Annals of the Institute of Statistical Mathematics manuscript No.
(will be inserted by the editor)

Testing for uniform stochastic ordering via empirical likelihood

Hammou El Barmi and Ian W. McKeague

Received: date / Revised: date

Abstract This paper develops an empirical likelihood approach to testing for the presence of uniform stochastic ordering (or hazard rate ordering) among univariate distributions based on independent random samples from each distribution. The proposed test statistic is formed by integrating a localized empirical likelihood statistic with respect to the empirical distribution of the pooled sample. The asymptotic null distribution of this test statistic is found to have a simple distribution-free representation in terms of standard Brownian motion. The approach is extended to the case of right-censored survival data via multiple imputation. Two applications are discussed: 1) uncensored survival time data of mice exposed to radiation, and 2) right-censored time-to-infection data from a human HIV vaccine trial comparing a placebo group with a vaccine group.

Keywords distribution-free, order restricted inference, nonparametric likelihood ratio testing.

1 Introduction

The nonparametric comparison of univariate distributions is an extensive field, but empirical likelihood methods have yet to be fully exploited when order-restricted comparisons are needed. There are many types of ordering for the comparison of distributions. These include, with increasing generality, likelihood ratio ordering, uniform stochastic ordering (equivalent to hazard rate

Hammou El Barmi (corresponding author)
Department of Statistics and Computer Information Systems, Baruch College, The City University of New York, One Baruch Way New York, NY 10010, U.S.A. E-mail: hammou.elbarmi@baruch.cuny.edu

Ian W. McKeague
Department of Biostatistics, Mailman School of Public Health, Columbia University, 722 West 168th Street, 6th Floor, New York, NY 10032, U.S.A. E-mail: im2131@columbia.edu

ordering), stochastic ordering, and increasing convex ordering (of interest in economics and actuarial science); see Shaked and Shanthikumar (2006). The aim of this paper is to develop an empirical likelihood approach to testing for the presence of uniform stochastic ordering. Such ordering often arises in the engineering and biomedical sciences, for example in reliability studies, or in the comparison of survival outcomes in randomized clinical trials.

Let X_1 and X_2 be nonnegative random variables with respective distribution functions F_1 and F_2, and survival functions $\bar{F}_1$ and $\bar{F}_2$. Then X_1 is said to be *uniformly stochastically smaller* than X_2 if $t \mapsto \bar{F}_1(t)/\bar{F}_2(t)$ is nonincreasing for $t \in [0, \tau_{F_2})$, where $\tau_{F_2} \equiv \inf\{x: F_2(x) = 1\}$. We denote this by $X_1 \preceq X_2$ or $F_1 \preceq F_2$. It is well known that if F_1 and F_2 are absolutely continuous with densities f_1 and f_2, respectively, then $F_1 \preceq F_2$ is equivalent to their corresponding hazard rates being ordered: $f_1(t)/\bar{F}_1(t) \geq f_2(t)/\bar{F}_2(t)$ for all $t \geq 0$. We also note that this is equivalent to

$$P[X_1 > t + s | X_1 > t] \leq P[X_2 > t + s | X_2 > t], \quad \text{for all } s, t \geq 0.$$

That is, the conditional distribution of X_1 given that $X_1 \geq t$ is stochastically smaller than that corresponding to X_2. For this reason, uniform stochastic ordering has a more useful and practical interpretation than the classical form of stochastic ordering (Lehmann, 1955), which orders survival functions rather than hazard functions.

Uniform stochastic ordering is useful in applications in which risks change dynamically (over time), or in which the rates of extreme events (in the upper-tail of the distributions) need to be compared, as with risk assessment studies in engineering. Dykstra et al. (1991) note that uniform stochastic ordering is especially of interest in clinical trials involving competing medical treatments. Even if the corresponding survival times are stochastically ordered initially, they may not be when patients are examined at a later time. However, under uniform stochastic ordering, one of the treatments can be considered the best.

Many tests for *classical* stochastic ordering are available in the literature, including an empirical likelihood test for uncensored data developed by El Barmi and McKeague (2013), but testing for the presence of *uniform* stochastic ordering has received little attention (formal theory is only available for discrete discrete distributions). Although our present approach is also based on empirical likelihood, the formulation of the problem and the derivation of the test statistic are completely different. Dykstra et al. (1991) obtained the nonparametric maximum likelihood estimators (NPMLE) of uniformly stochastically ordered distribution functions, and derived the likelihood ratio test for equality of multinomial distributions against the alternative that they are uniformly stochastically ordered.

Ad hoc hypothesis tests for uniform stochastic ordering in k-sample right-censored data settings go back to Tarone (1975) and Tarone and Ware (1977). Versions of such tests for the comparison of counting process intensities ("trend tests") were developed by Andersen et al. (1993, p. 388), generalizing the one-sided log-rank test to the k sample setting; see also Andersen et al. (1982). A choice of of k "scores" is needed to balance the contribution of each sample

in the overall test statistic. Unfortunately, such tests are only well powered for proportional hazards alternatives, and are not guaranteed to perform well under general hazard rate ordering. As far as we know, fully nonparametric tests for uniform stochastic ordering are not available in the literature (even in uncensored settings). Our aim in the present study is to fill this gap.

Estimation under uniform stochastic ordering has received considerable attention. In the one-sample case, Rojo and Samaniego (1991) derived the NPMLE of F_1 when $F_1 \preceq F_2$ with F_2 known, continuous and strictly increasing and showed that it is inconsistent. Mukerjee (1996) studied alternative estimators that are consistent in the one- and two-sample cases. Their weak convergence properties were studied by Arcones and Samaniego (2000).

We provide k-sample tests for general uniform stochastic ordering initially in the uncensored case, and then extend to right-censored data by adapting a multiple imputation approach considered by Taylor et al. (2002). The test statistics are constructed via the empirical likelihood (EL) method (Owen, 1988, 1990), which has flexibility of being nonparametric while preserving the efficiency of likelihood-ratio-based inference. Inference based on EL has many attractive properties: estimation of variance is typically not required, the range of the parameter space is automatically respected, and confidence regions have greater accuracy than those based on the Wald approach.

EL has an extended formulation for right-censored data (Thomas and Grunkemeier, 1975), but our proposed approach of adapting the uncensored EL test statistic to the right-censored setting via multiple imputation has the advantage of being much more straightforward and transparent — the censored-data version of the EL statistic does not have an explicit expression when uniformly stochastically ordered alternatives are involved, and its asymptotic properties are not known. Einmahl and McKeague (2003) developed a localized version of EL to allow nonparametric hypothesis testing in uncensored settings, and showed via simulation studies that it outperforms (in terms of power) the corresponding Cramér–von Mises statistics for a variety of classical testing problems. However, their approach is restricted to omnibus alternatives, whereas ordered alternatives are often more useful because they can provide a more direct interpretation of the result of the test. As mentioned earlier, El Barmi and McKeague (2013) adapted the EL approach for testing the classical form of stochastic ordering in the uncensored case. There is also an extensive literature on non-EL based tests for classical stochastic ordering, including the early paper of Schmid and Trede (1996), and recent papers of Ledwina and Wyłupek (2012, 2014). As far as we know, however, (nonparametric) tests for *uniform* stochastic ordering are not yet available.

The development of the proposed test statistic and results on its asymptotic null distribution are given in Section 2. First we consider the special case of testing whether a distribution function is uniformly stochastically larger than a specified distribution function, based on a single sample. Once the theory has been developed in this one-sample case, it is relatively straightforward to extend the approach to the general k-sample setting in which all the distribution functions are unknown. We also describe a simple extension of the

proposed test to a two-sided alternative, in which the ordering can be in either direction. In the case of right-censored survival data, we provide a multiple imputation procedure that allows the proposed test to be applied directly to the imputed complete data. In Section 3 we provide critical values for the proposed test, and illustrate the method in a simulation study and in two real data examples. Discussion is provided in Section 4, and the proofs of all the results are in Section 5.

2 Localized empirical likelihood tests

2.1 Comparison with a specified distribution

Suppose we are given a random sample $X_1, X_2, \ldots, X_n$ from the cdf F, and we want to test the null hypothesis $H_0 : F = F_0$ versus $H_1 : F \prec F_0$, where F_0 is a (pre-)specified continuous cdf. Here $\prec$ denotes $\preceq$ with equality excluded.

Our approach is based on translating the problem into testing a family of "local" null hypotheses of the form

$$H_0^{x,y} : \bar{F}(x)/\bar{F}_0(x) = \bar{F}(y)/\bar{F}_0(y) \quad \text{versus} \quad H_1^{x,y} : \bar{F}(x)/\bar{F}_0(x) > \bar{F}(y)/\bar{F}_0(y),$$

where $x < y$. The local empirical likelihood procedure (at fixed $x < y$) rejects $H_0^{x,y}$ for small values of

$$\mathcal{R}(x,y) = \frac{\sup\{L(F) : \theta = \theta_0\}}{\sup\{L(F) : \theta \leq \theta_0\}}, \tag{1}$$

where $L(F)$ is the nonparametric likelihood function, the parameter of interest is $\theta = \theta(F) \equiv \bar{F}(y)/\bar{F}(x) \in [0,1]$, and its null value is $\theta_0 \equiv \bar{F}_0(y)/\bar{F}_0(x)$. Here the suprema are restricted to cdfs F that are supported by the data points, and, by convention, $\sup \emptyset = 0$ and $0/0 = 1$.

First we decompose the nonparametric likelihood function, using p_i to denote the point mass that F places at X_i, as

$$\begin{aligned} L(F) = \prod_{i=1}^{n} p_i &= \left\{ \prod_{i:X_i \leq x} p_i \right\} \left\{ \prod_{i:x<X_i\leq y} p_i \right\} \left\{ \prod_{i:X_i>y} p_i \right\} \\ &= \left\{ \prod_{i:X_i \leq x} \frac{p_i}{F(x)} \right\} \left\{ \prod_{i:x<X_i\leq y} \frac{p_i}{F(y)-F(x)} \right\} \left\{ \prod_{i:X_i>y} \frac{p_i}{\bar{F}(y)} \right\} \\ &\quad \times \quad [F(x)]^{n\hat{F}(x)} [F(y)-F(x)]^{n(\hat{F}(y)-\hat{F}(x))} [\bar{F}(y)]^{n\hat{\bar{F}}(y)}, \end{aligned}$$

where $\hat{F}$ and $\hat{\bar{F}}$ are the empirical cdf and survival functions, respectively. The terms in braces in the above expression can be maximized separately from the remaining terms (by distributing the available mass uniformly over the data points in the relevant interval), and the constraints on θ in the numerator

and denominator of (1) have no effect. Thus the terms in braces make no contribution to $\mathcal{R}(x,y)$. The remaining terms can be written as

$$[1-\theta]^{n(\hat{\bar{F}}(x)-\hat{\bar{F}}(y))}\theta^{n\hat{\bar{F}}(y)}[F(x)]^{n\hat{F}(x)}[\bar{F}(x)]^{n\hat{\bar{F}}(x)}$$

and, similarly, we only need to consider the terms involving θ, as the last two terms again cancel and make no contribution to $\mathcal{R}(x,y)$. Indeed, the terms involving $F(x)$ and $\bar{F}(x)$ can be maximized separately from those involving θ: specifying $\bar{F}(x)$ does not restrict the values of $\theta = \bar{F}(y)/\bar{F}(x)$ in $[0,1]$, since $0 \le \bar{F}(y) \le \bar{F}(x)$ is the only constraint on $\bar{F}(y)$.

We have now reduced the evaluation of $\mathcal{R}(x,y)$ to an elementary constrained optimization problem, resulting in

$$\mathcal{R}(x,y) = \frac{[1-\theta_0]^{n(\hat{\bar{F}}(x)-\hat{\bar{F}}(y))}\theta_0^{n\hat{\bar{F}}(y)}}{[1-\theta_n]^{n(\hat{\bar{F}}(x)-\hat{\bar{F}}(y))}{\theta_n}^{n\hat{\bar{F}}(y)}},$$

where $\theta_n = \hat{\theta} \wedge \theta_0$, $\hat{\theta} = \hat{\bar{F}}(y)/\hat{\bar{F}}(x)$ and we make the convention that any term raised to a zero power is set to 1. Using a second-order Taylor expansion of $\log(1+y)$ about $y=0$, it can be shown (see the proof of the theorem below) that for given $x < y$ such that $0 < \theta_0 < 1$, under $H_0^{x,y}$,

$$\begin{aligned} -2\log\mathcal{R}(x,y) &= n\hat{\bar{F}}(x)(\hat{\theta}-\theta_0)^2\left[\frac{1}{\hat{\theta}} + \frac{1}{1-\hat{\theta}}\right] I[0 < \hat{\theta} \le \theta_0] + o_p(1) \\ &\xrightarrow{d} \frac{\bar{F}_0(x)}{\theta_0(1-\theta_0)} U^2 I(U \ge 0) = Z^2 I(Z \ge 0), \end{aligned}$$

where the delta method was used to obtain $\sqrt{n}(\hat{\theta}-\theta_0) \xrightarrow{d} U = \sigma Z$, where $Z \sim N(0,1)$, $\sigma^2 = \theta_0(1-\theta_0)/\bar{F}_0(x)$, and we also applied the continuous mapping theorem. That is, the asymptotic null distribution of $-2\log\mathcal{R}(x,y)$ is chi-bar square.

To test H_0 against H_1, we introduce the integral-type test statistic:

$$\begin{aligned} T_n &= -2\int_0^\infty\int_x^\infty \log(\mathcal{R}(x,y))\, d\hat{F}(y)\, d\hat{F}(x) \\ &= -\frac{2}{n^2}\sum_{X_i<X_j} \log(\mathcal{R}(X_i,X_j)). \end{aligned}$$

Note that the above definition does not require modification when there are ties in the data because $\log\mathcal{R}(x,x) = 0$. We only use the assumption of a continuous F_0 (under which there would be no ties a.s.) to help in the derivation of the asymptotic distribution. Use of the functional delta method (in place of the standard delta method) in the above argument leads to the following result giving the asymptotic null distribution of both of these test statistics.

Theorem 1 *If F_0 is continuous and H_0 holds, then T_n converges in distribution to*

$$\int_0^1 \int_s^1 \frac{W_+\left(f(s,t)\right)^2}{f(s,t)}\, dt\, ds,$$

where W is a standard Brownian motion, $W_+ = \max(W, 0)$ and

$$f(s,t) = \frac{t-s}{(1-s)(1-t)}.$$

2.2 Uniform stochastic ordering among k distributions

In this section we study the extension of the one-sample test of Section 2.1 to the case of a uniform stochastic ordering of $k \geq 2$ unknown distributions. That is, given independent random samples with cdfs F_j, $j = 1, \ldots, k$, we wish to test

$$H_0\colon F_1 = \ldots = F_k \quad \text{versus} \quad H_1 : F_1 \preceq F_2 \preceq \cdots \preceq F_k,$$

where at least one of the uniform stochastic orderings is strict under the alternative.

For simplicity, assume that the proportion $\gamma_j = n_j/n$ of observations in the jth sample remains fixed as the total sample size $n \to \infty$, with $0 < \gamma_j < 1$, $j = 1, \ldots, k$. The localized empirical likelihood ratio is given by

$$\mathcal{R}(x,y) = \frac{\sup\left\{\prod_{j=1}^k L(F_j) : \theta_j = \theta_{j+1},\ j = 1, \ldots, k-1\right\}}{\sup\left\{\prod_{j=1}^k L(F_j) : \theta_j \leq \theta_{j+1},\ j = 1, \ldots, k-1\right\}}, \tag{2}$$

where in each supremum F_j is supported by the observations in the jth sample, and $\theta_j = \theta(F_j) \equiv \bar{F}_j(y)/\bar{F}_j(x) \in [0,1]$ for given $x \leq y$. Using the same parameterization as before, separately for each F_j, and making the same cancelation in the numerator and denominator, it suffices to maximize

$$\prod_{j=1}^k \theta_j^{n_j \hat{\bar{F}}_j(y)} [1-\theta_j]^{n_j(\hat{\bar{F}}_j(x) - \hat{\bar{F}}_j(y))} \tag{3}$$

subject to the constraint $0 < \theta_1 = \ldots = \theta_k < 1$, or $0 < \theta_1 \leq \ldots \leq \theta_k < 1$, depending on whether it is the numerator or the denominator of (2). Under the first of these constraints, (3) is maximized by $\theta_j = \hat{\theta}_0 = \hat{\bar{F}}(y)/\hat{\bar{F}}(x)$, where $\hat{\bar{F}} = \sum_{j=1}^k \gamma_j \hat{\bar{F}}_j$ is the empirical survival function of the pooled sample, whereas in the absence of any constraint it is maximized by $\theta_j = \hat{\theta}_j = \hat{\bar{F}}_j(y)/\hat{\bar{F}}_j(x)$. Under the second constraint, we are in the setting of the classical bioassay problem (see Robertson et al., 1988, p. 32), and it follows that (3) is maximized by

$$\theta_j = \tilde{\theta}_j = E_{\mathbf{w}}(\hat{\boldsymbol{\theta}}|\mathcal{I})_j,$$

here $\tilde{\boldsymbol{\theta}} = E_{\mathbf{w}}(\hat{\boldsymbol{\theta}}|\mathcal{I})$ is the weighted least squares projection of $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_k)^T$ onto $\mathcal{I} = \{\mathbf{z} \in \mathbf{R}^k : z_1 \leq z_2 \leq \ldots \leq z_k\}$, with random weights $w_j = w_j(x) =$

$\gamma_j \hat{\tilde{F}}_j(x)$. The pool-adjacent-violators algorithm can be used to compute this projection, see Robertson et al. (1988). We now have

$$\mathcal{R}(x,y) = \prod_{j=1}^{k} \left[\frac{\hat{\theta}_0}{\tilde{\theta}_j}\right]^{n_j \hat{\tilde{F}}_j(y)} \left[\frac{1-\hat{\theta}_0}{1-\tilde{\theta}_j}\right]^{n_j(\hat{\tilde{F}}_j(x)-\hat{\tilde{F}}_j(y))} \tag{4}$$

under the convention that any term raised to a zero power is set to 1. It can be shown that the asymptotic null distribution of the EL statistic $-2\log\mathcal{R}(x,y)$ is chi-bar square.

To test H_0 against H_1, we propose the integrated EL statistic of the same form as the T_n in the previous section:

$$\begin{aligned} T_n &= -2\int_0^\infty\int_x^\infty \log\mathcal{R}(x,y)\, d\hat{F}(y) d\hat{F}(x) \\ &= -\frac{2}{n^2}\sum_{Z_i<Z_j} \log(\mathcal{R}(Z_i,Z_j)) \end{aligned} \tag{5}$$

where $\{Z_i, i = 1,\ldots,n\}$ is the pooled sample. Note that this test statistic has the attractive property of being distribution-free: its finite sample null-distribution does not depend on the common distribution function in the groups (given that it is continuous).

Theorem 2 *Under H_0 and assuming that the common distribution function F is continuous,*

$$T_n \xrightarrow{d} \sum_{j=1}^{k} \gamma_j \int_0^1\int_s^1 \frac{(E_{\boldsymbol{\gamma}}[\mathbf{Z}(f(s,t))|\mathcal{I}]_j - \overline{Z}(f(s,t)))^2}{f(s,t)}\, dt\, ds, \tag{6}$$

where $\mathbf{Z} = (W_1/\sqrt{\gamma_1},\ldots,W_k/\sqrt{\gamma}_k)^T$, the W_j are independent standard Brownian motions, $\overline{Z} = \boldsymbol{\gamma}^T\mathbf{Z}$, $\boldsymbol{\gamma} = (\gamma_j)$ and f is defined in Theorem 1.

2.3 Two-sided tests

In the two-sample case, it can be of interest to test $H_0\colon F_1 = F_2$ against the alternative $H_2\colon F_1 \preceq F_2$ or $F_2 \preceq F_1$, the union of the two one-sided alternatives. The two-sided test is especially relevant when there is no *a priori* information as to the direction of the ordering, for example in comparing two medical treatments. The EL statistic in this case can be formed as a union-intersection statistic, i.e., the maximum of the two one-sided test statistics of the form T_n constructed above. Similarly, in the k sample case we can easily construct a test for the broader alternative consisting of the union of the $k!$ possible ordered alternatives.

2.4 Right-censored data

In this section we discuss how the proposed tests can be adapted to handle right-censored data using a multiple imputation approach (Taylor et al. 2002). We restrict attention to the k-sample case, but the same approach works for the one-sample case.

First consider the situation of Type I censoring: all subjects enter at baseline and are followed for a set time-period, $[0, \tau]$ say, then at the end of follow-up τ any subjects remaining at risk are right-censored (and this is the only type of censoring). The right-censored subjects can be viewed as failing in some (unknown) random order after the end of follow-up. Moreover, note that it is only the *order* in which they fail that affects the complete-data test statistic T_n (which would be available if all failure times are observed), as implied by the distribution-free property of T_n. Our proposal is simply to average T_n over all possible permutations of these unobserved survival times. An average based on Monte Carlo sampling could be used to reduce the computational cost when the rate of censoring is high. The null distribution is unchanged.

A similar idea can be used to handle the case of random right-censoring. First note that the Kaplan–Meier estimator based on the jth sample can be plugged-in to provide an estimate $\hat{\theta}_j$ of the residual survival function $\theta_j(y) = \bar{F}_j(y)/\bar{F}_j(x)$, for $y \geq x$, provided $\hat{\bar{F}}_j(x) > 0$. If we specify $\tau > 0$ such that $\hat{\bar{F}}_j(\tau) > 0$ for all $j = 1, \ldots, k$, we have that for any right-censored observation at $x < \tau$, the estimated residual survival function $\hat{\theta}_j$ (for its corresponding sample) is well-defined. Simulating from this estimated residual survival distribution produces a new "uncensored" observation (in the jth sample) if it falls in $[x, \tau)$, but otherwise a Type I censored observation. Any observations (censored or non-censored) at $x \geq \tau$ become right-censored at τ. In this way we reduce the problem to the Type I censored case discussed above. Clearly it would be important to set the value of τ as large as possible to minimize the amount of extraneous right-censoring at τ, and in practice this could be achieved by setting it slightly to the left of $\min_j \tau_j$, where τ_j is the largest uncensored observation in the jth sample. In the sequel, when we use this procedure we average the complete-data test statistic T_n over 1000 simulated "complete" data samples.

Justification for the proposed imputation procedure can be derived using a result of Akritas (1986, Theorem 2.2) on bootstrapping the Kaplan–Meier estimator. Reid (1981) had discussed a bootstrap procedure based on an iid sample from the Kaplan–Meier estimator, and Akritas pointed out that such a resampling plan is inconsistent: its limiting distribution (over a fixed follow-up period $[0, \tau]$, conditional on the data) in fact coincides with the limit of the empirical process for the underlying survival times prior to censoring, rather than that of the Kaplan–Meier estimator. On the other hand, Aktitas's result shows that resampling based on the Kaplan–Meier estimator reproduces the asymptotic behavior of the uncensored case, which is exactly what we need. In this way it can be shown (by adapting the proof of Theorem 2) that the

limit distribution of the imputed version of T_n (conditional on the data) has the same limit as T_n provided in Theorem 2.

3 Numerical results

3.1 Calibrating the tests

Table 1 gives some approximate cut-off points for the k-sample test statistic T_n, at various significance levels and for k in the range 2–5. They are obtained by simulating 10,000 samples of size 100 (in each group) from the unit-parameter exponential distribution, and computing the quantiles of the resulting test statistics.

For the two-sided statistic discussed earlier, the approximate cut-off points corresponding to the first row of Table 1 are 1.448, 1.224, 0.977 and 0.783.

3.2 Simulation study

The simulation study compares the performance of the proposed test T_n with the test S_n developed by El Barmi and McKeague (2013) for testing equality of k distributions against the (broader) alternative that they are stochastically ordered. The test S_n is a natural competitor for T_n because it is also based on empirical likelihood, and the question arises as to whether the new test has greater power than S_n under the alternative of uniform stochastic ordering that it is specfically designed to detect. For $k = 2$ we also compare T_n with the standard one-sided log-rank test, and for $k = 3$ with the (Tarone–Ware) log-rank trend test (in the form given by Andersen et al. 1993, Example V.3.5); these tests are both denoted R_n.

We specify each cdf F_j in terms of its hazard function $r_j(x)$, setting $r_1(x) = x$, $r_2(x) = xI(0 < x < 1) + axI(x \geq 1)$, and $r_3(x) = xI(0 < x < 1) + bxI(x \geq 1)$, for various choices of $a \geq 1$ (when $k = 2$) and $b \geq a$ (when $k = 3$). In each simulation run, 10,000 data sets were used to approximate the rejection probabilities at a nominal level of $\alpha = 0.05$; the critical value for T_n is taken from Table 1. In the right-censored case, the imputed versions of T_n and S_n were obtained following the procedure described in Section 2.4. The censoring distribution in each sample is taken to be Weibull with shape parameter 2, and scale parameter specified in a way that produces 25% or 50% censoring.

The results are given in Tables 2–5. All the tests are slightly anti-conservative (see the first row in each table), except that T_n becomes conservative in the censored data case (Tables 4 and 5) in which multiple imputation is applied to T_n and S_n. The proposed test T_n has substantially greater power than S_n and R_n in Tables 2 and 3 (uncensored examples). In the right-censored case, R_n has slightly greater power than T_n close to the null hypothesis. This may be due in part to the slightly conservative nature of T_n in this case, but also because the log-rank test is well adapted to censoring compared to the multiple imputation procedure that we use. We have highlighted the entries in the

Tables where T_n has the largest increase in power over S_n and R_n (the next best being indicated in a box). The censoring severely reduces the power of both T_n and R_n, as can be seen when when comparing the highlighted entries in the bottom rows of Tables 2, 4 and 5.

3.3 Survival time under exposure to radiation

Our first real data example illustrates our approach in the case of uncensored data. We consider survival time data from a tumorigenicity study in which 181 mice were exposed to radiation (Hoel, 1972). One group (99 mice) lived in a conventional laboratory environment, while the others (82 mice) lived in a germ-free environment. Clearly, it is reasonable that the life distribution of the group in the germ-free environment (F_2) should be uniformly stochastically larger than that of the group in the conventional lab setting (F_1). Indeed, the prime motivation for the original study was to examine this very question.

Figure 1 shows the empirical survival functions for the two groups, and strongly suggests the presence of stochastic ordering: the estimate of $\bar{F}_1$ falls completely below the estimate of $\bar{F}_2$. Although it would be difficult to infer *hazard rate ordering* simply from inspection of Figure 1, our test statistic $T_n = 6.219$, with a p-value of less than 10^{-5}, provides extremely strong evidence.

3.4 Time to infection in an HIV vaccine efficacy trial

Our second real data example involves right-censored survival time data from an HIV vaccine trial. We consider data from the Step Study (Duerr et al., 2012), a double-blind, Phase II study in which 3000 HIV-1-seronegative male participants were recruited and randomly assigned to either an HIV vaccine arm or a placebo arm. The aim of the study was to assess the effect of treatment on the risk of HIV infection. Data are available on the 1836 participants who finished follow-up; 172 of these were infected before the end of follow-up. The time-to-event variable was defined as time from first vaccination until estimated time of infection. Random right-censorship is present due to the staggered-entry of the participants (91% being right-censored). Kaplan–Meier estimates of the survival functions in the vaccine and placebo arms are displayed in Figure 2.

Duerr et al. examined four subgroups specified by uncircumcised or circumcised (Uncirc, Circ), and Ad5-seropositive or Ad5-seronegative (Ad5-pos, Ad5-neg). Based on adjusted-Cox model analyses, they found a vaccine to placebo hazard ratio of 0.69 in the Ad5-seropositive and uncircumcised subgroup (with a 2-sided Bonferroni adjusted p-value of 0.02). Our (unadjusted) subgroup results, on the other hand, do not reveal any evidence of a uniformly larger hazard rate in the vaccine arm (based on T_n), nor any evidence of stochastic ordering (based on S_n), see Table 5. For comparison, we carried out an analysis on the full data set, and also found no evidence of an increase in

the hazard rate in the vaccine arm: $T_n = 0.298$ (p-value 0.30), $S_n = 1.017$ (p-value 0.13); the 1-sided Kolmogorov–Smirnov (KS) test gave a p-value of 0.21. Kaplan–Meier estimates of the survival functions in the vaccine and placebo arms corresponding to the four groups are displayed in Figure 3.

4 Concluding remarks

In this paper, we have developed an empirical likelihood ratio approach for testing equality of k-distributions against the alternative that they are uniformly stochastically ordered. The proposed test has the important feature that it does not involve *smoothing*, in contrast to the standard approach of comparing plots of Ramlau-Hansen estimates (i.e., smoothed Nelson–Aalen estimates) to infer hazard rate ordering.

We showed in the uncensored case that the asymptotic null distribution of the EL test statistic is distribution-free and can be characterized in terms of a standard Brownian motion. We also showed how to extend the test to right-censored data via multiple imputation. To illustrate our results, we considered examples involving survival-times for mice in two different lab environments, and right-censored time-to-infection data from an HIV vaccine trial comparing placebo and vaccine groups. Through a simulation study, we compared the performance of the proposed test with an EL test of classical stochastic ordering developed by El Barmi and McKeague (2013), and also with the one-sided log-rank test (or the Tarone–Ware log-rank trend test when $k \geq 3$). The results indicate, on balance, that the new test has substantially greater power than all the competing tests in the uncensored case. In the right-censored case, the log-rank test has slightly greater power than the new test close to the null hypothesis, but this may be due in part to the slightly conservative nature of the new test under multiple imputation.

A careful inspection of our results shows that the theory we have developed extends in a natural way to testing H_0 against a non-monotonic alternative. Specifically, if $\preceq$ is a quasi-order (i.e., a binary relation that is reflexive and symmetric) on $\{1, 2, \ldots, k\}$ and if $(F_1, F_2, \ldots, F_k)$ is defined to be isotonic with respect $\preceq$ if $F_i \prec F_j$ whenever $i \preceq j$, then our approach can be adapted to test $H_0 : F_1 = F_2 = \ldots = F_k$ against $H_1 : (F_1, F_2, \ldots, F_k)$ isotonic with respect to $\preceq$. The only modification to the localized empirical likelihood is that $\theta_j = E_{\mathbf{w}}(\hat{\boldsymbol{\theta}}|\mathcal{I})_j$ where $\mathcal{I}$ is now the isotonic cone corresponding to $\preceq$. Examples of H_1 include the tree ordering $F_1 \prec F_i, i \geq 2$ (for which $\mathcal{I} = \{\mathbf{x} \in \mathbf{R}^k : x_1 \leq x_i, i = 2, \ldots, k\}$) and the umbrella order $F_1 \succ F_2 \ldots \succ F_{i_0} \prec F_{i_0+1} \prec \ldots \prec F_k$, where i_0 is known (for which $.\mathcal{I} = \{\mathbf{x} \in \mathbf{R}^k : x_1 \leq x_1 \ldots \leq x_{i_0-1} \leq x_{i_0} \leq x_{i_0+1} \leq \ldots \leq x_k\}$). Several algorithms exist in the literature for computing $E_{\mathbf{w}}(\hat{\boldsymbol{\theta}}|\mathcal{I})$ and they are described in Robertson and Wright (1988).

A referee asked whether an extension of the test statistic T_n to involve *three* different locations $x < y < z$ is reasonable. If such an extension is created, it would need to be constructed by constraining the ratio of the survival functions to increase simultaneously at those locations. However, that would

be unnatural because as long at the ratio increases at all pairs of locations, it is increasing, so a third location is superfluous. In addition, we would be unable to re-parameterize the EL ratio in terms of θ as in (1) in order to reduce the EL statistic to a form that makes the asymptotic theory tractable. The same referee asked whether it may be useful to introduce a weight function into the integrand of T_n. Unfortunately, however, such weighting would distort the interpretation of the uniform stochastic ordering, and also produce a non-distribution-free test statistic, creating difficulties in calibrating the test.

5 Proofs

Proof of Theorem 1. As mentioned in the discussion leading up to the statement of the theorem, a key step in the proof is to apply the functional delta method to find the limiting distribution of the process $\hat{\theta}(x,y) = \hat{\bar{F}}(y)/\hat{\bar{F}}(x)$ by representing it as a functional of the empirical survival function: $\hat{\theta} = \theta(\hat{\bar{F}})$, where $\theta = \psi \circ \phi$, $\phi(A)(x,y) = (1/A(x), A(y))$ and $\psi(A,B)(x,y) = A(x)B(y)$. Here A and B belong to appropriate domains of functions in $\mathbf{D} = D[a,b]$, and it is assumed that $0 < F_0(a) < F_0(b) < 1$. The domain of ϕ is $\mathbf{D}_\phi = \{g \in \mathbf{D} : g > 0\}$. The functions ϕ and ψ are Hadamard differentiable (ϕ tangentially to $\mathbf{D}_\phi$), with the derivative of ϕ at $A \in \mathbf{D}_\phi$ given by

$$\phi'_A(h)(x,y) = (-h(x)/A(x)^2, h(y)), \quad h \in \mathbf{D},$$

and the derivative of ψ at $(A,B) \in \mathbf{D} \times \mathbf{D}$ given by

$$\psi'_{(A,B)}(h_1,h_2)(x,y) = A(x)h_2(y) + B(y)h_1(x), \quad (h_1,h_2) \in \mathbf{D} \times \mathbf{D},$$

where $x, y \in [a,b]$. From the chain rule for Hadamard differentiable functions, θ is Hadamard differentiable tangentially to $\mathbf{D}_\phi$, with derivative at $A \in \mathbf{D}_\phi$ given by

$$\theta'_A(h)(x,y) = [\psi'_{\phi(A)} \circ \phi'_A](x,y) = \frac{A(y)}{A(x)}\left[\frac{h(y)}{A(y)} - \frac{h(x)}{A(x)}\right], \quad h \in \mathbf{D}.$$

Noting that θ'_A is defined and continuous on the whole space $\mathbf{D}$, it follows from the above display with $A = \bar{F}_0$, and applying the functional delta method (see Section 20.2 of van der Vaart, 2000), that

$$\sqrt{n}(\hat{\theta}(x,y) - \theta_0(x,y)) = \theta_0(x,y)\left[\frac{\hat{G}(x)}{\bar{F}_0(x)} - \frac{\hat{G}(y)}{\bar{F}_0(y)}\right] + o_p(1) \tag{7}$$

uniformly over $(x,y) \in [a,b]^2$ with $x \le y$, where

$$\hat{G} = \sqrt{n}(\hat{F} - F_0) = -\sqrt{n}(\hat{\bar{F}} - \bar{F}_0)$$

is the empirical process.

Let

$$T_n^* = -2\int_0^\infty\int_x^\infty \log(\mathcal{R}(x,y))\,dF_0(y)\,dF_0(x),$$

which is easier to analyze than T_n, yet is asymptotically equivalent (cf. the last paragraph of the proof of Theorem 2). For $0 < \epsilon < 1$, let $0 < a_\epsilon < b_\epsilon$ be such that $F_0(a_\epsilon) = 1 - F_0(b_\epsilon) = \epsilon/2$. We decompose T_n^* as

$$T_n^* = T_{1n} + T_{2n} + T_{3n}, \tag{8}$$

where

$$\begin{aligned}
T_{1n} &= -2\int_{a_\epsilon}^{b_\epsilon}\int_x^{b_\epsilon} \log(\mathcal{R}(x,y))\,dF_0(y)\,dF_0(x),\\
T_{2n} &= -2\int_0^{a_\epsilon}\int_x^{b_\epsilon} \log(\mathcal{R}(x,y))\,dF_0(y)\,dF_0(x),\\
T_{3n} &= -2\int_0^{\infty}\int_{\max(x,b_\epsilon)}^{\infty} \log(\mathcal{R}(x,y))\,dF_0(y)\,dF_0(x).
\end{aligned}$$

Using Theorem 4.2 of Billingsley (1968), to complete the proof of the theorem it suffices to show that for each (sufficiently small) $\epsilon > 0$,

$$T_{1n} \xrightarrow{d} \int_{\epsilon/2}^{1-\epsilon/2}\int_s^{1-\epsilon/2} \frac{W_+\left(f(s,t)\right)^2}{f(s,t)}\,dt\,ds, \tag{9}$$

and that the two remainder terms are asymptotically negligible in the sense that

$$\limsup_{n\to\infty} P(|T_{jn}| \geq \delta) \to 0 \tag{10}$$

as $\epsilon \to 0$, for $j = 2, 3$ and all $\delta > 0$.

First consider T_{1n}. Using the inequality $|\log(1+y) - y + y^2/2| \leq |y|^3/3$ when $|y| \leq 1/2$, Donsker's theorem, and (7), we have (suppressing the dependence of $\hat\theta$ and θ_0 on (x,y)) that

$$\begin{aligned}
&\sup_{a_\epsilon \leq x \leq y \leq b_\epsilon} \left| \log(\mathcal{R}(x,y)) + \frac{n\hat{\bar{F}}(x)}{2}(\hat\theta - \theta_0)^2\left[\frac{1}{\hat\theta} + \frac{1}{1-\hat\theta}\right] I[0 < \hat\theta \leq \theta_0] \right|\\
&\qquad \leq \sup_{a_\epsilon \leq x \leq y \leq b_\epsilon} \frac{n}{3}|\hat\theta - \theta_0|^3\left[\frac{1}{\hat\theta} + \frac{1}{1-\hat\theta}\right] = o_p(1).
\end{aligned}$$

Note that $\hat F = \hat\Gamma(F_0)$ and $\hat G = \hat U(F_0)$, where $\hat\Gamma$ is the empirical cdf of $F_0(X_i) \sim \mathrm{Unif}(0,1)$, $i = 1, \ldots, n$, and $\hat U(t) = \sqrt{n}(\hat\Gamma(t) - t)$ is the uniform empirical process. Thus, using (7), Donsker's theorem, Slutsky's lemma, the above display, and making the change of variables $s = F_0(x)$ and $t = F_0(y)$ in the double integral, we obtain

$$T_{1n} = \int_{\epsilon/2}^{1-\epsilon/2}\int_s^{1-\epsilon/2} \frac{(1-s)(1-t)}{t-s}\hat V^2 I\big[\hat V \leq 0\big]\,dt\,ds + o_p(1), \tag{11}$$

where

$$\hat{V}(s,t) = \frac{\hat{U}(s)}{1-s} - \frac{\hat{U}(t)}{1-t}$$

and we have made use of the identity

$$\theta_0 \left[\frac{1}{\theta_0} + \frac{1}{1-\theta_0} \right] = \frac{\bar{F}_0(y)}{F_0(y) - F_0(x)} = \frac{1-t}{t-s}.$$

Since $B = \{(1-t)W[t/(1-t)],\ t \in [0,1]\}$ is a standard Brownian bridge,

$$\frac{B(s)}{1-s} - \frac{B(t)}{1-t} \stackrel{d}{=} W(f(s,t))$$

as processes indexed by $(s,t) \in [0,1]^2$. Thus (9) follows from (11) by application of the continuous mapping theorem to $\hat{U} \stackrel{d}{\to} B$ (Donsker's theorem), since the functional

$$h \mapsto \int_{\epsilon/2}^{1-\epsilon/2} \int_{s}^{1-\epsilon/2} h^2 I(h \le 0)/f \, dt \, ds, \quad h \in D[0,1]^2,$$

is continuous when the Skorohod space $D[0,1]^2$ is equipped with the uniform norm.

The asymptotic negligibility of the remainder terms (10) is established in a similar way to analogous terms in Einmahl and McKeague (2003), and we do not include the details. □

Remark. The expectation of the non-negative limiting random variable T in the statement of the theorem is less than $\sqrt{6}/4 \approx 0.61$. Indeed, using Fubini's theorem (without loss of generality W is assumed to be jointly measurable) and repeated use of the Cauchy–Schwarz inequality,

$$\begin{aligned} ET \le (ET^2)^{1/2} &\le \int_0^1 \int_s^1 \frac{\sqrt{EW_+\left(f(s,t)\right)^4}}{f(s,t)} \, dt \, ds \\ &= \int_0^1 \int_s^1 \frac{\sqrt{3f(s,t)^2/2}}{f(s,t)} \, dt \, ds = \sqrt{6}/4. \end{aligned}$$

Proof of Theorem 2. The proof is similar to the previous one. First note that it suffices to consider T_n^*, the asymptotically equivalent version of T_n in which integration over $\hat{F}$ is replaced by integration over F_0 (the common cdf under H_0), and we can use the same decomposition of T_n as in (8). Again, the main part of the argument concerns the leading term T_{1n}, in which the range of double integration is restricted to

$$D_\epsilon = \{(x,y) \colon a_\epsilon < x \le y < b_\epsilon\}.$$

Denoting $\theta_0(x,y) = \bar{F}_0(y)/\bar{F}_0(x)$ for $x \le y$, and using a property of isotonic regression (see Robertson et al, 1988), we have

$$\max_{1\le j\le k} |\tilde{\theta}_j(x,y) - \theta_0(x,y)| \le \max_{1\le j\le k} |\hat{\theta}_j(x,y) - \theta_0(x,y)|,$$

where for clarity we now make (x,y) explicit in the notation. Consequently, by the Glivenko–Cantelli theorem, $\tilde{\boldsymbol{\theta}}(x,y) \to \boldsymbol{\theta}_0(x,y)$ uniformly over $(x,y) \in D_\epsilon$ almost surely. Also, it follows from (7) that

$$\sqrt{n}(\hat{\boldsymbol{\theta}}(x,y) - \boldsymbol{\theta}_0(x,y)) = \theta_0(x,y)\left[\frac{\hat{\mathbf{G}}(x)}{\bar{F}_0(x)} - \frac{\hat{\mathbf{G}}(y)}{\bar{F}_0(y)}\right] + o_p(1)$$

uniformly over $(x,y) \in D_\epsilon$, where $\hat{\mathbf{G}} = (\hat{G}_1, \ldots, \hat{G}_k)$ and

$$\hat{G}_j = \sqrt{n}_j(\hat{F}_j - F_0)/\sqrt{\gamma_j} = -\sqrt{n}_j(\hat{\bar{F}}_j - \bar{F}_0)/\sqrt{\gamma_j}$$

is a scaled version of the empirical process for the jth sample. Therefore, from Donsker's theorem and a similar argument to the way we represented the limiting distribution of the process $\hat{V}$ in the proof of Theorem 1, we have

$$\sqrt{n}(\hat{\boldsymbol{\theta}}(x,y) - \boldsymbol{\theta}_0(x,y)) \xrightarrow{d} \theta_0(x,y)\mathbf{Z}(\rho(x,y)) \tag{12}$$

as processes indexed by $(x,y) \in D_\epsilon$, where $\mathbf{Z}$ is as defined in the statement of the theorem and

$$\rho(x,y) = \frac{F_0(x)}{\bar{F}_0(x)} - \frac{F_0(y)}{\bar{F}_0(y)}.$$

The least squares projection is continuous in all its arguments, as well as in the weights, so it can be shown using the continuous mapping theorem that

$$\sqrt{n}[\tilde{\boldsymbol{\theta}}(x,y) - \boldsymbol{\theta}_0(x,y)] \xrightarrow{d} \theta_0(x,y)E_{\boldsymbol{\gamma}}[\mathbf{Z}(\rho(x,y))|\mathcal{I}] \tag{13}$$

as processes indexed by $(x,y) \in D_\epsilon$. To justify the above limit, also note that the weights $w_j = w_j(x)$ that appear in $\tilde{\boldsymbol{\theta}} = E_{\mathbf{w}}(\hat{\boldsymbol{\theta}}|\mathcal{I})$ converge to $\gamma_j\bar{F}_0(x)$ uniformly over $(x,y) \in D_\epsilon$ almost surely (by the Glivenko–Cantelli theorem); the factor $\bar{F}_0(x)$ is common to all the weights in the limit, so it is not needed. Also, it can be shown that, jointly with the weak convergence in (13), we have

$$\begin{aligned}\sqrt{n}(\hat{\theta}_0(x,y) - \theta_0(x,y)) &\xrightarrow{d} \theta_0(x,y)\sum_{j=1}^{k}\sqrt{\gamma_j}\,W_j(\rho(x,y)) \\ &= \theta_0(x,y)\bar{Z}(\rho(x,y))\end{aligned} \tag{14}$$

as processes indexed by $(x,y) \in D_\epsilon$.

Again suppressing the dependence on $x \le y$, now write

$$\log\left(\frac{\tilde{\theta}_j}{\hat{\theta}_0}\right) = \log\left(1 + \frac{\tilde{\theta}_j - \hat{\theta}_j}{\hat{\theta}_j}\right) - \log\left(1 + \frac{\hat{\theta}_0 - \hat{\theta}_j}{\hat{\theta}_j}\right)$$

and

$$\log\left(\frac{1-\tilde{\theta}_j}{1-\hat{\theta}_0}\right) = \log\left(1+\frac{\hat{\theta}_j-\tilde{\theta}_j}{1-\hat{\theta}_j}\right) - \log\left(1+\frac{\hat{\theta}_j-\hat{\theta}_0}{1-\hat{\theta}_j}\right).$$

Using the inequality $|\log(1+y)-y+y^2/2| \le |y|^3/3$ when $|y| \le 1/2$, and the fact that

$$\sum_{j=1}^{k} n_j \hat{\bar{F}}_j(x)\tilde{\theta}_j = \sum_{j=1}^{k} n_j \hat{\bar{F}}_j(x)\hat{\theta}_j = \sum_{j=1}^{k} n_j \hat{\bar{F}}_j(x)\hat{\theta}_0,$$

where the first equality follows from the properties of isotonic regression (Robertson et al, 1988), we get

$$\begin{aligned}
&\sup_{(x,y)\in D_\epsilon} \left| \log(\mathcal{R}(x,y)) - \frac{1}{2}\sum_{j=1}^{k} \frac{n_j \hat{\bar{F}}_j(x)}{\hat{\theta}_j(1-\hat{\theta}_j)} \left[\left(\hat{\theta}_j - \hat{\theta}_0\right)^2 - \left(\tilde{\theta}_j - \hat{\theta}_j\right)^2 \right] \right| \\
&\qquad \le \frac{1}{3} \sup_{(x,y)\in D_\epsilon} \sum_{j=1}^{k} n_j \hat{\bar{F}}_j(x) \left[|\tilde{\theta}_j - \hat{\theta}_j|^3 + |\hat{\theta}_j - \hat{\theta}_0|^3 \right] \left[\frac{1}{\hat{\theta}_j^2} + \frac{1}{(1-\hat{\theta}_j)^2} \right],
\end{aligned}$$

which tends to zero in probability from (12), (13) and (14). Also, noting that in the present version of T_{1n}, when we replace $\hat{F}$ by F_0 there is a $o_p(1)$ remainder term (see the end of the proof for justification of this step), we obtain

$$\begin{aligned}
T_{1n} &= \int_{a_\epsilon}^{b_\epsilon}\int_{x}^{b_\epsilon} \sum_{j=1}^{k} \frac{n_j \hat{\bar{F}}_j(x)}{\hat{\theta}_j(1-\hat{\theta}_j)} \left[\left(\hat{\theta}_j - \hat{\theta}_0\right)^2 - \left(\tilde{\theta}_j - \hat{\theta}_j\right)^2 \right] dF_0(y)\, dF_0(x) + o_p(1) \\
&= \sum_{j=1}^{k} \gamma_j \int_{a_\epsilon}^{b_\epsilon}\int_{x}^{b_\epsilon} \frac{\hat{\bar{F}}_j(x)}{\hat{\theta}_j(1-\hat{\theta}_j)} \left(\sqrt{n}(\hat{\theta}_j - \theta_0) - \sqrt{n}(\hat{\theta}_0 - \theta_0) \right)^2 dF_0(y)\, dF_0(x) \\
&- \sum_{j=1}^{k} \gamma_j \int_{a_\epsilon}^{b_\epsilon}\int_{x}^{b_\epsilon} \frac{\hat{\bar{F}}_j(x)}{\hat{\theta}_j(1-\hat{\theta}_j)} \left(\sqrt{n}(\tilde{\theta}_j - \theta_0) - \sqrt{n}(\hat{\theta}_j - \theta_0) \right)^2 dF_0(y)\, dF_0(x) + o_p(1) \\
&\xrightarrow{d} \sum_{j=1}^{k} \gamma_j \int_{a_\epsilon}^{b_\epsilon}\int_{x}^{b_\epsilon} \frac{\theta_0 \bar{F}_0(x)}{1-\theta_0} \left(Z_i(\rho(x,y)) - \overline{Z}(\rho(x,y)) \right)^2 dF_0(y)\, dF_0(x) \\
&\qquad - \sum_{j=1}^{k} \gamma_j \int_{a_\epsilon}^{b_\epsilon}\int_{x}^{b_\epsilon} \frac{\theta_0 \bar{F}_0(x)}{1-\theta_0} \left(Z_j(\rho(x,y)) - E_{\boldsymbol{\gamma}}[\mathbf{Z}(\rho(x,y)|\mathcal{I}]_j \right)^2 dF_0(y)\, dF_0(x)
\end{aligned}$$

and, since

$$\begin{aligned}
&\sum_{j=1}^{k} \gamma_j \left(Z_j(\rho(x,y)) - \bar{Z}(\rho(x,y)) \right)^2 - \sum_{j=1}^{k} \gamma_j \left(Z_j(\rho(x,y)) - E_{\boldsymbol{\gamma}}[\mathbf{Z}(\rho(x,y)|\mathcal{I}]_j \right)^2 \\
&= \sum_{j=1}^{k} \gamma_j \left(E_{\boldsymbol{\gamma}}[\mathbf{Z}(\rho(x,y)|\mathcal{I}]_j - \bar{Z}(\rho(x,y)) \right)^2
\end{aligned}$$

by the properties of isotonic regression, it follows that

$$T_{1n} \xrightarrow{d} \sum_{j=1}^{k} \gamma_j \int_{a_\epsilon}^{b_\epsilon} \int_{x}^{b_\epsilon} \frac{\theta_0 \bar{F}_0(x)}{1-\theta_0} \left(E_{\boldsymbol{\gamma}}[\mathbf{Z}(\rho(x,y)|\mathcal{I}]_j - \bar{Z}(\rho(x,y)) \right)^2 dF_0(y)\, dF_0(x).$$

Using the change of variables $s = F_0(x)$ and $t = F_0(y)$ in the above display, and noting that

$$\frac{\theta_0 \bar{F}_0(x)}{1-\theta_0} = \frac{(1-s)(1-t)}{t-s} = 1/f(s,t) \quad \text{and} \quad \rho(x,y) = f(s,t),$$

we conclude that

$$T_{1n} \xrightarrow{d} \sum_{j=1}^{k} \gamma_j \int_{\epsilon/2}^{1-\epsilon/2} \int_{s}^{1-\epsilon/2} \frac{\left(E_{\boldsymbol{\gamma}}[\mathbf{Z}(f(s,t))|\mathcal{I}]_j - \bar{Z}(f(s,t)) \right)^2}{f(s,t)}\, dt\, ds.$$

The step involving the replacement of $\hat{F}$ by F_0 follows by noting that, given cadlag functions F and G having bounded variation, the map $(F, G) \mapsto \int F\, dG$ is Hadamard differentiable (see Kosorok 2008, p. 238). Each of the three remainder terms that arise from replacing F_0 by $\hat{F}$ involve double integrals with respect to $\hat{F} - F_0$, so we can write them in the form of $n^{-1/2}$ times a Hadamard differentiable function of an empirical process, and using the functional delta method we conclude that each term is of order $O_P(n^{-1/2})$. □

Acknowledgements

The authors thank the associate editor and two referees for their helpful comments. The authors thank also Peter Gilbert for helpful advice on using data from the Step Study and Ørnulf Borgan for help in providing R code for the Tarone–Ware statistic. The work of Ian McKeague was partially supported by NIH Grant R01GM095722-01 and NSF Grant DMS-1307838 and the work of Hammou El Barmi was partially supported by The City University of New York through a PSC-CUNY grant and the Spanish Ministry of Economy and Competitivity and FEDER under research grant MTM2013-40941-P.

References

Akritas, M. G. (1986). Bootstrapping the Kaplan–Meier estimator. *Journal of the American Statistical Association*, 81, 1032–1038.

Andersen, P. K., Borgan, Ø., Gill, R. D. and Keiding, N. (1982). Linear nonparametric tests for comparison of counting processes, with application to censored survival data (with discussion). *International Statistical Review*, 50, 219–258.

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

Arcones, M. A., and Samaniego, F. J. (2000). On the asymptotic distribution theory of a class of consistent estimators of a distribution satisfying a uniform stochastic ordering constraint. *The Annals of Statistics*, 28, 116–1150.

Billingsley, P. (1968). *Convergence of Probability Measures.* Wiley, New York.

Duerr, A., Huang, Y., Buchbinder, S., Coombs, R. W., Sanchez, J., del Rio, C., Casapia, M., Santiago, S., Gilbert, P., Corey, L., Robertson, M. N., and for the Step/HVTN 504 Study Team (2012). Extended follow-up confirms early vaccine-enhanced risk of HIV acquisition and demonstrates waning effect over time among participants in a randomized trial of recombinant adenovirus HIV vaccine (Step Study). *The Journal of Infectious Diseases*, 206, 258–266.

Dykstra, R., Kochar, S. and Robertson, T. (1991). Statistical inference for uniform stochastic ordering in several populations. *The Annals of Statistics*, 19, 870–888.

Einmahl, J. H. J. and McKeague, I. W. (2003). Empirical likelihood based hypothesis testing. *Bernoulli*, 9, 267–290.

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

El Barmi, H. and Mukerjee, H. (2013). Consistent estimation of survival functions under uniform stochastic ordering; the k-sample case. *Submitted for publication.*

Hoel, D. G. (1972). A representation of mortality data by competing risks. *Biometrics*, 28, 475–488.

Kosorok, M. R. (2008). *Introduction to Empirical Processes and Semiparametric Inference.* Springer, New York.

Ledwina, T. and Wyłupek, G. (2012). Nonparametric tests for stochastic ordering, *TEST*, 21, 730–756.

Ledwina, T. and Wyłupek, G. (2014). Tests for first-order stochastic dominance. Preprint, http://www.impan.pl/Preprints/p746.pdf

Lehmann, E. L. (1955). Ordered families of distributions. *The Annals of Mathematical Statistics*, 26, 399–419.

Mukerjee, H. (1996). Estimation of survival functions under uniform stochastic ordering. *Journal of the American Statistical Association*, 91, 1684–1689.

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

Owen, A. B. (1990). Empirical likelihood ratio confidence regions. *The Annals of Statistics*, 18, 90–120.

Reid, N. (1981). Estimating the median survival time. *Biometrika*, 68, 601–608.

Robertson, T., Wright, F. T. and Dykstra, R. L. (1988). *Order Restricted Inferences*, Wiley, New York.

Rojo, J. and Samaniego, F. J. (1991). On nonparametric maximum likelihood estimation of a distribution uniformly stochastically smaller than a standard. *Statistics & Probability Letters*, 11, 267–271.

Rojo, J. and Samaniego, F. J. (1993). On estimating a survival curve subject to a uniform stochastic ordering constraint. *Journal of the American Statistical Association*, 88, 566–572.

Shaked, M. and Shanthikumar, J. G. (2006). *Stochastic Orders*. Springer, New York.

Schmid, F. and Trede, M. (1996). Testing for first-order stochastic dominance: a new distribution-free test. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 45, 371–380.

Tarone, R. (1975). Tests for trend in life table analysis. *Biometrika*, 62, 679–682.

Tarone, R. and Ware, J. (1977). On distribution free tests for equality of survival distributions. *Biometrika*, 64, 156–160.

Taylor, J. M. G., Murray, S. and Hsu, C.-H. (2002). Survival estimation and testing via multiple imputation. *Statistics & Probability Letters*, 58, 221–232.

Thomas, D. and Grunkemeir, G. (1975). Confidence interval estimation of survival probabilities for censored data.*Journal of the American Statistical Association*, 70, 865–871.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes with Applications to Statistics*. Springer, New York.

Table 1 Selected critical points of T_n

	Significance level α			
k	0.01	0.02	0.05	0.10
2	1.243	1.014	0.777	0.592
3	1.613	1.412	1.112	0.892
4	1.956	1.728	1.373	1.111
5	2.113	1.829	1.490	1.231

Table 2 Power comparison of T_n, S_n and R_n (log-rank), no censoring, $k = 2$

	$n_1 = n_2 = 30$			$n_1 = n_2 = 50$		
a	T_n	S_n	R_n	T_n	S_n	R_n
1.0	0.057	0.062	0.053	0.053	0.055	0.050
1.2	0.123	0.095	0.105	0.156	0.103	0.129
1.4	0.227	0.144	0.193	0.297	0.174	0.259
1.6	0.332	0.197	0.262	0.471	0.269	0.379
1.8	0.451	0.263	0.356	0.617	0.370	0.502
2.0	0.552	0.324	0.438	0.739	0.474	0.615
2.2	0.637	0.386	0.497	**0.831**	0.569	0.706
2.4	0.719	0.448	0.570	0.899	0.662	0.774
2.6	0.771	0.499	0.626	0.939	0.729	0.826
2.8	0.824	0.554	0.670	0.958	0.776	0.857
3.0	**0.863**	0.608	0.709	0.971	0.828	0.893

Table 3 Power comparison of T_n, S_n and R_n (Tarone–Ware), no censoring, $k = 3$

		$n_1 = n_2 = n_3 = 30$			$n_1 = n_2 = n_3 = 50$		
a	b	T_n	S_n	R_n	T_n	S_n	R_n
1.0	1.0	0.058	0.056	0.059	0.057	0.059	0.053
1.2	1.4	0.216	0.129	0.110	0.292	0.171	0.140
1.4	1.8	0.446	0.243	0.194	0.605	0.344	0.264
1.6	2.2	0.633	0.360	0.278	**0.819**	0.524	0.401
1.8	2.6	0.766	0.469	0.375	0.927	0.668	0.530
2.0	3.0	**0.861**	0.559	0.455	0.971	0.791	0.638
2.2	3.4	0.918	0.643	0.523	0.989	0.864	0.726
2.4	3.8	0.944	0.705	0.596	0.995	0.912	0.793
2.6	4.2	0.968	0.760	0.658	0.999	0.944	0.852
2.8	4.6	0.978	0.808	0.708	0.999	0.965	0.887
3.0	5.0	0.985	0.842	0.744	1.000	0.975	0.912

Table 4 Power comparison of T_n and S_n and R_n (log-rank), 25% censoring, $k = 2$

	$n_1 = n_2 = 30$			$n_1 = n_2 = 50$		
a	T_n	S_n	R_n	T_n	S_n	R_n
1.0	0.042	0.056	0.055	0.038	0.053	0.052
1.2	0.070	0.076	0.095	0.094	0.092	0.108
1.4	0.115	0.104	0.133	0.168	0.128	0.171
1.6	0.167	0.128	0.194	0.265	0.183	0.256
1.8	0.216	0.155	0.230	0.356	0.229	0.324
2.0	0.278	0.199	0.286	0.453	0.280	0.404
2.2	0.334	0.223	0.333	0.541	0.341	0.481
2.4	0.388	0.253	0.370	0.618	0.395	0.545
2.6	0.440	0.282	0.418	0.696	0.460	0.597
2.8	0.489	0.316	0.461	0.747	0.507	0.645
3.0	**0.542**	0.347	0.488	**0.801**	0.564	0.694

Table 5 Power comparison of T_n and S_n and R_n (log-rank), 50% censoring, $k = 2$

	$n_1 = n_2 = 30$			$n_1 = n_2 = 50$		
a	T_n	S_n	R_n	T_n	S_n	R_n
1.0	0.036	0.053	0.051	0.039	0.054	0.054
1.2	0.056	0.067	0.069	0.065	0.075	0.071
1.4	0.071	0.080	0.091	0.093	0.098	0.105
1.6	0.092	0.097	0.113	0.135	0.124	0.138
1.8	0.116	0.114	0.131	0.180	0.153	0.167
2.0	0.141	0.138	0.154	0.228	0.186	0.206
2.2	0.161	0.154	0.172	0.274	0.216	0.228
2.4	0.183	0.167	0.198	0.316	0.241	0.265
2.6	0.205	0.182	0.210	0.360	0.275	0.292
2.8	0.227	0.187	0.226	0.383	0.281	0.307
3.0	**0.256**	0.216	0.251	**0.430**	0.316	0.328

Table 6 Subgroup analyses for HIV vaccine efficacy trial

Subgroup	T_n	p-value	S_n	p-value	1-sided KS	p-value
Ad5-neg and Uncirc	0.271	0.315	1.085	0.119	0.799	0.279
Ad5-neg and Circ	0.282	0.312	0.365	0.364	0.578	0.513
Ad5-pos and Uncirc	0.277	0.317	0.942	0.137	0.818	0.262
Ad5-pos and Circ	0.312	0.295	0.892	0.147	0.822	0.259

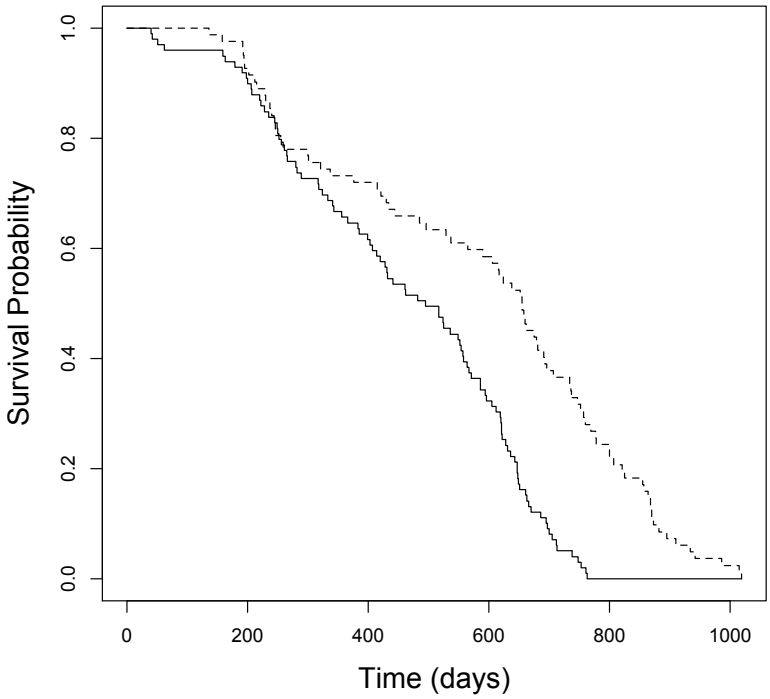

Fig. 1 Estimates of the survival functions $\bar{F}_1$ (solid line) and $\bar{F}_2$ (dashed line) for the two groups of mice.

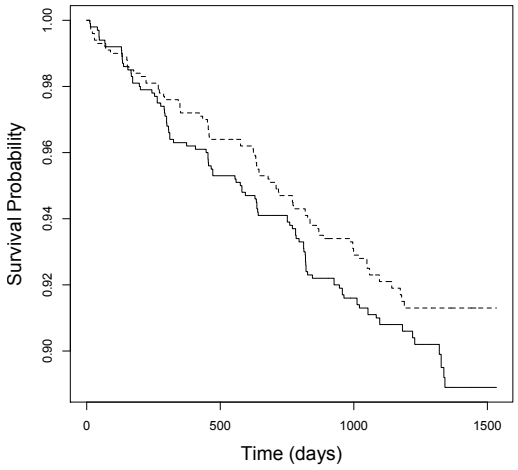

Fig. 2 Kaplan–Meier estimates of the survival functions for the placebo arm (dashed line) and the HIV vaccine arm (solid line) based on the full data set.

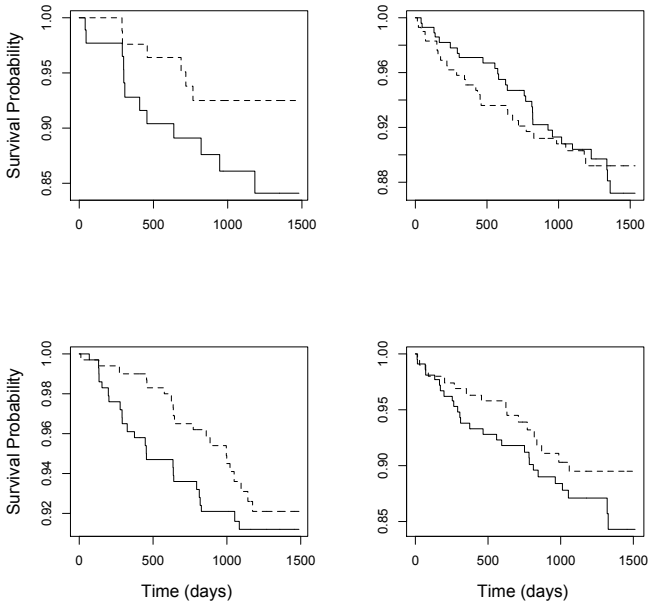

Fig. 3 Kaplan–Meier estimates of the survival functions for the four subgroups (same order as Table 4); the dashed lines represent the placebo arm and the solid lines the HIV vaccine arm; the Ad5-seropositive and uncircumcised subgroup is on the lower left.