absence of a crossing (1) is a straightforward adaptation of a simultaneous confidence band for the difference $S_1 - S_2$ on $[t_1, t_2]$, which crosses the time axis when the survival functions cross. The most challenging part of the problem is to produce a test of (2), or even more simply for the one-sided alternative

$$H_0 \colon S_1 = S_2 \text{ versus } H_1 \colon S_1 \succ S_2. \tag{3}$$

This test is tractable, however, given that S_1 and S_2 do not cross, since in that case it suffices to detect whether $S_1(t) > S_2(t)$ at some t. Developing this second test constitutes the main part of the paper. A test of the two-sided alternative in (2) is then easily constructed using the union-intersection principle applied to the two one-sided test statistics for H_1 and H_1' (i.e., (3) in each direction).

Commonly used two-sample tests for censored data include the log-rank test and weighted Kaplan–Meier (WKM) tests [34], and these tests can be one-sided or two-sided. The log-rank test is based on an integrated weighted difference between hazard functions, and is thus designed to detect ordered hazards instead of more general stochastic ordering. Other tests based on weighted differences between hazard functions, such as the $\mathcal{K}$-class of weighted log-rank statistics [17, 16], also share this property. The WKM class of tests targets stochastically ordered alternatives by estimating an integrated weighted difference between survival functions, but such test statistics depend on an ad hoc weight function that needs to be specified throughout follow-up.

We derive our procedure for testing (3) using the empirical likelihood (EL) method. EL involves forming a ratio of two nonparametric likelihoods subject to constraints on the parameters of interest. The method originates with Thomas

and Grunkemeier [35], who constructed pointwise confidence intervals for survival functions from right-censored data. EL has also been used to provide confidence regions for parameters defined by estimating equations [29, 30], in numerous censored and uncensored settings. EL enjoys many appealing properties: highly accurate confidence regions, self-studentization and the possibility of Bartlett correctability. There is also evidence that EL-based tests have optimal power [see, e.g., 21]. On the other hand, order restricted inference is known to be challenging for EL [see, e.g., 30, Ch. 10], and much less has been done in this direction. El Barmi [12] explored EL tests for order-restricted hypotheses of the form $g(\eta) \leq 0$, where g is some smooth function and η is a finite-dimensional parameter specified by estimating equations [see also 38]. Other recent contributions in this direction have been made by Andrews and Guggenberger [2] and Canay [5]. As for order restrictions on distribution functions, El Barmi and McKeague [13] studied EL-based tests for stochastic ordering, while Davidov et al. [7] investigated EL-based tests for likelihood ratio ordering under a semiparametric biased sampling model. However, these tests are limited to uncensored data.

As already mentioned, the main part of our proposed procedure is an EL-test for one-sided stochastic ordering (3). The idea is to construct a localized EL statistic for $H_0^t\colon S_1(t) = S_2(t)$ versus $H_1^t\colon S_1(t) > S_2(t)$ at each given t . The key step in this construction is to recast the stochastic ordering constraint into an inequality involving a single Lagrange multiplier. Then the proposed test rejects H_0 for large values of the maximally selected EL statistic. A maximally selected test statistic is used (as opposed to integral-type) because it is more sensitive to local differences between the survival functions. Kolmogorov–Smirnov type test statistics (not based on EL) for stochastic ordering have been proposed by El Barmi and Mukerjee [14] and Davidov and Herman [8]. Besides localization, another possible approach might be to use the full nonparametric likelihood [10, 31] and compute its ratio under $S_1 \succ S_2$ versus $S_1 = S_2$. However, we find the localization approach to be much more tractable. The localization approach has been used in Einmahl and McKeague [11], Davidov and Herman [9] and El Barmi and McKeague [13] for testing various nonparametric hypotheses, except they considered an integral type test statistic and restricted attention to uncensored data. Park et al. [32] proposed a localized NPMLE under stochastic ordering (for right-censored data), but its asymptotic distribution is not known, so it is unclear how a formal test could be developed using their approach.

Various ways of formulating EL in right-censored data settings have been proposed. The standard approach for censored data [35, 23] maximizes the censored data likelihood subject to contraint(s) on the parameter of interest. Wang and Jing [37] instead used the nonparametric likelihood for uncensored data and plug-in of the Kaplan–Meier (KM) estimator of the censoring distribution. We use the former approach as it is tractable and more natural in our setting. There are in fact two different versions of EL for censored data, namely the binomial and Poisson versions [see, e.g., 26]. We utilize the binomial version.

The paper is organized as follows. In Section 2.1 we set up the general framework and notation to be used throughout the paper. The main focus of our

procedure, the two-sample test for stochastic ordering, is developed in Section 2.2. The first part of our procedure, the test that the survival functions do not cross, is then discussed in Section 2.3. Related tests are discussed in Section 2.4: stochastic ordering in the null hypothesis, a test for crossing survival functions, and an integral type statistic. Section 3 presents the results of a simulation study: the proposed two-sample EL test is shown to outperform the log-rank and WKM tests under different stochastically ordered alternatives, including alternatives with crossing hazards. We have also shown the effectiveness of our combined procedure in ruling out H_c when testing for stochastic ordering. Application of the proposed test to the RCT mentioned earlier is given in Section 4, and some concluding remarks are placed in Section 5.

2. EL tests for stochastic ordering under right censorship

2.1. Preliminaries

We introduce notation for the one-sample setup, then add *a further subscript j indicating the j-th sample* ($j = 1, 2$) for the two-sample case in the corresponding notation. Let X_i and C_i for $i = 1, \ldots, n$ be i.i.d. from unknown survival functions S and G, respectively; only $\min(X_i, C_i)$ and $I(X_i \leq C_i)$ are observed. The lifetimes X_i and the censoring times C_i are assumed to be independent. Also, $S(0) = G(0) = 1$. Order the uncensored lifetimes as $0 < T_1 < \ldots < T_m < \infty$. For each T_i $(i = 0, \ldots, m)$, let r_i be the number alive just before T_i, d_i be the number of deaths at T_i and h_i be the hazard at T_i. Let $N(t)$ be the number of observed lifetimes that are less than or equal to t. Then the nonparametric likelihood (depending on the unknown survival function) supported on the observed lifetimes is proportional to

$$L(S) \equiv \prod_{i=1}^{m} h_i^{d_i} (1 - h_i)^{r_i - d_i} \tag{4}$$

for $h_i \in [0, 1]$. The NPMLE for $S(t)$, namely the KM estimator $\hat{S}(t) = \prod_{i \leq N(t)} (1 - d_i/r_i)$, is asymptotically normal with variance $S^2(t)\sigma^2(t)$, where $\sigma^2(t) = -\int_0^t dS(u)/\{S(u)S(u-)G(u-)\}$. This variance can be consistently estimated by the well-known Greenwood formula, $\hat{S}^2(t)\hat{\sigma}^2(t)$, where $\hat{\sigma}^2(t) = n \sum_{i \leq N(t)} [d_i/\{r_i (r_i - d_i)\}]$.

For the two-sample case, the nonparametric likelihood (denoted as $L(S_1, S_2)$) is proportional to $L(S_1)L(S_2)$ by independence between the two samples. The sample proportion n_j/n is assumed to converge to some $p_j > 0$, where $n = n_1 + n_2$. The $\hat{\sigma}^2(t)$ now equals the weighted average $n\{\hat{\sigma}_1^2(t)/n_1 + \hat{\sigma}_2^2(t)/n_2\}$, consistently estimating $\sigma_1^2(t)/p_1 + \sigma_2^2(t)/p_2$.

2.2. Two-sample test for stochastic ordering

Now we develop the main part of our combined procedure, the two-sample test for stochastic ordering. As described in the Introduction, we focus on the one-

sided test for (3), then construct a test of the two-sided alternative in (2) using the union-intersection principle applied to the two one-sided test statistics for H_1 and H_1' (i.e., (3) in each direction).

Consider the "local" hypotheses H_0^t: $S_1(t) = S_2(t)$ versus H_1^t: $S_1(t) > S_2(t)$ for a given t, and the EL ratio

$$\mathcal{R}(t) = \frac{\sup\{L(S_1, S_2)\colon S_1(t) = S_2(t)\}}{\sup\{L(S_1, S_2)\colon S_1(t) \geq S_2(t)\}}, \tag{5}$$

where we use the conventions $\sup\emptyset = 0$ and $0/0 = 1$. Note that the numerator and denominator of $\mathcal{R}(t)$ maximize $L(S_1)L(S_2)$ over $(h_{11}, \ldots, h_{m_1 1}, h_{12}, \ldots, h_{m_2 2}) \in [0,1]^m$ $(m = m_1 + m_2)$ subject to the constraints

$$\prod_{i \leq N_1(t)} (1 - h_{i1}) = \text{ or } \geq \prod_{i \leq N_2(t)} (1 - h_{i2}), \tag{6}$$

respectively. We solve this constrained maximization problem using the Karush–Kuhn–Tucker (KKT) method [4], a generalization of the Lagrange method that allows inequality constraints. As the constraints are placed only on the lifetimes up to t, the terms after t turn out to be the same in both the numerator and denominator and thus cancel out. Also, for some t the maximum is attained on the boundary of the constraint set, in which case $\mathcal{R}(t) = 1$. Specifically, in Appendix A we establish the following expression for the EL ratio:

$$\mathcal{R}(t) = \begin{cases} 1, & \hat{\lambda} \geq 0, \\ \displaystyle\prod_{j=1}^{2} \prod_{i \leq N_j(t)} \frac{\hat{h}_{ij}^{d_{ij}} \left(1 - \hat{h}_{ij}\right)^{r_{ij} - d_{ij}}}{\bar{h}_{ij}^{d_{ij}} \left(1 - \bar{h}_{ij}\right)^{r_{ij} - d_{ij}}}, & \hat{\lambda} < 0, \end{cases} \tag{7}$$

where $\bar{h}_{ij} = d_{ij}/r_{ij}$, $\hat{h}_{ij} = d_{ij}/(r_{ij} + (-1)^{j-1}\hat{\lambda})$, and the Lagrange multiplier $\hat{\lambda}$ is determined by the equality in (6) with h_{ij} replaced by $\hat{h}_{ij}$. Here we have suppressed the dependence of $\hat{\lambda}$ and $\hat{h}_{ij}$ on t.

Based on the above expression, we can derive large sample properties of the local EL test statistic, $-2\log\mathcal{R}(t)$. This is done by approximating $-2\log\mathcal{R}(t)$ via a Taylor expansion as a function of the difference between $\log\hat{S}_1(t)$ (recall from Section 2.1 that $\hat{S}(t)$ is the KM estimator) and $\log\hat{S}_2(t)$. We then make use of asymptotic properties of $\hat{S}_j(t)$ $(j = 1, 2)$ to establish the weak convergence of $-2\log\mathcal{R}(t)$. The asymptotic null distribution turns out to be chi-bar square Namely, for t such that $0 < S_0(t) < 1$ and $G_j(t) > 0$ for $j = 1, 2$,

$$-2\log\mathcal{R}(t) \xrightarrow{d} Z_+^2$$

under H_0^t, where $Z \sim N(0,1)$ and $Z_+ = \max(Z, 0)$. This result can be used to test the local hypotheses H_0^t versus H_1^t.

To test for the alternative of stochastic ordering, we propose the following maximally selected EL statistic:

$$K_n = \sup_{t\in[t_1,t_2]} \{-2\log \mathcal{R}(t)\}, \tag{8}$$

where $0 < t_1 < t_2 < \infty$ are to be specified. We suppress the dependence of K_n on t_1 and t_2. Guidance on the choice of $[t_1, t_2]$ is provided later.

Our first result gives the asymptotic null distribution of K_n (see Appendix B for the proof).

Theorem 1. *Suppose H_0 holds and the common survival function S_0 is continuous. For t_1 and t_2 satisfying $S_0(t_1) < 1$ and $S_0(t_2)G_j(t_2) > 0$ for $j = 1, 2$,*

$$K_n \xrightarrow{d} \sup_{x\in[x_1,x_2]} \left\{ \frac{B_+^2(x)}{x(1-x)} \right\},$$

where B is a standard Brownian bridge on $[0, 1]$, $B_+ = \max(B, 0)$, $x_j = b(t_j)$ for $j = 1, 2$, and $b(t) = \sigma^2(t)/\{1 + \sigma^2(t)\}$.

To implement the test, we pre-specify one of the intervals $[t_1, t_2]$ or $[x_1, x_2] = [b(t_1), b(t_2)]$ and determine the other via $b(t)$ or $b^{-1}(x) = \inf\{t : b(t) \geq x\}$. However, b is unknown, so one of the two intervals has to be estimated.

We can choose $[t_1, t_2]$ based on the smallest and the largest observed lifetimes [see, e.g., 3] or some other biological considerations [33], and then estimate $[x_1, x_2]$ (by $[\hat{x}_1, \hat{x}_2]$ say). But we cannot tabulate critical values in advance, because $[\hat{x}_1, \hat{x}_2]$ varies across different data sets. In this case, instead of using the tabulated critical values, R code supplied in a supplementary file [6] can be used to compute the critical value based on $[\hat{x}_1, \hat{x}_2]$.

On the other hand, pre-determining $[x_1, x_2]$ allows "universal" critical values, and this is the approach we take. Both the choice of $[x_1, x_2]$ and details of implementation will be provided in the next subsection.

2.2.1. Calibrating the test

This section discusses issues in calibrating the test. The first one is the choice of $[x_1, x_2]$. Secondly, having chosen $[x_1, x_2]$, we explain how to estimate $[t_1, t_2]$ and implement the proposed EL test. Justification for this calibration procedure will be provided in Appendix C, where a statistic K_n^* is defined for K_n with estimated $[t_1, t_2]$. Critical values for the test are then obtained via simulation in Section 3.

The choice of $[x_1, x_2]$ is important because the interval width can affect power of the EL test. In a similar context, this issue has been discussed by Davidov and Herman [8]; they proposed a (non-EL-based) test of stochastic ordering for uncensored data via localization, and point out that a narrower $[x_1, x_2]$ gives smaller critical values, but may fail to capture deviations (from H_0) outside the interval. We have investigated their recommendation $[x_1, x_2] = [0.2, 0.8]$ in a

simulation study (see Section 3 and Table 4), and found that the performance is not very sensitive to the choice of x_2, so our preference is to choose x_2 close to 1 to utilize the local statistics on the right tail. Our simulation study (in Section 3) shows that the choice $x_1 = 0.2$ and $x_2 = 0.98$ performs well in terms of balancing power and accuracy, and this is what we recommend in practice.

Having specified $[x_1, x_2]$, we need to estimate $[t_1, t_2]$. Under suitable conditions on b^{-1}, t_l can be consistently estimated by $\hat{b}^{-1}(x_l) = \inf\{t : \hat{b}(t) \geq x_l\}$ for $l = 1, 2$, where

$$\hat{b}(t) = \frac{\hat{\sigma}^2(t)}{1 + \hat{\sigma}^2(t)}$$

is a consistent estimator of $b(t)$. We can then compute K_n^* accordingly, based on the estimates $\hat{t}_1$ and $\hat{t}_2$. To ensure stability of K_n^* in small samples, we consider only values of $-2\log\mathcal{R}(t)$ inside the interval formed by the smallest and the largest observed lifetimes, $[\max(T_{11}, T_{12}), \min(T_{m_1 1}, T_{m_2 2})]$. Such restriction has been used in simultaneous inference of survival functions in censored data [see, e.g., 3]. This leads to considering only $t \in [\max(\hat{t}_1, T_{11}, T_{12}), \min(\hat{t}_2, T_{m_1 1}, T_{m_2 2})]$. Note that this modification makes no difference asymptotically, since $[\max(T_{11}, T_{12}), \min(T_{m_1 1}, T_{m_2 2})] \supset [\hat{b}^{-1}(x_1), \hat{b}^{-1}(x_2)]$ eventually.

2.2.2. Two-sided testing

The above one-sided test for stochastic ordering has an immediate extension to a two-sided test for (2), as needed for the second part of the composite testing procedure described in the Introduction. The two-sided alternative in (2) is the union of the two one-sided alternatives ($S_1 \succ S_2$ or $S_2 \succ S_1$). Based on the union-intersection principle, the test statistic is the maximum of the two one-sided test statistics. The asymptotic null distribution of this test statistic is $\sup_{x\in[x_1,x_2]}[B^2(x)/\{x(1-x)\}]$, where B is a standard Brownian bridge, as in Theorem 1. The test can therefore be calibrated in much the same way as we did for the one-sided test.

2.3. Crossing survival functions

As explained in the Introduction, if there is no prior information that excludes the possibility of crossing survival functions, we conduct the first part of our combined testing procedure. It calls for a consistent test of (1), and this can be done using a $100(1-\alpha)\%$ simultaneous confidence band for the difference $S_1 - S_2$, and showing that the asymptotic level of the resulting test is bounded above by α. The follow-up interval $[t_1, t_2]$ to be used in this connection will be specified in a later section. A band for the ratio S_1/S_2 could be used in a similar fashion [see, e.g., the EL band given by 24], but here for simplicity we only consider the difference approach.

Consider the random multiplier bootstrap band $\mathcal{B}_n$ for $S_1 - S_2$ developed by Parzen et al. [33]. Intuitively, the data would support the presence of crossing

(H_c) when the lower boundary of $\mathcal{B}_n$ is > 0 at some time point (implying $S_1(u) > S_2(u)$ for some $u \geq 0$), and its upper boundary is < 0 at another time point (implying $S_1(v) < S_2(v)$ for some $v \geq 0$). Therefore, the opposite should lead to rejection of the null hypothesis H_c of a crossing: if the lower boundary of $\mathcal{B}_n$ is ≤ 0 or its upper boundary ≥ 0.

Note that $\mathcal{B}_n$ is centered on the difference $\hat{S}_1 - \hat{S}_2$ of the KM estimators, and the results of Parzen et al. [33] imply (under the same conditions assumed here) that it has coverage $P(S_1 - S_2 \in \mathcal{B}_n) \to 1 - \alpha$, and maximal width $M_n = O_p(1/\sqrt{n})$. This leads to an asymptotic level α test as follows. Under H_c, there exist $u, v \geq 0$ such that $S_1(u) - S_2(u) = \sup_t\{S_1(t) - S_2(t)\} > 0$ and $S_1(v) - S_2(v) = \inf_t\{S_1(t) - S_2(t)\} < 0$. Let $\epsilon = \min\{S_1(u) - S_2(u), -S_1(v) + S_2(v)\} > 0$. Clearly H_c is not rejected in the event that $S_1 - S_2 \in \mathcal{B}_n$ and $M_n < \epsilon$, so

$$P(H_c \text{ rejected}) \leq P(S_1 - S_2 \notin \mathcal{B}_n) + P(M_n \geq \epsilon) \to \alpha.$$

To obtain the family-wise error of conducting this test of (1) along with the test for stochastic ordering (2), we appeal to the partitioning principle of Finner and Strassburger [15]. This principle holds when the null hypotheses are disjoint (in our case H_c and H_0 are indeed disjoint), and shows that the level of each test can be chosen to be the same as the desired family-wise error rate (α).

2.4. Related tests

2.4.1. Stochastically ordered null

We have developed our test for stochastic ordering (3) under the null hypothesis $S_1 = S_2$ and under the assumption that S_1 and S_2 do not cross. Here we describe how our approach can be extended to the stochastically ordered null hypothesis $S_1 \preceq S_2$ under the same assumption (i.e., testing $H_0 \cup H_1'$ versus H_1). The local EL ratio is

$$\mathcal{R}'(t) = \frac{\sup\{L(S_1, S_2)\colon S_1(t) \leq S_2(t)\}}{\sup\{L(S_1, S_2)\}},$$

where the denominator maximizes over the union of the local (null and alternative) hypotheses and results in no constraint on $S_1(t)$ and $S_2(t)$. Since the KM estimator is the NPMLE, if $\hat{S}_1(t) \leq \hat{S}_2(t)$ the numerator of $\mathcal{R}'(t)$ coincides with the unconstrained maximum and thus equals the denominator. If $\hat{S}_1(t) > \hat{S}_2(t)$, it can be shown that the numerator attains its maximum on the boundary $S_1(t) = S_2(t)$ of the constraint set (using log-concavity of (4)). We then have

$$\mathcal{R}'(t) = \begin{cases} 1, & \hat{S}_1(t) \leq \hat{S}_2(t), \\ \dfrac{\sup\{L(S_1, S_2)\colon S_1(t) = S_2(t)\}}{\sup\{L(S_1, S_2)\}}, & \hat{S}_1(t) > \hat{S}_2(t). \end{cases}$$

Thus $\mathcal{R}'(t) = \mathcal{R}(t)$ by (7), since $\hat{\lambda} \geq 0$ is the same as $\hat{S}_1(t) \leq \hat{S}_2(t)$ by Appendix A. Hence K_n does not change under this broader null hypothesis. The same calibration method can be used to obtain an asymptotic level α test because

$$P(K_n > c_\alpha | S_1 \preceq S_2) \leq P(K_n > c_\alpha | S_1 = S_2) \to \alpha,$$

where c_α is the upper α-quantile of the limiting distribution in Theorem 1.

2.4.2. Detecting crossing survival functions

In some applications it can be of interest to test for crossing survival functions, i.e., reversing the null and alternative hypotheses in (1). This can be done by carrying out the one-sided test of Section 2.2 in *both* possible directions. The reason is here the parameter space for the one-sided tests allows for crossing, so that the test of Section 2.2 is interpreted instead as testing $S_1(t) \leq S_2(t)$ for all t versus $S_1(t) > S_2(t)$ for some t, based on Section 2.4.1 and the union-intersection principle. If both tests reject, then there is evidence of crossing survival functions. Then, using the intersection-union principle, we take the minimum of the two one-sided test statistics as the test statistic. The R code (provided online) for implementing the one-sided test is readily adapted for this purpose, with critical values obtained from simulating

$$\min \left[\sup_{x \in [x_1, x_2]} \left\{ \frac{B_-^2(x)}{x(1-x)} \right\}, \sup_{x \in [x_1, x_2]} \left\{ \frac{B_+^2(x)}{x(1-x)} \right\} \right],$$

where B_- is the negative part of the Brownian bridge B.

2.4.3. An integral type statistic

An integral type EL statistic could be developed as well, as an extension of the integrated statistics provided by El Barmi and McKeague [13] for uncensored data. However, it is challenging to find a suitable integrator that (a) is interpretable, (b) leads to an easily calibrated test. For example, a direct extension of their integrator $\hat{F}$ to the censored case (i.e., using the Kaplan–Meier estimates), as far as we know, will not lead to an asymptotically distribution free test statistic as our Lemma 3; so this extension does not satify criterion (b).

We have also tried using the following test statistic:

$$\int_{t_1}^{t_2} \{-2 \log \mathcal{R}(t)\} \, d\left(\frac{\hat{\sigma}^2(t)}{1 + \hat{\sigma}^2(t)} \right),$$

with the very unintuive integrator $\hat{\sigma}^2(t)/(1+\hat{\sigma}^2(t))$. The integrator is chosen so that the limiting distribution

$$\int_{x_1}^{x_2} \left\{ \frac{B_+^2(x)}{x(1-x)} \right\} dx$$

TABLE 1
Critical values for K_n^* for selected x_1, x_2 and α.

x_1		0.1			0.15			0.2	
$x_2\backslash\alpha$	0.01	0.05	0.1	0.01	0.05	0.1	0.01	0.05	0.1
0.975	11.822	8.255	6.648	11.672	8.074	6.489	11.542	7.953	6.365
0.98	11.912	8.329	6.720	11.758	8.159	6.556	**11.619**	**8.028**	**6.442**
0.985	11.996	8.415	6.807	11.851	8.253	6.658	11.739	8.131	6.532

is the same as the asymptotic null distribution in El Barmi and McKeague [13] with the $[x_1, x_2]$ restriction. However, the integrator is weighting the local EL statistics in a way that is difficult to interpret (criterion (a) violated), and is rather ad hoc and needs to be specified throughout follow-up, just like the WKM class of tests.

Due to the undesirable properties of these integrated statistics, we do not pursue this direction further. In comparison, our proposed maximally selected EL statistic does not need to be weighted throughout follow-up and the test is easily calibrated.

3. Simulation study

In this section, we report the results of a simulation study to evaluate various aspects of the proposed method. We start with the second (i.e., main) step in the combined procedure. We restrict our attention to one-sided tests, but results for the two-sided tests are similar. We first tabulate selected critical values, and then compare the performance of K_n^* with the (one-sided) log-rank and WKM tests in terms of accuracy and power. Finally we assess the performance of the combined procedure under models of equal and crossing survival functions, and most importantly, stochastic ordering.

3.1. Critical values and accuracy

Quantiles of the limiting distribution in Lemma 3 of Appendix C are used as critical values for K_n^*. These are computed by simulation based on 100,000 replications of standard Brownian bridge over a fine grid on $[0, 1]$ (100,000 equidistant points), for selected values of x_1 and x_2 (see Table 1).

To compute empirical significance levels, we simulate lifetimes from the piecewise exponential distribution displayed as solid line in upper left panel of Figure 1. We consider exponential censoring distribution: $G_1 = G_2 = \text{Exp}(\theta)$, where θ is chosen to give a censoring rate (CR) of 10% or 25%. Our one-sided EL statistic (K_n^*) is compared with the one-sided log-rank statistic. Another class of tests for comparison is the one-sided WKM, and we follow recomendations of Pepe and Fleming [34] and select the WKM statistic with the pooled variance estimator and the weight function denoted by $\hat{w}_c(t)$ in their paper.

Results on the size of our EL test are given in Table 2, where we use $[x_1, x_2] = [0.2, 0.98]$. The test is slightly conservative in small samples but approaches

TABLE 2
Empirical significance levels based on 10,000 replications.

CR	group size	$\alpha = 0.05$			$\alpha = 0.01$		
		K_n^*	log-rank	WKM	K_n^*	log-rank	WKM
10%	50	0.040	0.057	0.055	0.007	0.013	0.011
	80	0.041	0.052	0.054	0.008	0.010	0.010
	200	0.045	0.051	0.048	0.009	0.011	0.011
25%	50	0.037	0.057	0.054	0.006	0.012	0.012
	80	0.041	0.051	0.056	0.008	0.009	0.010
	200	0.046	0.054	0.050	0.010	0.010	0.011

the nominal level as the sample size increases. Such conservativeness has been seen in other maximal deviation-type statistics for stochastic ordering [8]. The empirical significance levels of the one-sided log-rank test and the WKM test under the same settings are closer to the nominal level, but sometimes on the anticonservative side.

3.2. Power comparisons

In this section, we compare the small sample power of the proposed test with the one-sided WKM and log-rank tests. Two models of lifetime distributions are considered, both with piecewise-constant hazards. In Model A, the hazard functions cross while the survival functions still remain stochastically ordered (see Figure 1, first column). In this case, the one-sided log-rank test can fail to detect the difference between the survival curves because it is designed to detect ordered hazards. In Model B, the two groups have different hazards initially but the same hazard later on, so the difference between the survival functions gradually wears off (see Figure 1, second column). This is a common phenomenon which is also seen in our real data example in Section 4. For both models, we consider exponential and uniform censoring distributions: $G_1 = G_2 = \text{Exp}(\theta_1)$ or $\text{Uniform}(0, c_1)$, with θ_1 or c_1 chosen to give a CR of 10% or 25% for group 1.

Results are given in Table 3 for K_n^* using $[x_1, x_2]$=$[0.2, 0.98]$. Note that K_n^* outperforms the other tests in all the cases considered, especially in the crossing hazards scenario (Model A). The much lower power of WKM in Model A is surprising, because this test were shown to work well under crossing hazard alternatives in some previous simulation examples [34]. The superior performance of our test may be due to two factors: first, our test is based on nonparametric likelihood, so it can be expected to be more powerful than tests that depend on an ad hoc weight function; second, we are using a maximal deviation-type statistic, rather than a weighted average, so our test may be more sensitive to local differences in the survival functions.

We have also investigated power under proportional hazards configurations, and our test closely matches the performance of the log-rank and WKM tests (see supplementary tables). These results show that for stochastically ordered alternatives, the proposed EL test can compete effectively with the log-rank and WKM tests, especially when the hazard functions cross.

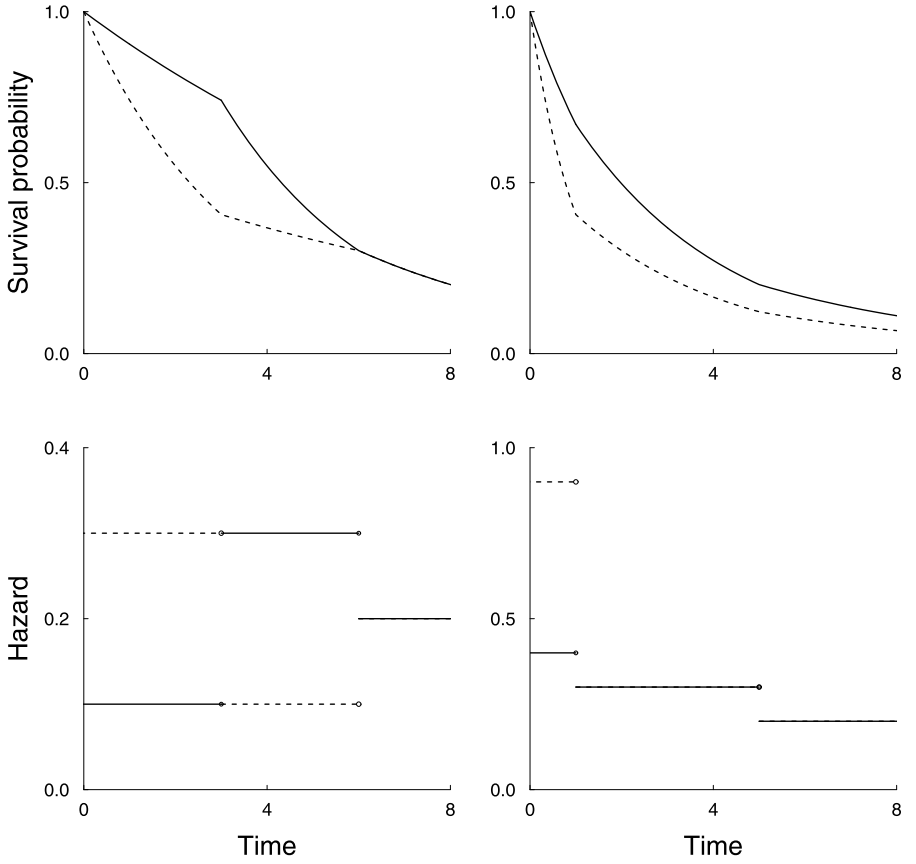

FIG 1. *The piecewise exponential survival functions (top row) and the hazard functions (bottom row) in Model A (first column): S_1 (solid) and S_2 (dashed), and in Model B (second column): S_1 (solid) and S_2 (dashed).*

Table 4 gives size and power for various choices of x_1 and x_2 reflecting light or heavy truncation. It is clear from the last two rows that light truncation on the left results in both poor accuracy and power compared with the top row, which corresponds to our recommendation $[x_1, x_2]=[0.2, 0.98]$. Yet the performance is not very sensitive to the choice of x_2, so our preference is to choose x_2 close to 1 in order to reduce truncation.

3.3. Combined procedure

Finally we conducted simulations for the combined procedure described in the Introduction: (1) testing the survival functions do not cross, and (2) testing stochastic ordering under the assumption of no crossing. The goal is to see if the combined procedure has good performance under models of equal and crossing survival functions, and most importantly, stochastic ordering.

TABLE 3

Power at $\alpha = 0.05$ based on 10,000 replications. **Model A***: survival functions as in Figure 1, upper left panel.* **Model B***: survival functions as in Figure 1, upper right panel.*

model	group size	test	exp. censoring		unif. censoring	
			10%	25%	10%	25%
Model A	50	K_n^*	0.851	0.833	0.849	0.834
		log-rank	0.318	0.379	0.314	0.373
		WKM	0.328	0.391	0.330	0.431
	80	K_n^*	0.975	0.968	0.975	0.971
		log-rank	0.416	0.503	0.415	0.501
		WKM	0.426	0.507	0.433	0.569
Model B	50	K_n^*	0.689	0.672	0.688	0.676
		log-rank	0.625	0.659	0.621	0.650
		WKM	0.521	0.583	0.521	0.613
	80	K_n^*	0.876	0.862	0.877	0.869
		log-rank	0.782	0.815	0.784	0.812
		WKM	0.660	0.729	0.675	0.775

TABLE 4

Size and power for various choices of x_1 and x_2 based on 10,000 replications, $\alpha = 0.05$, $n_1 = n_2 = 50$, and exponential censoring with censoring rate 10%. **Model A***: survival functions as in Figure 1, upper left panel.* **Model B***: survival functions as in Figure 1, upper right panel. For size, only the solid survival functions are used.*

x_1	x_2	critical value	size		power	
			Model A	Model B	Model A	Model B
0.2	0.98	8.028	0.040	0.040	0.851	0.689
0.2	0.8	6.879	0.037	0.039	0.890	0.703
0.02	0.98	8.829	0.029	0.028	0.806	0.628
0.02	0.8	8.048	0.023	0.025	0.838	0.612

TABLE 5

Proportion of decisions from the combined procedure based on 10,000 replications, $\alpha = 0.05$, $n_1 = n_2 = 80$, and exponential censoring with censoring rate 10%. **Model A***: survival functions as in Figure 1, upper left panel.* **Model A.0***: solid survival function in Figure 1, upper left panel.* **Model C***: survival functions as in Figure 2.*

Decision	Model A.0	Model A	Model C
Crossing	0	0.007	0.996
Equality	0.960	0.046	0.000
Stochastic ordering	0.040	0.947	0.004

Three models of lifetime distributions are considered: the one from Section 3.1 under H_0, the Model A from Section 3.2 under H_1, and a new Model C under H_c (see Figure 2). Decisions are labeled as "crossing" if H_c is not rejected in testing (1), "equality" if H_c is rejected but H_0 is not rejected in testing (2), and "stochastic ordering" if both H_c and H_0 are rejected. The empirical proportion of decisions from the combined procedure is then reported.

The results are summarized in Table 5. Our combined procedure is effective in correctly identifying H_0, H_c and most importantly, H_1. The results from Model A, in particular, shows the ability of our combined procedure to rule out H_c in testing for stochastic ordering.

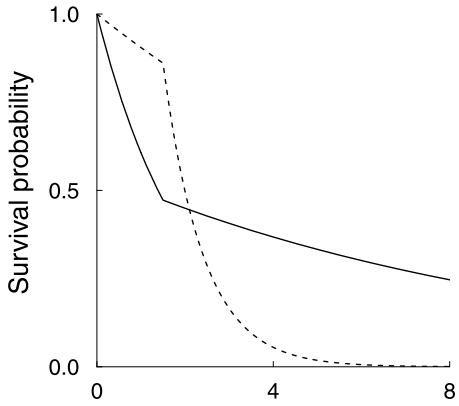

FIG 2. *The piecewise exponential survival functions in Model C: S_1 (solid) and S_2 (dashed).*

4. Application

A RCT for treatment of severe alcoholic hepatitis [28] is analyzed. The data are obtained by digitizing the published KM curves and reconstructing survival and censoring information using the algorithm developed by Guyot et al. [18]. The purpose of the trial was to assess whether a combination therapy of prednisolone plus *N*-acetylcysteine is better than prednisolone alone (the currently recommended treatment). A total of 174 patients were randomized to taking the combination ($n_1 = 85$) or only prednisolone ($n_2 = 89$), and the primary endpoint is their 6-month survival. The KM curves (see the top panel of Figure 3) suggest a stochastic ordering between the two groups.

The case of crossing survival functions is precluded via a rejection of H_c in testing (1) in our composite procedure. Application of the one-sided EL test indicates that the combination therapy group has stochastically larger survival pattern than patients receiving only prednisolone ($K_n^* = 10.36$, $p = 0.018$). In comparison, the WKM and the one-sided log-rank tests yield p-values of 0.021 and 0.037, respectively. Examining the cumulative hazards plot (see the bottom panel of Figure 3), we can see that the slopes (i.e. hazards) of the two curves only differ noticeably during the initial 40 days. Such a scenario of an initial hazard difference has been considered in Model B of Section 3.2, where we show our EL test is better adapted to detecting a difference between the two treatment groups.

Nguyen-Khac et al. [28] actually used the two-sided log-rank test and reported a p-value of 0.07. They concluded that the combination therapy does not improve the 6-month survival. In contrast, our two-sided EL test shows that the two treatment groups are significantly different and there is a uniformly higher survival function in one of the groups ($p = 0.036$, computed by the supplementary R program that implements the two-sided EL test). In this case the EL test shows a more significant result that leads to a completely different conclusion than the log-rank test.

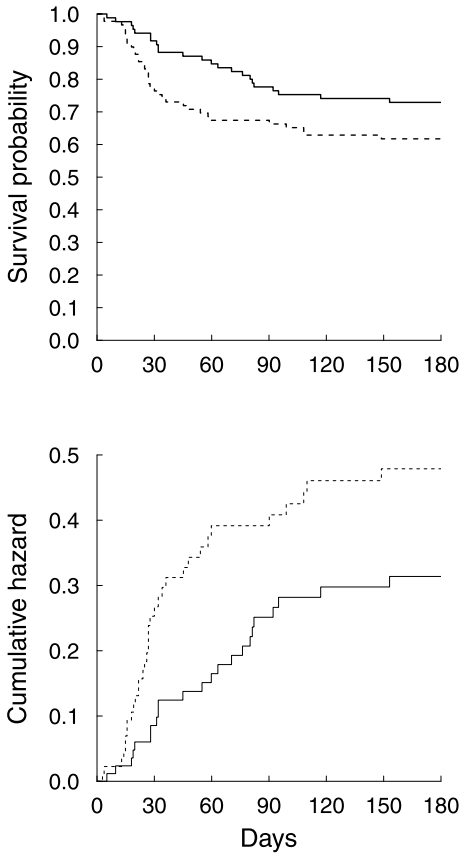

FIG 3. *Estimates of survival functions (top) and cumulative hazards (bottom) for prednisolone plus N-acetylcysteine (solid line) versus prednisolone alone (dashed line).*

5. Discussion

We have developed a class of EL-based tests for both one- and two-sided stochastically ordered alternatives under right censoring. The procedure involves first checking that the survival functions do not cross. The proposed test statistic for one-sided stochastic ordering is a maximally selected local EL statistic and is shown to be asymptotically distribution-free. The test statistic for two-sided stochastic ordering is taken as the maximum of the two one-sided test statistics. A simulation study shows that our test can be much more powerful than the log-rank and WKM tests under alternatives with crossing hazards. We applied our test to a RCT involving patients with severe alcoholic hepatitis and found a more significant result than the log-rank and WKM tests.

Our test statistics utilize a data-dependent interval $[t_1, t_2]$, much like the data-dependent weight-function used in integral-type tests based on hazard or survival functions. Such restriction has been used in simultaneous inference of

survival functions in censored data [see, e.g., 27, 3, 33], such as Nair's equal precision confidence band [27] and empirical likelihood based confidence band [20]. This cannot be avoided in procedures that are (asymptotically) based on standardized statistics, as far as we know. However, in contrast to methods that rely on the selection of a complete weight function throughout follow-up (e.g., the WKM test), it is actually much easier and more transparent to select just the two tuning parameters (x_1 and x_2) needed in our case. Although t_1 and t_2 could be specified using a data-dependent rule (such as 5% of the data in each tail), this approach would have the disadvantage of needing tailor-made critical values for each dataset. In this case, instead of using the tabulated critical values in Table 1, one can use the supplementary R code to compute a critical value based on $[\hat{x}_1, \hat{x}_2]$.

Our test targets stochastically ordered alternatives through construction of a nonparametric likelihood ratio (EL). It can be expected to be more powerful than commonly used two-sample tests that either are not tailored for such alternatives or depend on an ad hoc weight function. When combined with a test for the absence of a crossing (1), it provides more information about the nature of the difference between S_1 and S_2 compared to the omnibus alternative $S_1 \neq S_2$, in which case the functional parameters S_1 and S_2 may be ordered in one direction at certain time points, but ordered in the reverse direction at other time points.

Our central contribution is the development of the first EL-based test for ordered survival functions in right-censored data settings, and we envision the test to be useful in clinical trials, in reliability engineering, and health policy applications. It would also be of interest to extend our approach to allow the testing of stochastic ordering in k-sample censored data settings, and to explore how it could be used for other types of ordering between distributions, such as increasing convex ordering, likelihood ratio ordering and uniform stochastic ordering (or hazard rate ordering). Another direction is to generalize our approach to cover the situation with left-truncation. This can be done by using the empirical likelihood formulated in Li [22] (who considered the case of one sample, two-sided situation at one time point only), although a full derivation is well beyond the scope of this article.

Appendix A: Derivation of the local EL statistic

Here we derive the local EL ratio (7). First, we will obtain a closed-form expression for the denominator of (5) by the KKT method. After a log transformation, the optimization problem becomes minimizing

$$-\sum_{j=1}^{2}\sum_{i=1}^{m_j}\{d_{ij}(\log h_{ij}) + (r_{ij} - d_{ij})\log(1 - h_{ij})\}$$

over $(h_{11}, \ldots, h_{m_1 1}, h_{12}, \ldots, h_{m_2 2}) \in [0,1]^m$ $(m = m_1 + m_2)$ subject to the constraints

$$\sum_{i\le N_2(t)} \log(1-h_{i2}) - \sum_{i\le N_1(t)} \log(1-h_{i1}) \le 0.$$

Since the domain $[0,1]^m$ is convex, the objective and constraint functions are convex and differentiable, and Slater's condition is satisfied, the KKT conditions are necessary and sufficient for optimality. More specifically, the Lagrangian is defined as a function $\mathcal{L}: [0,1]^m \times \mathbb{R} \to \mathbb{R}$ such that

$$\begin{aligned}
&\mathcal{L}(h_1,\dots,h_m,\lambda)\\
&\equiv -\sum_{j=1}^{2}\sum_{i=1}^{m_j}\{d_{ij}(\log h_{ij}) + (r_{ij}-d_{ij})\log(1-h_{ij})\}\\
&\quad + \lambda\left\{\sum_{i\le N_2(t)} \log(1-h_{i2}) - \sum_{i\le N_1(t)} \log(1-h_{i1})\right\}.
\end{aligned}$$

The optimal solution is denoted as $(\hat{h}^1_{11},\dots,\hat{h}^1_{m_1 1},\hat{h}^1_{12},\dots,\hat{h}^1_{m_2 2},\hat{\lambda}^1)$, with the superscript indicating the correspondence of the denominator with H_1. The dependence of the solution on t is omitted here for simplicity but will appear in the proof of Theorem 1 (see Appendix B) when the EL ratio is considered as a process indexed by t. The optimal solution must satisfy the KKT conditions:

$$\nabla_h \mathcal{L}(\hat{h}_{11},\dots,\hat{h}_{m_1 1},\hat{h}_{12},\dots,\hat{h}_{m_2 2},\hat{\lambda}^1) = 0, \tag{9}$$

$$\prod_{i\le N_1(t)}\left(1-\hat{h}^1_{i1}\right) \ge \prod_{i\le N_2(t)}\left(1-\hat{h}^1_{i2}\right), \tag{10}$$

$$\hat{\lambda}^1 \ge 0, \tag{11}$$

$$\hat{\lambda}^1\left\{\sum_{i\le N_2(t)} \log(1-\hat{h}^1_{i2}) - \sum_{i\le N_1(t)} \log(1-\hat{h}^1_{i1})\right\} = 0, \tag{12}$$

which are known as stationarity, primal feasibility, dual feasibility, and complementary slackness, respectively. The stationarity condition yields $\hat{h}^1_{ij} = d_{ij}/r_{ij}$ for $i = N_j(t)+1,\dots,m_j$ and

$$\hat{h}^1_{ij} = \frac{d_{ij}}{r_{ij} + (-1)^{j-1}\hat{\lambda}^1}$$

for $i = 1,\dots,N_j(t)$, for each $j = 1,2$. Define $D_j = \max_{i=1,\dots,N_j(t)}(d_{ij} - r_{ij})$. Since $(\hat{h}^1_{11},\dots,\hat{h}^1_{m_1 1},\hat{h}^1_{12},\dots,\hat{h}^1_{m_2 2})$ should be in the domain $[0,1]^m$, we have that $D_1 \le \hat{\lambda}^1 \le -D_2$, where $D_j \le 0$ for $j = 1,2$.

The numerator of $\mathcal{R}(t)$ can be handled in a similar fashion. Denoting the optimal solution to the Lagrangian by $(\hat{h}^0_{11},\dots,\hat{h}^0_{m_1 1},\hat{h}^0_{12},\dots,\hat{h}^0_{m_2 2},\hat{\lambda}^0)$, it turns out $\hat{h}^0_{ij}$ has the same form as $\hat{h}^1_{i,j}$ but with $\hat{\lambda}^1$ replaced by $\hat{\lambda}^0$, and $\hat{\lambda}^0$ only needs to satisfy $D_1 \le \hat{\lambda}^0 \le -D_2$ and

$$\prod_{i\le N_1(t)}\left(1-\hat{h}^0_{i1}\right) = \prod_{i\le N_2(t)}\left(1-\hat{h}^0_{i2}\right). \tag{13}$$

Note that the estimated hazards after time t under no constraints, namely $\hat{h}_{ij}^{v}$ for $v = 0, 1$ and $i = N_j(t) + 1, \ldots, m_j$, are the same in the numerator and denominator, and so these terms cancel out. This leads to

$$\mathcal{R}(t) = \prod_{j=1}^{2} \prod_{i \le N_j(t)} \frac{\left(\hat{h}_{ij}^{0}\right)^{d_{ij}} \left(1 - \hat{h}_{ij}^{0}\right)^{r_{ij} - d_{ij}}}{\left(\hat{h}_{ij}^{1}\right)^{d_{ij}} \left(1 - \hat{h}_{ij}^{1}\right)^{r_{ij} - d_{ij}}}. \tag{14}$$

We next further simplify $\mathcal{R}(t)$ by analyzing the relationship between $\hat{\lambda}^0$ and $\hat{\lambda}^1$, namely by showing that $\hat{\lambda}^1 = 0$ when $\hat{\lambda}^0 < 0$ and $\hat{\lambda}^1 = \hat{\lambda}^0$ when $\hat{\lambda}^0 \ge 0$. Defining

$$a_j(\lambda) \equiv \prod_{i \le N_j(t)} \left\{1 - \frac{d_{ij}}{r_{ij} + (-1)^{j-1}\lambda}\right\}$$

for $j = 1, 2$ and

$$a(\lambda) \equiv \frac{a_1(\lambda)}{a_2(\lambda)},$$

we can see that $a_j(0) = \hat{S}_j(t)$, $\hat{\lambda}^0$ satisfies $a(\hat{\lambda}^0) = 1$, and $\hat{\lambda}^1$ satisfies $a(\hat{\lambda}^1) \ge 1$. Notice that $a(\lambda)$ is strictly increasing in λ on $(D_1, -D_2)$, tending to 0 and ∞ as $\lambda \downarrow D_1$ and $\uparrow -D_2$, respectively. Also, condition (12) implies either $\hat{\lambda}^1 = 0$ or

$$\sum_{i \le N_2(t)} \log(1 - h_{i2}) - \sum_{i \le N_1(t)} \log(1 - h_{i1}) = 0 \tag{15}$$

must hold, and since (15) is equivalent to $\hat{\lambda}^1 = \hat{\lambda}^0$, we obtain that $\hat{\lambda}^1$ is either 0 or $\hat{\lambda}^0$. These observations along with (10) and (11) imply the following:

Case 1: If $\hat{\lambda}^0 < 0$, then by (11) we have $\hat{\lambda}^1 \ne \hat{\lambda}^0$. Since $\hat{\lambda}^1$ is either 0 or $\hat{\lambda}^0$, we obtain that $\hat{\lambda}^1 = 0$.

Case 2: If $\hat{\lambda}^0 > 0$, then by monotonicity of $a(\lambda)$ we have $a(0) < 1$. Suppose $\hat{\lambda}^1 = 0$, then $a(0) \ge 1$ by (10), which contradicts $a(0) < 1$. So we have $\hat{\lambda}^1 = \hat{\lambda}^0$.

Case 3: If $\hat{\lambda}^0 = 0$, then because $\hat{\lambda}^1$ is either 0 or $\hat{\lambda}^0$, we can see that $\hat{\lambda}^1 = \hat{\lambda}^0 = 0$.

Then from (14) we have

$$\mathcal{R}(t) = \begin{cases} 1, & \hat{\lambda}^0 \ge 0, \\ \displaystyle\prod_{j=1}^{2} \prod_{i \le N_j(t)} \frac{\left(\hat{h}_{ij}^{0}\right)^{d_{ij}} \left(1 - \hat{h}_{ij}^{0}\right)^{r_{ij} - d_{ij}}}{\left(\frac{d_{ij}}{r_{ij}}\right)^{d_{ij}} \left(1 - \frac{d_{ij}}{r_{ij}}\right)^{r_{ij} - d_{ij}}}, & \hat{\lambda}^0 < 0. \end{cases}$$

This is exactly (7). We use the simplified notation $\hat{h}_{ij}$ and $\hat{\lambda}$ to replace $\hat{h}_{ij}^{0}$ and $\hat{\lambda}^0$, respectively.

Another version of (7) will be used in the proof of Theorem 1: replacing $\hat{\lambda}^0 \geq 0$ and $\hat{\lambda}^0 < 0$ in (7) by $\hat{S}_1(t) \leq \hat{S}_2(t)$ and $\hat{S}_1(t) > \hat{S}_2(t)$, respectively. This version is based on the equality of the events $\hat{\lambda}^0 < 0$ and $\hat{S}_1(t) > \hat{S}_2(t)$, which can be seen by noting that $a(\lambda)$ is strictly increasing, $a(\hat{\lambda}^0) = 1$ and $a(0) = \hat{S}_1(t)/\hat{S}_2(t)$.

Appendix B: Proof of Theorem 1

We will need the following lemma giving an asymptotic expansion of the localized EL statistic in terms of $\hat{S}_1(t)$ and $\hat{S}_2(t)$.

Lemma 2.

$$-2\log\mathcal{R}(t) = \frac{n}{\hat{\sigma}^2(t)}\left\{\log\hat{S}_1(t) - \log\hat{S}_2(t)\right\}^2 I\left\{\hat{S}_1(t) > \hat{S}_2(t)\right\} + O_p(n^{-1/2}),$$

where the O_p term holds uniformly in t over $[t_1, t_2]$.

Proof. We first find the asymptotic order of $\hat{\lambda}(t)$ uniformly for $t \in [t_1, t_2]$, then we derive an asymptotic expansion of $\hat{\lambda}(t)$ uniformly for $t \in [t_1, t_2]$. Next, by a Taylor series expansion, we approximate $-2\log\mathcal{R}(t)$ as a function of $\hat{\lambda}(t)$. Based on the two expansions, we obtain the desired result.

First, we find the asymptotic order of the Lagrange multiplier $\hat{\lambda}(t)$. Since $\hat{\lambda}(t)$ comes from the numerator of the EL ratio (5), it satisfies the equality constraint (13). McKeague and Zhao [24] studied the same Lagrange multiplier derived from optimizing the nonparametric likelihood under an equality constraint on the ratio of two survival functions, so by their Lemma A.1,

$$\hat{\lambda}(t) = O_p(\sqrt{n}) \tag{16}$$

uniformly for $t \in [t_1, t_2]$.

Next we derive an asymptotic expansion of $\hat{\lambda}(t)$. The expansion is obtained by Taylor expanding the l.h.s. of

$$\sum_{i \leq N_1(t)} \log\left\{1 - \frac{d_{i1}}{r_{i1} + \hat{\lambda}(t)}\right\} - \sum_{i \leq N_2(t)} \log\left\{1 - \frac{d_{i2}}{r_{i2} - \hat{\lambda}(t)}\right\} = 0$$

and then rearranging terms. In detail, the j-th term ($j = 1, 2$) on the l.h.s., by a similar argument in Hollander et al. [20, p. 225], has the expansion

$$\log\hat{S}_j(t) + \Delta_j\hat{\lambda}(t)\frac{\hat{\sigma}_j^2(t)}{n_j} + O_p(n_j^{-1}),$$

where $\Delta_j = 1$ for $j = 1$ and -1 for $j = 2$. Combining the two terms and using $n_j/n \to p_j$ gives

$$\log\hat{S}_1(t) - \log\hat{S}_2(t) + \hat{\lambda}(t)\frac{\hat{\sigma}^2(t)}{n} + O_p(n^{-1}) = 0.$$

Rearranging the terms, we have

$$\hat{\lambda}(t) = -\frac{n}{\hat{\sigma}^2(t)} \left\{ \log \hat{S}_1(t) - \log \hat{S}_2(t) + O_p(n^{-1}) \right\}. \tag{17}$$

Next, we find an asymptotic expansion of $-2 \log \mathcal{R}(t)$ as a function of $\hat{\lambda}(t)$. We begin, based on (7), by writing $-2 \log \mathcal{R}(t)$ as

$$\begin{aligned} &-2 \sum_{j=1}^{2} \sum_{i \le N_j(t)} \left[(r_{ij} - d_{ij}) \log \left\{ 1 + \frac{\Delta_j \hat{\lambda}(t)}{r_{ij} - d_{ij}} \right\} \right] \\ &+2 \sum_{j=1}^{2} \sum_{i \le N_j(t)} \left[r_{ij} \log \left\{ 1 + \frac{\Delta_j \hat{\lambda}(t)}{r_{ij}} \right\} \right] \end{aligned}$$

times an indicator $I(\hat{\lambda}(t) < 0)$. The j-th term above, by a similar argument in Li [23, p.102], has the expansion

$$\hat{\lambda}^2(t) \sum_{i \le N_j(t)} \frac{d_{ij}}{r_{ij}(r_{ij} - d_{ij})} + O_p(n_j^{-1/2})$$

for $j = 1, 2$. Using $n_j/n \to p_j$, and the fact that $\hat{\lambda}(t) < 0$ is equivalent to $\hat{S}_1(t) > \hat{S}_2(t)$, we can combine the terms for $j = 1, 2$ and obtain

$$-2 \log \mathcal{R}(t) = \left\{ \hat{\sigma}^2(t) \frac{\hat{\lambda}^2(t)}{n} + O_p(n^{-1/2}) \right\} I \left\{ \hat{S}_1(t) > \hat{S}_2(t) \right\}.$$

This and (17) give the desired result. □

Remark. Lemma 2 shows that $-2 \log \mathcal{R}(t)$ is asymptotically equivalent to squaring the positive part of a scaled difference between the log of KM estimators from the two samples. The inclusion of only the positive part of the difference can be attributed to the stochastically ordered form of our alternative hypothesis. We have compared the small sample performance of K_n and its counterpart based on this squared difference (results not shown), and it turns out the latter tends to be too conservative.

The advantage of using the EL approach, as opposed to a test statistic derived from the first term in the expansion of Lemma 2, is that we expect higher-order accuracy [cf. 19]. This is parallel to the parametric result in which the likelihood ratio test is asymptotically equivalent to the Wald test, but the former has better higher-order accuracy [see, e.g., 25].

We now complete the proof of Theorem 1.

We first obtain the weak convergence of $-2 \log \mathcal{R}(t)$ as a process on $[t_1, t_2]$, based on Lemma 2 and large sample properties of the KM estimator. Then by a transformation of the limiting process and the continuous mapping theorem, we get the limiting distribution of K_n.

To obtain the limit process of $-2\log\mathcal{R}(t)$, we begin by finding the weak convergence of $\log\hat{S}_1 - \log\hat{S}_2$, as the asymptotic expansion of $-2\log\mathcal{R}(t)$ in Lemma 2 suggests. For each $j = 1, 2$, it has been shown [see, e.g., 1, p.191 and p.263] that

$$\sqrt{n_j}\left(\log\hat{S}_j - \log S_j\right) \xrightarrow{d} U_j$$

as $n \to \infty$ on $D[0, t_2]$, where $U_j(t)$ is a Gaussian martingale with $U_j(0) = 0$ and $\mathrm{Cov}(U_j(s), U_j(t)) = \sigma_j^2(\min(s,t))$. Therefore, under H_0, the continuous mapping theorem implies

$$\sqrt{n}\left(\log\hat{S}_1 - \log\hat{S}_2\right) \xrightarrow{d} \frac{U_1}{\sqrt{p_1}} - \frac{U_2}{\sqrt{p_2}} \equiv U, \tag{18}$$

where $U(t)$ is a Gaussian martingale with $U(0) = 0$ and $\mathrm{Cov}(U(s), U(t)) = \sigma^2(\min(s,t))$.

Next, we establish the weak convergence of $-2\log\mathcal{R}(t)$. By (18) and the continuous mapping theorem, we have

$$n\left\{\log\hat{S}_1(t) - \log\hat{S}_2(t)\right\}^2 I\left\{\hat{S}_1(t) > \hat{S}_2(t)\right\} \xrightarrow{d} U_+^2(t)$$

in $D[t_1, t_2]$, where $U_+ = \max(U, 0)$. Then by the uniform consistency of $\hat{\sigma}^2(t)$ with respect to $\sigma^2(t)$ and Slutsky's Lemma, we have

$$\frac{n}{\hat{\sigma}^2(t)}\left\{\log\hat{S}_1(t) - \log\hat{S}_2(t)\right\}^2 I\left\{\hat{S}_1(t) > \hat{S}_2(t)\right\} \xrightarrow{d} \frac{U_+^2(t)}{\sigma^2(t)}$$

in $D[t_1, t_2]$. This and Lemma 2 imply

$$-2\log\mathcal{R}(t) \xrightarrow{d} \frac{U_+^2(t)}{\sigma^2(t)} \tag{19}$$

in $D[t_1, t_2]$.

Lastly, the asymptotic null distribution of K_n is obtained as follows. First notice that

$$\frac{U(t)}{1+\sigma^2(t)} \quad \text{and} \quad B\left(\frac{\sigma^2(t)}{1+\sigma^2(t)}\right)$$

are both zero mean Gaussian processes with the same covariance function, so they have the same distribution. We then have $U_+^2(t)/\sigma^2(t)$ equal in distribution to

$$B_+^2\left(\frac{\sigma^2(t)}{1+\sigma^2(t)}\right)\frac{\left(1+\sigma^2(t)\right)^2}{\sigma^2(t)}.$$

This, together with (18) and the continuous mapping theorem, implies that $\sup_{t\in[t_1,t_2]}\{-2\log\mathcal{R}(t)\}$ converges in distribution to

$$\sup_{t\in[t_1,t_2]}\left\{B_+^2\left(\frac{\sigma^2(t)}{1+\sigma^2(t)}\right)\frac{\left(1+\sigma^2(t)\right)^2}{\sigma^2(t)}\right\}.$$

The result follows from noticing that the r.h.s. of the above is the same as

$$\sup_{x\in[x_1,x_2]}\left\{\frac{B_+^2(x)}{x(1-x)}\right\},$$

where $x_1 = b(t_1)$ and $x_2 = b(t_2)$.

Appendix C: Validating the calibration procedure

The following result justifies the approach of pre-specifying $[x_1, x_2]$ and estimating $[t_1, t_2]$, as outlined in Section 2.2.1.

Lemma 3. *Suppose S_0 is continuous. Then under H_0, for $0 < x_1 < x_2 < 1$,*

$$K_n^* \xrightarrow{d} \sup_{x\in[x_1,x_2]}\left\{\frac{B_+^2(x)}{x(1-x)}\right\},$$

provided $b^{-1}(\cdot)$ is continuous at x_1 and x_2, where K_n^* is just K_n with t_1 and t_2 replaced by $\hat{t}_1 = \max\{\hat{b}^{-1}(x_1), T_{11}, T_{12}\}$ and $\hat{t}_2 = \min\{\hat{b}^{-1}(x_2), T_{m_1 1}, T_{m_2 2}\}$, respectively.

Proof. The idea is to obtain the joint convergence of $-2\log\mathcal{R}(t)$, $\hat{t}_1$ and $\hat{t}_2$, and then to apply the continuous mapping theorem.

First, we show the weak convergence of $-2\log\mathcal{R}(t)$. We will apply (19) in the proof of Theorem 1, but we need to translate the conditions to be in terms of x_1 and x_2 instead of t_1 and t_2. Given $0 < x_1 < x_2 < 1$ at which $b^{-1}(\cdot)$ is continuous, it suffices to show that $t_1 = b^{-1}(x_1)$ and $t_2 = b^{-1}(x_2)$ satisfy the conditions $S_0(t_1) < 1$ and $S_0(t_2)G_j(t_2) > 0$ for $j = 1, 2$. To show $S_0(t_1) < 1$, we simply use $b(t_1) = x_1 > 0$, which implies $\sigma^2(t_1) > 0$ and thus $S_0(t_1) < 1$. To show $S_0(t_2)G_j(t_2) > 0$ for $j = 1, 2$, we argue by contradiction. Suppose $S_0(t_2)G_j(t_2) = 0$ for some $j = 1, 2$. Since b is continuous (by continuity of S_0) and nondecreasing, we can pick an $\epsilon < 1 - x_2$ and δ small enough such that $x_2 \le b(t_2 + \delta) < x_2 + \epsilon < 1$. Because b^{-1} is continuous at x_2, there is no "flat" of b around t_2, and thus δ can be chosen so that b is strictly increasing in $[t_2, t_2 + \delta]$. This and $S_0(t_2)G_j(t_2) = 0$ lead to $b(t_2 + \delta) = 1$, which contradicts $b(t_2 + \delta) < x_2 + \epsilon < 1$. So we have $S_0(t_2)G_j(t_2) > 0$ for $j = 1, 2$, as required.

Next, we show $\hat{t}_j \xrightarrow{P} t_j$ for $j = 1, 2$. The proof makes use of the theory of Z-estimators [see, e.g., 36, Theorem 5.9]. Let $\Psi_n(t) = \hat{b}(t) - x_1$, $\Psi(t) = b(t) - x_1$, and $\Theta = [\tau_1, \tau_2]$ such that $[t_1, t_2] \subset \Theta \subset (0, \infty)$. We already know $\Psi_n(t_1) = o_p(1)$ and $\Psi(t_1) = 0$. It suffices to show that $\sup_{t\in\Theta} |\Psi_n(t) - \Psi(t)| \xrightarrow{P} 0$ and $\inf_{t:|t-t_1|\ge\epsilon} |\Psi(t)| > 0$ for all $\epsilon > 0$. The former is implied by the uniform consistency of $\hat{\sigma}^2$ (and thus $\hat{b}$), and the latter by the continuity of b^{-1} at x_1. Therefore we have $\hat{t}_1 \xrightarrow{P} t_1$. The same argument applies to $\hat{t}_2$.

Lastly, the asymptotic null distribution of K_n^* is obtained as follows. From the weak convergence of $-2\log\mathcal{R}(t)$ and $\hat{t}_j \xrightarrow{P} t_j$ for $j = 1, 2$, we have the joint

convergence $[-2\log\mathcal{R}(t),\hat{t}_1,\hat{t}_2]^T \xrightarrow{d}[U_+^2(t)/\sigma^2(t),t_1,t_2]^T$ in $D[t_1,t_2]\times\Theta^2$ [see, e.g., 36, Theorem 18.10 (v)]. Then applying a similar argument in the last part of the proof for Theorem 1 and the continuous mapping theorem, we get the desired result. □

Acknowledgements

Computing resources for this paper came from the Extreme Science and Engineering Discovery Environment (XSEDE) supported by NSF Grant OCI-1053575. Hsin-wen Chang was partially supported by Ministry of Science and Technology of Taiwan under grant 104-2118-M-001-001. Ian McKeague was partially supported by NIH Grant R01GM095722-01 and NSF Grant DMS-1307838. The authors thank Hammou El Barmi for his helpful comments.

Supplementary Material

Supplement to "Empirical likelihood based tests for stochastic ordering under right censorship"
(doi: 10.1214/16-EJS1180SUPP; .zip). R programs implementing the procedures developed in this article are available online. Supplementary tables with simulation results under the setup of proportional hazards are also provided.

References

[1] Andersen, P. K., Borgan, Ø., Gill, R. D., and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. New York: Springer. MR1198884
[2] Andrews, D. W. K. and Guggenberger, P. (2009). Validity of subsampling and "plug-in asymptotic" inference for parameters defined by moment inequalities. *Econometric Theory*, 25:669–709. MR2507528
[3] Borgan, Ø. and Liestøl, K. (1990). A note on confidence intervals and bands for the survival function based on transformations. *Scandinavian Journal of Statistics*, 17(1):35–41. MR1062844
[4] Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press. MR2061575
[5] Canay, I. A. (2010). EL inference for partially identified models: large deviations optimality and bootstrap validity. *Journal of Econometrics*, 156(2):408–425. MR2609942
[6] Chang, H. and McKeague, I. W. (2016). Supplement to "Empirical likelihood based tests for stochastic ordering under right censorship." DOI: 10.1214/16-EJS1180SUPP.
[7] Davidov, O., Fokianos, K., and Iliopoulos, G. (2010). Order-restricted semiparametric inference for the power bias model. *Biometrics*, 66(2):549–557. MR2758835

[8] Davidov, O. and Herman, A. (2009). New tests for stochastic order with application to case control studies. *Journal of Statistical Planning and Inference*, 139(8):2614–2623. MR2523652

[9] Davidov, O. and Herman, A. (2012). Ordinal dominance curve based inference for stochastically ordered distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(5):825–847. MR2988908

[10] Dykstra, R. L. (1982). Maximum likelihood estimation of the survival functions of stochastically ordered random variables. *Journal of the American Statistical Association*, 77(379):621–628. MR0675889

[11] Einmahl, J. H. J. and McKeague, I. W. (2003). Empirical likelihood based hypothesis testing. *Bernoulli*, 9(2):267–290. MR1997030

[12] El Barmi, H. (1996). Empirical likelihood ratio test for or against a set of inequality constraints. *Journal of Statistical Planning and Inference*, 55(2):191–204. MR1423966

[13] El Barmi, H. and McKeague, I. W. (2013). Empirical likelihood based tests for stochastic ordering. *Bernoulli*, 19:295–307. MR3019496

[14] El Barmi, H. and Mukerjee, H. (2005). Inferences under a stochastic ordering constraint: the k-sample case. *Journal of the American Statistical Association*, 100(469):252–261. MR2156835

[15] Finner, H. and Strassburger, K. (2002). The partitioning principle: a powerful tool in multiple decision theory. *The Annals of Statistics*, 30(4):1194–1213. MR1926174

[16] Fleming, T. R. and Harrington, D. P. (1991). *Counting Processes and Survival Analysis*. John Wiley & Sons, Inc. MR1100924

[17] Gill, R. D. (1980). *Censoring and Stochastic Integrals*. Mathematisch Centrum. MR0596815

[18] Guyot, P., Ades, A. E., Ouwens, M. J. N. M., and Welton, N. J. (2012). Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan–Meier survival curves. *BMC Medical Research Methodology*, 12(1):9.

[19] Hall, P. and La Scala, B. (1990). Methodology and algorithms of empirical likelihood. *International Statistical Review // Revue Internationale de Statistique*, 58(2):109–127.

[20] Hollander, M., McKeague, I. W., and Yang, J. (1997). Likelihood ratio-based confidence bands for survival functions. *Journal of the American Statistical Association*, 92:215–226. MR1436110

[21] Kitamura, Y., Santos, A., and Shaikh, A. M. (2012). On the asymptotic optimality of empirical likelihood for testing moment restrictions. *Econometrica*, 80(1):413–423. MR2920761

[22] Li, G. (1995a). Nonparametric likelihood ratio estimation of probabilities for truncated data. *Journal of the American Statistical Association*, 90(431):997–1003. MR1354016

[23] Li, G. (1995b). On nonparametric likelihood ratio estimation of survival probabilities for censored data. *Statistics & Probability Letters*, 25:95–104. MR1365025

[24] McKeague, I. W. and Zhao, Y. (2002). Simultaneous confidence bands for ratios of survival functions via empirical likelihood. *Statistics & Probability Letters*, 60:405–415. MR1947180

[25] Mukerjee, R. (1994). Comparison of tests in their original forms. *Sankhyā: The Indian Journal of Statistics, Series A (1961–2002)*, 56(1):118–127. MR1343961

[26] Murphy, S. A. (1995). Likelihood ratio-based confidence intervals in survival analysis. *Journal of the American Statistical Association*, 90(432):1399–1405. MR1379483

[27] Nair, V. N. (1984). Confidence bands for survival functions with censored data: a comparative study. *Technometrics*, 26(3):265–275.

[28] Nguyen-Khac, E., Thevenot, T., Piquet, M.-A., Benferhat, S., Goria, O., Chatelain, D., Tramier, B., Dewaele, F., Ghrib, S., Rudler, M., Carbonell, N., Tossou, H., Bental, A., Bernard-Chabert, B., and Dupas, J.-L. (2011). Glucocorticoids plus *N*-acetylcysteine in severe alcoholic hepatitis. *New England Journal of Medicine*, 365(19):1781–1789.

[29] Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75(2):237–249. MR0946049

[30] Owen, A. B. (2001). *Empirical Likelihood*. Chapman & Hall/CRC, Boca Raton.

[31] Park, Y., Kalbfleisch, J. D., and Taylor, J. M. G. (2012a). Constrained nonparametric maximum likelihood estimation of stochastically ordered survivor functions. *Canadian Journal of Statistics*, 40(1):22–39. MR2896928

[32] Park, Y., Taylor, J. M. G., and Kalbfleisch, J. D. (2012b). Pointwise nonparametric maximum likelihood estimator of stochastically ordered survivor functions. *Biometrika*, 99(2):327–343. MR2931257

[33] Parzen, M. I., Wei, L. J., and Ying, Z. (1997). Simultaneous confidence intervals for the difference of two survival functions. *Scandinavian Journal of Statistics*, 24(3):pp. 309–314. MR1481417

[34] Pepe, M. S. and Fleming, T. R. (1989). Weighted Kaplan–Meier statistics: a class of distance tests for censored survival data. *Biometrics*, 45(2):497–507. MR1010515

[35] Thomas, D. R. and Grunkemeier, G. L. (1975). Confidence interval estimation of survival probabilities for censored data. *Journal of the American Statistical Association*, 70:865–871. MR0405766

[36] van der Vaart, A. W. (2000). *Asymptotic Statistics*. Cambridge Series on Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge. MR1652247

[37] Wang, Q.-H. and Jing, B.-Y. (2001). Empirical likelihood for a class of functionals of survival distribution with censored data. *Annals of the Institute of Statistical Mathematics*, 53:517–527. MR1868888

[38] Yu, W., El Barmi, H., and Ying, Z. (2011). Restricted one way analysis of variance using the empirical likelihood ratio test. *Journal of Multivariate Analysis*, 102(3):629–640. MR2755020