The Annals of Statistics
2019, Vol. 47, No. 1, 205–232
https://doi.org/10.1214/18-AOS1686

NONPARAMETRIC TESTING FOR MULTIPLE SURVIVAL FUNCTIONS WITH NONINFERIORITY MARGINS

BY HSIN-WEN CHANG[1] AND IAN W. MCKEAGUE[2]

Academia Sinica and Columbia University

New nonparametric tests for the ordering of multiple survival functions are developed with the possibility of right censorship taken into account. The motivation comes from noninferiority trials with multiple treatments. The proposed tests are based on nonparametric likelihood ratio statistics, which are known to provide more powerful tests than Wald-type procedures, but in this setting have only been studied for pairs of survival functions or in the absence of censoring. We introduce a novel type of pool adjacent violator algorithm that leads to a complete solution of the problem. The limit distributions can be expressed as weighted sums of squares involving projections of certain Gaussian processes onto the given ordered alternative. A simulation study shows that the new procedures have superior power to a competing combined-pairwise Cox model approach. We illustrate the proposed methods using data from a three-arm noninferiority trial.

1. Introduction. This article introduces new nonparametric tests for the ordering of multiple survival functions. The motivation comes from applications to noninferiority trials with time-to-event outcomes and multiple competing treatments. The aim of noninferiority testing is to demonstrate that the efficacy of an experimental treatment is within a tolerable margin of a standard treatment. Such trials have become increasingly popular in recent years because, as safer and less costly therapies become available, there is a need to show that they are no less effective than reference therapies [Wellek (2010), Rothmann, Wiens and Chan (2011)].

Let $S_1, \ldots, S_k$ be unknown survival functions corresponding to $k \geq 2$ treatments. In noninferiority trials, it is of interest to establish orderings such as

$$H_1 \colon S_1^{M_1} \succ S_2^{M_2} \succ \cdots \succ S_k^{M_k}, \tag{1.1}$$

or

$$H_2 \colon S_1^{M_1} \succ S_j^{M_j} \qquad \text{for all } j = 2, \ldots, k, \tag{1.2}$$

Received May 2017; revised November 2017.

[1]Supported in part by Ministry of Science and Technology of Taiwan Grants 105-2118-M-001-012 and 106-2118-M-001-015-MY3.

[2]Supported in part by NSF Grant DMS-1307838 and NIH Grant R01GM095722.

MSC2010 subject classifications. Primary 62G10; secondary 62G20, 62F30.

Key words and phrases. Empirical likelihood, equivalence test, order-restricted inference.

where $M_1, \ldots, M_k > 0$ are the prespecified margins (which are informed by regulatory guidelines and previous clinical studies). Here, for functions $f(t)$ and $g(t)$ of t over some given interval, we define $f \succ g$ to mean $f(t) \geq g(t)$ for all t with a strict inequality for some t; the time domain is restricted to a given follow-up period $[t_1, t_2]$, as is conventional with simultaneous inference in censored data settings [see, e.g., Borgan and Liestøl (1990), Parzen, Wei and Ying (1997)]. The ordering $S_1^{M_1} \succ S_2^{M_2}$ would result from $M_1 h_1 \prec M_2 h_2$, where h_j is the hazard function of S_j. In this sense, M_1/M_2 represents a tolerable margin for the hazard ratio between treatments 1 and 2; note that this ratio is the only relevant quantity in the comparison between these two treatments, but to keep the notation balanced we preserve both M_1 and M_2. If $M_j \equiv 1$, then (1.1) and (1.2) reduce to simple linear and tree stochastic orderings, respectively [Chi (2002), El Barmi and Mukerjee (2005)]. When shorter survival is desirable (which is the case in our real data example in Section 7), $M_1 \geq M_2$ represents superiority of treatment 2 over 1, otherwise noninferiority of 2 over 1.

The ordering in (1.1) is of special interest in the case of a three-arm noninferiority clinical trial ($k = 3$), where S_1 represents a placebo, S_2 a standard therapy and S_3 an experimental therapy. In this case, (1.1) consists of two parts: (a) noninferiority $S_2^{M_2} \succ S_3^{M_3}$, where $M_2 < M_3$ indicates noninferiority of the experimental to the standard therapy, and (b) assay sensitivity $S_1^{M_1} \succ S_2^{M_2}$, where $M_1 \geq M_2$ indicates superiority of the standard therapy over the placebo. Here, (b) indicates that the trial is capable of differentiating between effective and ineffective treatments, which is required to ensure the quality of the whole study [Hauschke and Pigeot (2005)]. In (1.2), S_1 may refer to a standard therapy, and the remaining S_j to experimental therapies [see, e.g., White et al. (2011), for uncensored outcomes]. The ratio of the margins M_1/M_j could be specified as 0.8, for example, which would mean that the largest tolerable decrease in the hazard of any experimental treatment (over the standard treatment) is 25%. There are variants of (1.2), such as both $S_1^{M_1} \succ S_3^{M_3}$ and $S_2^{M_2} \succ S_3^{M_3}$, which would be useful in comparing the experimental treatment to the placebo and the standard therapy in a three-arm noninferiority clinical trial. In all of the above settings, a nonparametric approach has not been developed, except for $k = 2$ using a comparison of Kaplan–Meier (KM) estimates [Freitag, Lange and Munk (2006)].

In (1.1) and (1.2), we focus on the case of prespecified (i.e., absolute) margins to formalize the noninferiority of the experimental treatment(s) to the reference treatment. Absolute margins are usually given a priori in the study protocol [see, e.g., Hida and Tango (2011), for normally distributed endpoints]. As an alternative to the absolute-margin method, the noninferiority margins can be specified in terms of the relative effect of the standard therapy to the placebo. This is known as the *retention-of-effect* method [see, e.g., Mütze et al. (2017), for a nonparametric approach for uncensored data], which for right-censored endpoints has been developed by Kombrink, Munk and Friede (2013) using the Cox model with treatment effects measured on the log-relative-risk scale. Our proposed approach can

be adapted to provide a retention-of-effect analysis as well, namely in terms of log-relative survival (see Sections 5.1 and 7). For some background on the use of relative survival and its relation to the analysis of hazard differences in noninferiority trials, see Dignam et al. (2009), Dickman et al. (2004), McDaniel, Yu and Chappell (2016).

Our main contribution is to construct nonparametric likelihood ratio (NPLR) tests for ordered alternative hypotheses of the form (1.1) and (1.2). Existing approaches to noninferiority testing without parametric assumptions are all based on Wald-type procedures. We use the classic nonparametric likelihood for right-censored data [Thomas and Grunkemeier (1975), Li (1995)] that inspired the broad class of empirical likelihood methods [Owen (1988)], and is known to have greater power than Wald-type procedures [see, e.g., Kitamura, Santos and Shaikh (2012)]. Empirical likelihood has been used for superiority testing (stochastic ordering) for uncensored data [El Barmi and McKeague (2013)], but the right-censored case has yet to be studied, except when $k = 2$ and $M_j \equiv 1$ [Chang and McKeague (2016)].

Numerical methods for computing the NPLR are readily available (e.g., Newton–Raphson-type algorithms). However, when $k \geq 3$, it is difficult to find the asymptotic distributions of the test statistics when the NPLR is expressed as a limit of such an algorithm. The central difficulty for the asymptotic theory is that evaluating the NPLR involves solving two constrained optimization problems having no closed-form solutions—there are $2(k-1)$ time-dependent Lagrange multipliers having complex dependencies. To solve this problem, we introduce a novel type of pool adjacent violator algorithm (PAVA) to characterize the NPLR, leading to a tractable way of analyzing the Lagrange multipliers and hence obtaining the asymptotic distribution of the test statistics.

Our approach to testing H_1 specified in (1.1) is based on partitioning the parameter space for $(S_1, \ldots, S_k)$ into $H_{01} \cup H_{01}^c$, where $H_{01} = H_0 \cup H_1$ and

$$H_0 \colon S_1^{M_1} = \cdots = S_k^{M_k}. \tag{1.3}$$

We propose a two-step procedure: test the null H_{01}^c versus H_{01}, then H_0 versus H_1. Rejection of both of these null hypotheses gives support for H_1 versus the overall null $H_1^c = H_{01}^c \cup H_0$. The development of our NPLR method in the sequel is aimed at the second test. The first test is more standard, in that it can be based on a simultaneous confidence tube for the $k-1$ functions $M_j \log S_j(t) - M_{j+1} \log S_{j+1}(t)$; an analogous method was introduced in Chang and McKeague (2016). We will show in Section 3.2 that the family-wise error rate of the proposed two-step procedure can be controlled at the same alpha-level as the individual tests. The case of testing H_2 in (1.2) is similar.

For a competing method to testing H_1, we consider the combination of all $k-1$ pairwise tests of the form

$$H_0^j \colon S_j^{M_j} \preceq S_{j+1}^{M_{j+1}} \quad \text{versus} \quad H_1^j \colon S_j^{M_j} \succ S_{j+1}^{M_{j+1}}, \tag{1.4}$$

with $j = 1, \ldots, k-1$. The combined procedure tests that at least one of the H_0^j holds versus the alternative that all of the H_1^j hold (i.e., H_1). We can combine our pairwise NPLR tests to formulate such a test of H_1, using the intersection-union principle which does not require multiplicity adjustment [see, e.g., Berger and Hsu (1996)]. We will show via simulations that this competing procedure is less powerful than our two-step test. On the other hand, multiplicity adjustment can be employed if (in addition to testing H_1) there is interest in testing the individual alternatives H_1^j. Such a procedure for $k = 3$ has been suggested by Kombrink, Munk and Friede (2013) using Cox model-based Wald-type tests, although their individual alternatives are based on relative risk instead of relative survival as in our case. We can also conduct a multiplicity-adjusted version of our pairwise NPLR tests for each H_1^j. Note that the aforementioned procedures do not take into account the possibility of crossings between $S_j^{M_j}$ and $S_{j+1}^{M_{j+1}}$, whereas our proposed two-step test mentioned above does not ignore this possibility.

The paper is organized as follows. In Section 2, we construct the NPLR and characterize it using a new type of PAVA. In Section 3, we develop the proposed two-step procedures and describe the competing combined-pairwise NPLR tests. A by-product of our methodology is given in Section 3.4: a new NPLR approach to the classical k-sample omnibus test. In Section 4, we outline the proofs of our asymptotic results using the characterization of the NPLR in terms of PAVA. Section 5 provides extensions of our approach to retention-of-effect analysis and stratified testing. Section 6 presents simulation results showing that the proposed two-step NPLR tests perform better than the competing methods based on the combination of pairwise tests, in both accuracy and power. Section 7 provides an application of the proposed methods to a three-arm noninferiority trial. Concluding remarks are given in Section 8.

2. Nonparametric likelihood ratio construction.

2.1. *Background.* First, we recall the standard one-sample right-censored framework. The survival times are i.i.d. from an unknown continuous survival function S, and we assume independent censoring with survival function G. Suppose $S(t_2)G(t_2) > 0$, where t_2 is the end of follow-up as mentioned in the Introduction. Only the minimum of the survival and censoring times is observed, along with the censoring indicator, in a sample of size n. The nonparametric likelihood [Thomas and Grunkemeier (1975)] is proportional to

$$L(S) \equiv \prod_{i=1}^{m} h_i^{d_i} (1 - h_i)^{r_i - d_i} \tag{2.1}$$

for $h_i \in [0, 1]$, where r_i is the number at risk just before T_i, d_i is the number of deaths at T_i and h_i is the hazard probability at T_i, $0 < T_1 < \cdots < T_m < \infty$ are

the ordered observed uncensored times (see Owen [2001, page 142], for more detail). The NPMLE for $S(t)$ is the KM estimator $\hat{S}(t) = \prod_{i \le N(t)}(1 - d_i/r_i)$, where $N(t)$ is the number of observed uncensored times that are less than or equal to t. It is known that $\sqrt{n}\{\hat{S}(\cdot) - S(\cdot)\}$ converges weakly in $\ell^\infty([t_1, t_2])$ to a mean-zero Gaussian process with covariance function $S(s)S(t)\sigma^2(\min(s,t))$, where $\sigma^2(t) = -\int_0^t dS(u)/\{S^2(u)G(u-)\} > 0$ and $\ell^\infty([t_1, t_2])$ is the space of all bounded functions (on $[t_1, t_2]$) endowed with the supremum norm. A uniformly consistent estimate of $\sigma^2(t)$ is given by $\hat{\sigma}^2(t) = n\sum_{i \le N(t)}[d_i/\{r_i(r_i - d_i)\}]$. Application of the functional delta method then gives $\hat{U}(t) = \sqrt{n}\{\log \hat{S}(t) - \log S(t)\}/\hat{\sigma}(t) \xrightarrow{d} U(t)$ in $\ell^\infty([t_1, t_2])$, where $U(t)$ is a mean-zero Gaussian process with covariance $\sigma^2(\min(s,t))/\{\sigma(s)\sigma(t)\}$.

For our setting of k independent samples as in the Introduction, we use the same notation except with a further subscript j indicating the jth sample. The Gaussian processes U_j, $j = 1, \ldots, k$, are independent by assumption. The proportion $p_j \equiv n_j/n > 0$ of data in the jth sample is assumed to be fixed, where $n = n_1 + \cdots + n_k$ is the total sample size; this assumption can naturally be relaxed to $n_j/n \to p_j > 0$ as $n \to \infty$. For future reference, we define the time-varying group-specific weights $w_j(t) \propto 1/\theta_j(t)$, where $\theta_j(t) = M_j^2\sigma_j^2(t)/p_j$ can be estimated by $\hat{\theta}_j(t) \equiv nM_j^2\hat{\sigma}_j^2(t)/n_j$. The weights $w_j(t)$ are normalized to sum to 1. That is, $w_j(t) = \prod_{g \in E_j}\theta_g(t)/\phi(t) > 0$, where $E_j = \{1, \ldots, k\} \setminus \{j\}$ and $\phi(t) = \sum_{l=1}^k \prod_{g \in E_l}\theta_g(t)$. In the uncensored case and $M_j \equiv 1$, the weights are just $w_j(t) \propto p_j S_j(t)/(1 - S_j(t))$. These weights play an important role in the sequel because asymptotically the PAVA solution is represented by a weighted least squares projection related to the particular ordered alternative.

2.2. *Nonparametric likelihood ratio for ordered survival functions.* This section develops the proposed test statistic of H_0 versus H_1, where H_0 is given in (1.3). The statistic for the alternative of tree ordering H_2 is similar. We start by constructing the local NPLR at a given time point t:

$$\mathcal{R}(t) = \frac{\sup\{L(S_1, \ldots, S_k)\colon S_1^{M_1}(t) = \cdots = S_k^{M_k}(t)\}}{\sup\{L(S_1, \ldots, S_k)\colon S_1^{M_1}(t) \ge \cdots \ge S_k^{M_k}(t)\}}, \tag{2.2}$$

where $L(S_1, \ldots S_k)$ is the product of the k marginal nonparametric likelihoods as in (2.1), under the conventions $\sup \varnothing = 0$ and $0/0 = 1$. By the method of Lagrange multipliers, the numerator of (2.2) is the constrained maximum $\prod_{j=1}^k \prod_{i=1}^{m_j} \bar{h}_{ij}^{d_{ij}}(1 - \bar{h}_{ij}^{d_{ij}})$, where

$$\bar{h}_{ij} = \frac{d_{ij}}{r_{ij} + M_j(\bar{\lambda}_j - \bar{\lambda}_{j-1})}, \tag{2.3}$$

for $i \le N_j(t)$, $\bar{h}_{ij} = d_{ij}/r_{ij}$ for $i > N_j(t)$, and the multipliers $\bar{\lambda}_1, \ldots, \bar{\lambda}_{k-1}$ ($\bar{\lambda}_0 \equiv \bar{\lambda}_k \equiv 0$) satisfy the equality constraints

$$\prod_{i \le N_{j+1}(t)} (1 - \bar{h}_{i,j+1})^{M_{j+1}} = \prod_{i \le N_j(t)} (1 - \bar{h}_{ij})^{M_j} \tag{2.4}$$

for $j = 1, \ldots, k-1$. On the other hand, the inequality constraints in the denominator of $\mathcal{R}(t)$ can be handled using the Karush–Kuhn–Tucker (KKT) method [Boyd and Vandenberghe (2004)]. The denominator of (2.2) is then $\prod_{j=1}^{k} \prod_{i=1}^{m_j} \tilde{h}_{ij}^{d_{ij}} (1 - \tilde{h}_{ij}^{d_{ij}})$, where

$$\tilde{h}_{ij} = \frac{d_{ij}}{r_{ij} + M_j(\tilde{\lambda}_j - \tilde{\lambda}_{j-1})}, \tag{2.5}$$

for $i \le N_j(t)$, $\tilde{h}_{ij} = d_{ij}/r_{ij}$ for $i > N_j(t)$, and the multipliers $\tilde{\lambda}_1, \ldots, \tilde{\lambda}_{k-1}$ ($\tilde{\lambda}_0 \equiv \tilde{\lambda}_k \equiv 0$) satisfy the conditions

$$\prod_{i \le N_{j+1}(t)} (1 - \tilde{h}_{i,j+1})^{M_{j+1}} \le \prod_{i \le N_j(t)} (1 - \tilde{h}_{ij})^{M_j}, \tag{2.6a}$$

$$\tilde{\lambda}_j \ge 0, \tag{2.6b}$$

$$\tilde{\lambda}_j \left\{ \prod_{i \le N_{j+1}(t)} (1 - \tilde{h}_{i,j+1})^{M_{j+1}} - \prod_{i \le N_j(t)} (1 - \tilde{h}_{ij})^{M_j} \right\} = 0. \tag{2.6c}$$

Intuitively, the denominators in (2.3) and (2.5) can be seen as adjustments to the size of the risk set (r_{ij}), in order to satisfy the equality and inequality constraints. After canceling out $\bar{h}_{ij} = d_{ij}/r_{ij}$ and $\tilde{h}_{ij} = d_{ij}/r_{ij}$ for $i > N_j(t)$, the local NPLR simplifies to

$$\mathcal{R}(t) = \prod_{j=1}^{k} \prod_{i \le N_j(t)} \frac{\bar{h}_{ij}^{d_{ij}} (1 - \bar{h}_{ij})^{r_{ij} - d_{ij}}}{\tilde{h}_{ij}^{d_{ij}} (1 - \tilde{h}_{ij})^{r_{ij} - d_{ij}}}. \tag{2.7}$$

Throughout this section, we have suppressed the dependence of $\bar{\lambda}_j$, $\bar{h}_{ij}$, $\tilde{\lambda}_j$ and $\tilde{h}_{ij}$ on t. In the special case of $k = 2$, there are only two multipliers in effect: $\bar{\lambda}_1$ and $\tilde{\lambda}_1$. It can be shown that $\tilde{\lambda}_1 = \bar{\lambda}_1$ if $\bar{\lambda}_1 \ge 0$, and $\tilde{\lambda}_1 = 0$ otherwise, which simplifies (2.7) to the one-sided form of the local NPLR in Chang and McKeague (2016), equation (7) for $M_1 = M_2 = 1$. When $k \ge 3$ the relationship among the (now more than two) Lagrange multipliers becomes much more complicated, so (2.7) no longer has a simple one-sided form.

Our approach of localizing the NPLR to a specific t will make the asymptotic theory tractable; as far as we know, no asymptotic distribution theory is available for an NPLR under the global ordering restriction $S_1 \succeq \cdots \succeq S_k$. An alternative way of approaching the constrained optimization problem in (2.2) is to use Fenchel

duality. This can be done by adapting the method of Dykstra and Feltz (1989) to apply to constraints at a given t and allowing general M_j. However, it seems difficult to obtain asymptotic results for the local NPLR based on Fenchel duality and its associated algorithm, and for that reason we develop the pool adjacent violator algorithm described in the next section.

2.3. *Pool adjacent violator algorithm.* At first sight, an asymptotic analysis of the denominator of $\mathcal{R}(t)$ in (2.7) appears challenging because of the $k-1$ inequality constraints in (2.6). In order to get around this difficulty, we establish equality between the solution of equations (2.5), (2.6) and that of a pool adjacent violator-type algorithm (PAVA) (see Appendix A). Here, instead of just taking weighted averages over adjacent blocks B_1 and B_2 of sample indices that violate the order constraints, as in standard PAVA, the *pooling* for $B \equiv B_1 \cup B_2$ involves solving the equality constrained optimization problem

$$\begin{aligned} \check{\boldsymbol{h}} &\equiv \arg\max_{\boldsymbol{h}} \sum_{j\in B}\sum_{i=1}^{m_j}[d_{ij}\log h_{ij} + (r_{ij}-d_{ij})\log(1-h_{ij})] \\ &\text{subject to } \prod_{i\le N_j(t)}(1-h_{ij})^{M_j} = \prod_{i\le N_l(t)}(1-h_{il})^{M_l}, \qquad j,l\in B, \end{aligned} \tag{2.8}$$

where $\boldsymbol{h} = (h_{ij})$, $i = 1,\dots,m_j$, $j \in B$. This solution takes the same form as (2.3) except that j is restricted to belong to B. The order constraints are *violated* if

$$\prod_{i\le N_j(t)}(1-\check{h}_{ij})^{M_j} < \prod_{i\le N_l(t)}(1-\check{h}_{il})^{M_l}$$

for all $j \in B_1$ and $l \in B_2$ when B_1 precedes B_2. Our PAVA (see Algorithm 1) starts by setting up k singleton blocks of sample indices, $\{1\},\dots,\{k\}$. The algorithm proceeds by sweeping through the blocks from the left until an adjacent pair of violators is found, and then pooling this pair of blocks. This process is repeated until no adjacent violators are found, which occurs after finitely many iterations. The main challenge in connecting this PAVA solution with the denominator of the local NPLR is to show that the Lagrange multipliers coming from PAVA are nonnegative [i.e., agreeing with (2.6b)], while it is relatively easy to show agreement of the PAVA solution with (2.5) and the remaining parts of (2.6). Uniqueness of the maximizer of the denominator of the local NPLR (as shown in Section 4.1) then completes the proof.

3. Noninferiority testing.

3.1. *Test statistics for simple linear ordering.* To test H_0 versus the alternative of simple linear ordering H_1, we propose the following maximally selected and integrated local NPLR statistics:

$$K_n = \sup_{t\in[t_1,t_2]}\big[-2\log\mathcal{R}(t)\big], \qquad I_n = \int_{t_1}^{t_2}\big[-2\log\mathcal{R}(t)\big]\,d\hat{F}_0(t), \tag{3.1}$$

Algorithm 1: PAVA for NPLR at t

Input : right-censored data from k groups
Output : $\mathcal{R}(t)$, $\mathcal{B}(t)$
Initialize: $q = 0$; $\mathcal{B}^{(0)} = \{\{1\}, \ldots, \{k\}\}$; $\check{h}_{ij}^{(0)} = d_{ij}/r_{ij}$ for all i, j

while *there exists a pair of adjacent violators in* $\mathcal{B}^{(q)}$ **do**
  let $B_1, B_2 \in \mathcal{B}^{(q)}$ be the first such pair
  solve (2.8) with $B = B_1 \cup B_2$
  $\check{h}_{ij}^{(q)} = \check{h}_{ij}, i = 1, \ldots, m_j, j \in B$
  $\check{h}_{ij}^{(q)} = \check{h}_{ij}^{(q-1)}, i = 1, \ldots, m_j, j \notin B$
  $\mathcal{B}^{(q+1)} = \mathcal{B}^{(q)} \cup \{B\} \setminus \{B_1, B_2\}$
  $q = q + 1$
end
replace $\tilde{h}_{ij}$ by $\check{h}_{ij}^{(q)}$ in (2.7) to get the denominator of $\mathcal{R}(t)$
solve (2.8) with $B = \{1, \ldots, k\}$ to get the numerator of $\mathcal{R}(t)$
$\mathcal{B}(t) = \mathcal{B}^{(q)}$.

where $\hat{F}_0(t) = 1 - \hat{S}_0(t)$, $\hat{S}_0(t)$ is a consistent estimate of the survival function $S_0(t) = \sum_{j=1}^k v_j(t) S_j^{M_j}(t)$ under H_0, and $[t_1, t_2]$ is as discussed in the Introduction. We take $v_j(t)$ to be inversely proportional to the asymptotic standard deviation of $\hat{S}_j^{M_j}(t)$ and normalized so that $\sum_{j=1}^k v_j(t) = 1$. Our first result gives the asymptotic null distributions of K_n and I_n (see Section 4.2 and Appendix B for the proof), as expressed in terms of the time-varying weights $w_1(t), \ldots, w_k(t)$ and the independent Gaussian processes $U_1(t), \ldots, U_k(t)$ defined in Section 2.1.

THEOREM 1. *Under* H_0, *if* $1 > S_0(t_1) > S_0(t_2) > 0$, *we have*

$$K_n \xrightarrow{d} \sup_{t \in [t_1, t_2]} \mathrm{SSB}(t), \qquad I_n \xrightarrow{d} \int_{t_1}^{t_2} \mathrm{SSB}(t)\, dF_0(t),$$

where $F_0(t) = 1 - S_0(t)$,

$$\mathrm{SSB}(t) = \sum_{j=1}^k w_j(t) \big[E_{\boldsymbol{w}}(\boldsymbol{U}_{\boldsymbol{w}}(t) | \mathcal{I})_j - \bar{U}(t) \big]^2,$$

$\boldsymbol{U}_{\boldsymbol{w}}(t) = [U_1(t)/\sqrt{w_1(t)}, \ldots, U_k(t)/\sqrt{w_k(t)}]^T$, $\bar{U}(t) = \sum_{l=1}^k \sqrt{w_l(t)} U_l(t)$, *and the* k*-vector* $E_{\boldsymbol{w}}(\boldsymbol{U}_{\boldsymbol{w}}(t) | \mathcal{I})$ *is the weighted least squares projection of* $\boldsymbol{U}_{\boldsymbol{w}}(t)$ *onto* $\mathcal{I} = \{z \in \mathbb{R}^k : z_1 \geq \cdots \geq z_k\}$ *with weights* $w_1(t), \ldots, w_k(t)$.

REMARK 1. The limiting distributions involve a similar weighted sum-of-squares-between-blocks (SSB) structure as appears in the uncensored case [El

Barmi and McKeague (2013)]. However, the weights $w_1, \ldots, w_k$ are much more complex, with time dependence and adjustment for censoring as well as the noninferiority margins. Further, our limiting distributions cannot be put into a distribution-free form because of this extra complexity (for further explanations, see the Supplementary Material Section S.3 [Chang and McKeague (2019)]).

REMARK 2. Asymptotically the local NPLR statistic $-2\log\mathcal{R}(t)$ behaves like a weighted SSB respecting the simple linear ordering constraint. This is in close analogy with standard ANOVA-type testing.

REMARK 3. In practice, t_1 and t_2 will often be specified in a study protocol. They could also be chosen in a data-driven fashion, as related to the condition $0 < S_0(t_2) < S_0(t_1) < 1$. For example, we could take $t_1 = \inf\{t : \hat{S}_j(t) < 1 \text{ for all } j\}$ and $t_2 = \sup\{t : \hat{S}_j(t) > 0 \text{ for all } j\}$ [cf. Chang, El Barmi and McKeague (2016)]; this is what we use in our simulation studies and data analysis.

REMARK 4. For $k = 2$, the limiting distribution of K_n reduces to $\sup_{t\in[t_1,t_2]}\{\Upsilon_+^2(t)\}$, where Υ is a mean-zero Gaussian process with covariance function $(\theta_1+\theta_2)(\min(s,t))/\sqrt{(\theta_1+\theta_2)(s)(\theta_1+\theta_2)(t)}$. When $M_1 = M_2 = 1$, this limiting distribution becomes the one in Theorem 1 of Chang and McKeague (2016), after a transformation in the time scale.

As mentioned in Remark 1, the asymptotic null distributions are not distribution-free, so we devise a bootstrap method to calibrate the tests. To obtain critical values for K_n and I_n, we utilize a multiplier bootstrap approach commonly used in survival analysis [see, e.g., Parzen, Wei and Ying (1997)]. Specifically, the limiting distribution of K_n and I_n can be bootstrapped by replacing each $U_j(t)$ by its multiplier bootstrap

$$U_j^*(t) = \sqrt{n_j}\sum_{i=1}^{n_j}\left[\left\{\sum_{l=1}^{n_j} I(X_{lj} \geq X_{ij})\right\}^{-1} \delta_{ij} I(X_{ij} \leq t)\xi_{ij}\right] \Big/ \hat{\sigma}_j(t) \tag{3.2}$$

and replacing $w_j(t)$ by its estimate $\hat{w}_j(t)$, where δ_{ij} is the censoring indicator and X_{ij} the minimum of the survival time and censoring time for the ith subject in the jth sample, and the ξ_{ij} are i.i.d. $N(0,1)$-multipliers independent of the data. This way, to calibrate the test we simulate K_n^* and I_n^* by repeatedly generating samples of Gaussian random multipliers $\{\xi_{ij}\}$. We then compare the empirical quantiles of these bootstrapped values K_n^* and I_n^* with our test statistic K_n and I_n, respectively.

3.2. *Testing H_{01}^c versus H_{01}.* As mentioned in the Introduction, our main procedure of testing H_0 versus H_1 needs to be preceded by a test of H_{01}^c versus $H_{01} = H_0 \cup H_1$. This initial test is used to exclude the possibility of crossings or alternative orderings that constitute H_{01}^c. Such a test can be based on

a confidence tube for $\psi_j(t) \equiv M_j \log S_j(t) - M_{j+1} \log S_{j+1}(t)$ simultaneously over $t \in [t_1, t_2]$ and $j = 1, \ldots, k-1$. We adapt the two-sample random multiplier bootstrap band of Parzen, Wei and Ying (1997) for this purpose. Specifically, using $\sqrt{n_j}(\log \hat{S}_j(t) - \log S_j(t))/\hat{\sigma}_j(t) \xrightarrow{d} U_j(t)$ jointly in $\ell^\infty([t_1, t_2])$ for $j = 1, \ldots, k$, along with the definition of $\theta_j(t)$ and the continuous mapping theorem, we have

$$\begin{aligned} &\sup_{j, t \in [t_1, t_2]} \sqrt{n} \big| \hat{\psi}_j(t) - \psi_j(t) \big| \\ &\quad \xrightarrow{d} \sup_{j, t \in [t_1, t_2]} \big| U_j(t) \sqrt{\theta_j(t)} - U_{j+1}(t) \sqrt{\theta_{j+1}(t)} \big|, \end{aligned} \tag{3.3}$$

where $\hat{\psi}_j(t) = M_j \log \hat{S}_j(t) - M_{j+1} \log \hat{S}_{j+1}(t)$ and j varies from 1 to $k-1$. This limiting distribution can be bootstrapped by replacing $U_j(t)$ by the $U_j^*(t)$ in (3.2). Then an asymptotic $100(1-\alpha)\%$ simultaneous confidence tube for $\psi_j(t)$ $(j = 1, \ldots, k-1)$ can be constructed as follows: simulate

$$\sup_{j=1,\ldots,k-1} \sup_{t \in [t_1, t_2]} \big| U_j^*(t) \sqrt{\hat{\theta}_j(t)} - U_{j+1}^*(t) \sqrt{\hat{\theta}_{j+1}(t)} \big|$$

by repeatedly generating samples of Gaussian random multipliers $\{\xi_{ij}\}$. Denote the upper α-quantile of these bootstrapped values by c_α^*. The proposed tube is $\mathcal{U}(t) = \mathcal{U}_1(t) \times \mathcal{U}_2(t) \times \cdots \times \mathcal{U}_{k-1}(t)$ for $t \in [t_1, t_2]$, where

$$\mathcal{U}_j(t) = \hat{\psi}_j(t) \pm n^{-1/2} c_\alpha^* = \big[\ell_{j-}(t), \ell_{j+}(t) \big]$$

for $j = 1, \ldots, k-1$.

We construct a rule to reject the null H_{01}^c as follows. Note that evidence for alternative H_{01} is provided by evidence for either H_0 or H_1. Support for H_1 arises on the intersection of the events $\mathcal{E}_{11} = \{\text{for all } j,\ \ell_{j-}(s_j) > 0 \text{ for some } s_j\}$ and $\mathcal{E}_{12} = \{\ell_{j+}(\cdot) \geq 0 \text{ for all } j\}$, and support for H_0 arises on the intersection of $\mathcal{E}_{01} = \{\ell_{j-}(\cdot) \leq 0 \text{ for all } j\}$ and $\mathcal{E}_{12}$; note that here j varies from 1 to $k-1$. Thus, we take the rejection region to be the union of these events, which can be expressed as $\mathcal{E} = \mathcal{E}_{12} \cap (\mathcal{E}_{01} \cup \mathcal{E}_{11})$. Note that in the two-sample case, $\mathcal{E}_{01} \cup \mathcal{E}_{11}$ is the whole parameter space [since we just need to consider $\mathcal{U}_1(t)$], so $\mathcal{E} = \mathcal{E}_{12}$; the test then coincides with the one-sided version of the test for no crossings given in Chang and McKeague (2016).

Now we show that this test has asymptotic level α. Note that (3.3) implies that $\mathcal{U}$ has coverage $P(\boldsymbol{\psi} \in \mathcal{U}) \to 1-\alpha$, with maximal width $W = O_p(1/\sqrt{n})$ over all $\mathcal{U}_j$, where $\boldsymbol{\psi} = [\psi_1, \ldots, \psi_{k-1}]^T$. It can be shown that under H_{01}^c at least one of the following occurs:

Case 1: $\inf_t \psi_\iota(t) < 0$ for some ι;
Case 2: $\sup_t \psi_\iota(t) > 0$ for some ι, and $\psi_l(\cdot) \leq 0$ for some l.

First, consider case 1 and let $\epsilon_t = -\inf_t \psi_t(t) > 0$. Since the acceptance region is $\mathcal{E}^c = \mathcal{E}_{12}^c \cup (\mathcal{E}_{01}^c \cap \mathcal{E}_{11}^c)$, and $\{\boldsymbol{\psi} \in \mathcal{U}\} \cap \{W < \varepsilon_t\} \subset \mathcal{E}_{12}^c$,

$$P\big(H_{01}^c \text{ rejected}\big) \le P(\mathcal{E}_{12}) \le P(\boldsymbol{\psi} \notin \mathcal{U}) + P(W \ge \varepsilon_t) \to \alpha.$$

A similar argument works for case 2, except that now the $\mathcal{E}_{01}^c \cap \mathcal{E}_{11}^c$ part of the acceptance region comes into play.

The family-wise error rate of the proposed two-step procedure: testing H_{01}^c followed by testing H_0, can be controlled at the same α-level as the individual tests. This is due to the partitioning principle of Finner and Strassburger (2002), which holds when the null hypotheses are disjoint (in our case H_{01}^c and H_0 are indeed disjoint).

3.3. *Combining pairwise tests for simple linear ordering.* As mentioned in the Introduction, a competing method for testing H_1 would be to use a combination of our pairwise NPLR tests. Consider the pairwise test (1.4) for a given j. By a slight extension of results in Chang and McKeague (2016), the local NPLR is

$$\mathcal{R}_j(t) = \frac{\sup\{L(S_j, S_{j+1}) \colon S_j^{M_j}(t) = S_{j+1}^{M_{j+1}}(t)\}}{\sup\{L(S_j, S_{j+1}) \colon S_j^{M_j}(t) \ge S_{j+1}^{M_{j+1}}(t)\}}. \tag{3.4}$$

Thus the approach developed in Sections 2.2–3.1 applies here, as we are dealing with the special case of $k = 2$. Then, for testing $\bigcup_{l=1}^k H_0^l$ versus H_1, the intersection-union principle can be used to construct an asymptotic level α test with the rejection region being the intersection of the individual rejection regions $\{I_{ln} > c_{\alpha,l}\}$ over $l = 1, \dots, k-1$, where

$$I_{ln} = \int_{t_1}^{t_2} \big[-2 \log \mathcal{R}_l(t)\big]\, d\hat{F}_0(t),$$

$c_{\alpha,l}$ is the upper α-quantile of the bootstrapped values based on replacing $U_l(t)$ in the limiting distribution of Theorem 1 (with only sample j and $j+1$ involved) by the $U_l^*(t)$ in (3.2).

3.4. *Testing the omnibus alternative to H_0.* Although our focus has been on noninferiority testing, a by-product of our methodology (when $M_j \equiv 1$) is a new NPLR approach to the classical k-sample problem of testing H_0 versus its complement, that is, the omnibus alternative that at least two of the survival functions are different [see, e.g., Andersen et al. (1993), Chapter 5]. When order restrictions are removed, the denominator of $\mathcal{R}(t)$ is just the unconstrained maximum (attained by the KM estimates), and thus unconstrained versions K_n^o and I_n^o of our earlier test statistics in (3.1) are readily constructed. Their asymptotic null distributions are given as follows.

THEOREM 2. *Under H_0 with $M_j \equiv 1$ and the assumptions of Theorem 1, we have*

$$K_n^o \xrightarrow{d} \sup_{t\in[t_1,t_2]} \mathrm{SSB}^o(t), \qquad I_n^o \xrightarrow{d} \int_{t_1}^{t_2} \mathrm{SSB}^o(t)\, dF_0(t),$$

where

$$\mathrm{SSB}^o(t) = \sum_{j=1}^{k} w_j(t)\Big[\frac{U_j(t)}{\sqrt{w_j(t)}} - \bar{U}(t)\Big]^2.$$

The limiting process $\mathrm{SSB}^o(t)$ can again be interpreted as a weighted sum of squares between blocks, but now each component of the sum only compares an individual group (rather than adjacent groups that have been pooled) with an overall weighted average, in parallel with standard ANOVA. In other words, each block only consists of an individual group because no order restrictions have been imposed. These tests can be calibrated by a similar multiplier bootstrap approach as described in Section 3.1.

Interestingly, the nonparametric-likelihood-optimized $\hat{w}_j$ that weight the information from the k groups do not seem to have been proposed previously. These weights take into account sampling variation, the sample proportions p_j, the margins and the censoring patterns, which are not fully incorporated into, for example, tests of Jonckheere–Terpstra-type [Gehan (1965), Liu and Tsai (1999)] or Bonferroni correction of pairwise comparisons.

4. Approximating the local NPLR.

4.1. *Solving the constrained optimization.* In this section, we derive the expression (2.7) for the local NPLR, which is a first step toward analyzing the asymptotic behavior of our test statistics. First, consider the denominator of the local NPLR (2.2) and apply the KKT method. After a log transformation, the optimization problem is to minimize

$$-\sum_{j=1}^{k}\sum_{i=1}^{m_j}\{d_{ij}(\log h_{ij}) + (r_{ij} - d_{ij})\log(1 - h_{ij})\} \tag{4.1}$$

over $\boldsymbol{h} \equiv (h_{11}, \ldots, h_{m_1 1}, \ldots, h_{1k}, \ldots, h_{m_k k}) \in (0,1)^m$, where $m = \sum_{j=1}^{k} m_j$, subject to the constraints

$$M_{j+1}\sum_{i\le N_{j+1}(t)} \log(1 - h_{i,j+1}) - M_j \sum_{i\le N_j(t)} \log(1 - h_{ij}) \le 0 \tag{4.2}$$

for $j = 1, \ldots, k-1$. 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. Define the Lagrangian $\mathcal{L}: (0,1)^m \times \mathbb{R}^{k-1} \to \mathbb{R}$ by

$$\mathcal{L}(\boldsymbol{h}, \boldsymbol{\lambda}) \equiv -\sum_{j=1}^{k}\sum_{i=1}^{m_j}\{d_{ij}(\log h_{ij}) + (r_{ij} - d_{ij})\log(1-h_{ij})\} + \sum_{j=1}^{k-1}\lambda_j\left\{M_{j+1}\sum_{i\le N_{j+1}(t)}\log(1-h_{i,j+1}) - M_j\sum_{i\le N_j(t)}\log(1-h_{ij})\right\},$$

where $\boldsymbol{\lambda} \equiv (\lambda_1, \ldots, \lambda_{k-1})$. An optimal solution (at the end of this paragraph we show it is unique), denoted as $(\tilde{\boldsymbol{h}}, \tilde{\boldsymbol{\lambda}})$, must satisfy the KKT conditions in (2.6) and $\nabla_{\boldsymbol{h}}\mathcal{L}(\tilde{\boldsymbol{h}}, \tilde{\boldsymbol{\lambda}}) = 0$ (stationarity). The stationarity condition yields $\tilde{h}_{ij} = d_{ij}/r_{ij}$ for $i = N_j(t)+1, \ldots, m_j$ and (2.5) for $i = 1, \ldots, N_j(t)$, $j = 1, \ldots, k$. Define $\mu_j = \max_{i=1,\ldots,N_j(t)}(d_{ij} - r_{ij})$. Since $\tilde{\boldsymbol{h}}$ is in the domain $(0,1)^m$, we have that $\mu_j < M_j(\tilde{\lambda}_j - \tilde{\lambda}_{j-1})$, where $\mu_j \le 0$ for $j = 1, \ldots, k$. This implies $\mathcal{L}(\boldsymbol{h}, \tilde{\boldsymbol{\lambda}})$ is a strictly convex function of $\boldsymbol{h}$, and thus the optimal solution is unique [see, e.g., Boyd and Vandenberghe (2004), page 248].

The numerator of $\mathcal{R}(t)$ can be handled in a similar fashion. Denoting the optimal solution in this case by $(\bar{\boldsymbol{h}}, \bar{\boldsymbol{\lambda}})$, it turns out $\bar{h}_{ij}$ has the same form as $\tilde{h}_{i,j}$ but with $\tilde{\lambda}_j$ replaced by $\bar{\lambda}_j$, and $\bar{\lambda}_j$ only needs to satisfy $\mu_j < M_j(\bar{\lambda}_j - \bar{\lambda}_{j-1})$ for $j = 1, \ldots, k$ (recall from Section 2.2 that $\bar{\lambda}_k = 0$) and (2.4). By a similar argument as used before, the optimal solution $(\bar{\boldsymbol{h}}, \bar{\boldsymbol{\lambda}})$ is unique. Note that the estimated hazards after time t under no constraints, namely $\tilde{h}_{ij}$ and $\bar{h}_{ij}$ for $i = N_j(t)+1, \ldots, m_j$, are the same in the numerator and denominator, and so these terms cancel out. This leads to (2.7).

4.2. *Limiting behavior of the local NPLR statistic.* To prove Theorem 1, we first approximate the local NPLR statistic in terms of the margins M_j and the Lagrange multipliers in the following lemma; see Supplementary Material, Section S.1.1 for the proof [Chang and McKeague (2019)].

LEMMA 3. *Under the conditions of Theorem 1,*

$$-2\log\mathcal{R}(t) = \sum_{j=1}^{k}\big(\bar{\Delta}_j^2(t) - \tilde{\Delta}_j^2(t)\big)\frac{\hat{\sigma}_j^2(t)}{n_j} + O_p\big(n^{-1/2}\big)$$

uniformly over $t \in [t_1, t_2]$, *where* $\tilde{\Delta}_j(t) = M_j(\tilde{\lambda}_j(t) - \tilde{\lambda}_{j-1}(t))$, $\bar{\Delta}_j(t) = M_j(\bar{\lambda}_j(t) - \bar{\lambda}_{j-1}(t))$ *for* $j = 1, 2, \ldots, k$.

Similar decompositions just involving terms of the form $\bar{\Delta}_j(t)$ have previously been found for $M_j \equiv 1$ and $k \le 2$ [see, e.g., Li (1995), McKeague and Zhao (2002)], but the presence of the $\tilde{\Delta}_j(t)$ term is new due to the order-restricted aspect of our problem. The special case $k = 2$ is studied in Chang and McKeague

(2016), but the general decomposition involving $\tilde{\Delta}_j(t)$ cannot be derived from that case. This reduces the problem to studying the large sample properties of $\bar{\Delta}_j(t)$ and $\tilde{\Delta}_j(t)$, since the limiting behavior of $\hat{\sigma}_j^2(t)$ is already known.

The processes $\bar{\Delta}_j(t)$ can be jointly approximated by a time-dependent random linear transformation of the processes $\hat{U}_j(t)$, $j = 1, \ldots, k$, where $\hat{U}_j(t)$ is defined in Section 2.1. We then make use of standard asymptotic properties of $\hat{S}_j(t)$ and the functional delta method to study the limiting behavior of $\bar{\Delta}_j(t)$. A similar approach will be used for $\tilde{\Delta}_j(t)$, but the problem is more challenging because each PAVA block has to be treated separately. These considerations and the fact that $\hat{\sigma}_j^2(t)$ is a uniformly consistent estimate of $\sigma_j^2(t)$ lead to the following lemma; see Appendices B.1 and B.2, and the Supplementary Material, Sections S.1.2–S.1.3 [Chang and McKeague (2019)] for the proof.

LEMMA 4. *Under the conditions of Theorem 1, the leading term of the decomposition in Lemma 3 is asymptotically equivalent to*

$$\sum_{B\in\mathcal{B}(t)}\sum_{l\in B} w_l(t)\{\hat{U}_B(t) - \check{U}(t)\}^2 \xrightarrow{d} \mathrm{SSB}(t)$$

as processes on $[t_1, t_2]$, *where* $\mathcal{B}(t)$ *is the set of blocks resulting from PAVA*,

$$\hat{U}_B(t) = \sum_{j\in B}\sqrt{w_j(t)}\hat{U}_j(t)\Big/\sum_{j\in B} w_j(t), \qquad \check{U}(t) = \sum_{j=1}^{k}\sqrt{w_j(t)}\hat{U}_j(t).$$

Lemmas 3 and 4 together connect the asymptotic behavior of $-2\log\mathcal{R}(t)$ with a weighted variation among $\hat{U}_B(t)$ over the PAVA blocks $B \in \mathcal{B}(t)$, and show that it converges in distribution to $\mathrm{SSB}(t)$. It is obvious that $\check{U}(t) \xrightarrow{d} \bar{U}(t)$ in $\ell^\infty[t_1, t_2]$, but it is more challenging to obtain the limiting distribution of $\hat{U}_B(t)$ for $B \in \mathcal{B}(t)$ because $\mathcal{B}(t)$ changes with n. The proof utilizes the sample version $E_{\boldsymbol{w}}(\hat{\boldsymbol{U}}_{\boldsymbol{w}}(t)|\mathcal{I})$ of $E_{\boldsymbol{w}}(\boldsymbol{U}_{\boldsymbol{w}}(t)|\mathcal{I})$, where $\hat{\boldsymbol{U}}_{\boldsymbol{w}}(t)$ is defined by replacing $U_j(t)$ by $\hat{U}_j(t)$ in $\boldsymbol{U}_{\boldsymbol{w}}(t)$. Specifically, by a similar PAVA argument as in Section 2.3 and Appendix A, $E_{\boldsymbol{w}}(\hat{\boldsymbol{U}}_{\boldsymbol{w}}(t)|\mathcal{I})$ determines a (possibly different) set of blocks $\mathcal{B}'(t)$ such that $\hat{U}_{B_1}(t) \geq \hat{U}_{B_2}(t)$ for $B_1 \in \mathcal{B}'(t)$ preceding $B_2 \in \mathcal{B}'(t)$; for the details, see the proof of Lemma S.2 in the Supplementary Material, Section S.1.4 [Chang and McKeague (2019)]. The key part of the proof is to show that

$$\begin{aligned}&\sum_{B\in\mathcal{B}(t)}\sum_{j\in B} w_j(t)\{\hat{U}_B(t) - \check{U}(t)\}^2\\ &\quad= \sum_{j=1}^{k} w_j(t)\{E_{\boldsymbol{w}}(\hat{\boldsymbol{U}}_{\boldsymbol{w}}(t)|\mathcal{I})_j - \check{U}(t)\}^2 + o_p(1)\end{aligned} \tag{4.3}$$

uniformly over $t \in [t_1, t_2]$. Then, since $E_{\boldsymbol{w}}(\cdot|\mathcal{I})$ is a continuous function on $\mathbb{R}^k$, by the continuous mapping theorem we obtain that the first term on the RHS of (4.3) converges weakly to SSB(t), which establishes Lemma 4. To show (4.3), let $B(j) \in \mathcal{B}(t)$, and $B'(j) \in \mathcal{B}'(t)$ denote the blocks containing j. Then (4.3) is equivalent to

$$\begin{aligned}(4.4)\qquad &\sum_{j=1}^{k} w_j(t)\{\hat{U}_{B(j)}(t) - \check{U}(t)\}^2 - \sum_{j=1}^{k} w_j(t)\{\hat{U}_{B'(j)}(t) - \check{U}(t)\}^2 \\ &\quad = \sum_{j=1}^{k} w_j(t)\{\hat{U}_{B(j)}(t) - \hat{U}_{B'(j)}(t)\}\{\hat{U}_{B(j)}(t) + \hat{U}_{B'(j)}(t)\} = o_p(1),\end{aligned}$$

since $\sum_{j=1}^{k} w_j(t)\check{U}(t)\{\hat{U}_{B(j)}(t) - \hat{U}_{B'(j)}(t)\} = 0$. Further, $\sum_{j=1}^{k} w_j(t)\{\hat{U}_{B(j)}(t) + \hat{U}_{B'(j)}(t)\}$ is uniformly bounded in probability, so (4.4) is a consequence of the following lemma.

LEMMA 5. *Under the conditions of Theorem 1, for each $j = 1, \ldots, k$, the processes $\hat{U}_{B(j)}(t)$ and $\hat{U}_{B'(j)}(t)$ are asymptotically equivalent uniformly over $t \in [t_1, t_2]$.*

The proof of Lemma 5 is based on understanding how the PAVA that solves the NPLR (as described in Section 2.3) relates to the PAVA from $E_{\boldsymbol{w}}(\hat{\boldsymbol{U}}_{\boldsymbol{w}}(t)|\mathcal{I})$; see the Supplementary Material, Sections S.1.3 and S.1.4 [Chang and McKeague (2019)].

5. Extensions.

5.1. *Retention-of-effect.* As mentioned in the Introduction, an alternative way of assessing noninferiority (of experimental to reference treatments) is in terms of log-relative-survival versus placebo. Here, we briefly present an extension of our method to deal with such retention-of-effect hypotheses in the three-arm setting described in the Introduction. That is, we are interested in testing

$$(5.1)\qquad H_0^r\colon S_2/S_1 \preceq (S_3/S_1)^M \quad \text{versus} \quad H_1^r\colon S_2/S_1 \succ (S_3/S_1)^M.$$

Here, the noninferiority margin $M > 1$ specifies how much the effect of the standard therapy relative to placebo is retained by the experimental treatment. Specifically, $S_2/S_1 \succ (S_3/S_1)^M$ means the experimental treatment achieves more than $1/M \times 100\%$ of the effect of the standard therapy (relative to placebo).

We can again devise a test statistic based on the local NPLR using the Lagrange/KKT method; see the Supplementary Material, Section S.4 [Chang and McKeague (2019)]. The $k = 3$ version of Theorem 1 extends to this setting, the only differences being that the projection is onto $\mathcal{I} = \{\boldsymbol{z} \in \mathbb{R}^3 : z_1 = z_2 \geq z_3\}$, and the weight functions w_1, w_2, w_3 are more complicated because the survival functions do not have equivalent roles (as they do in simple linear ordering).

5.2. *Stratified testing*. In multicenter clinical trials, randomization of subjects is often carried out separately by medical center or region. Our procedure involves two steps with different test statistics, so developing a unified approach to stratified testing raises some difficulties. If the interest is in testing whether H_1 holds in at least one stratum, then a straightforward approach would be to use our original procedure in each stratum, along with a Bonferroni adjustment. On the other hand, if the interest is in testing whether H_1 holds in all the strata, then by the intersection-union principle, the same alpha-level can be used for each stratum-specific procedure.

6. Simulation study. In this section, we compare the performance of the proposed procedures with the competing method based on the combination of pairwise tests (using the intersection-union principle). For the pairwise tests, we consider two possibilities: one based on our NPLR method, and the other based on a comparable Cox model absolute-margin analysis. As mentioned in the Introduction, these procedures do not need multiplicity adjustment, unless there is interest in testing the individual pairwise alternatives. For this need, we also consider multiplicity-adjustment in the pairwise tests and compare our NPLR approach with the corresponding Cox model absolute-margin analysis. We restrict to the case of noninferiority testing with simple linear ordering.

The simulation setup is a three-arm noninferiority trial ($k = 3$) in which shorter survival is desirable (as in our real data example in Section 7): S_1 represents a placebo, S_2 a standard therapy and S_3 an experimental therapy. We consider three Weibull scenarios: A (representing H_0), B and C (representing H_1), and in each define the S_j by specifying $S_j^{M_j}$ (see Figure 1) and M_j. In addition, there are three sets of margins: $(M_1, M_2, M_3) = (1, 1, 10/8)$, $(M_1, M_2, M_3) = (1, 1, 10/7)$ and $(M_1, M_2, M_3) = (1.1, 1, 10/8)$; this results in nine sets of (S_1, S_2, S_3). All of these margins represent superiority of the standard therapy over the placebo (i.e., $M_1 \geq M_2$), and a noninferiority of the experimental to the standard therapy (i.e.,

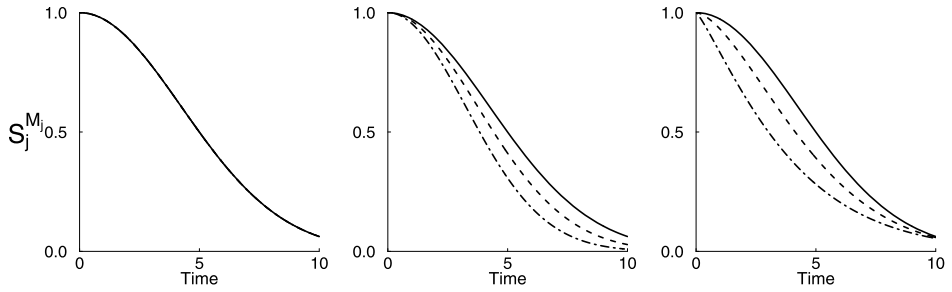

FIG. 1. *Simulation scenarios*: *A* (*left*) *representing* H_0, *B* (*middle*) *and C* (*right*) *representing* H_1. *Each* $S_j^{M_j}$ *is specified as Weibull*: *placebo* (*solid*), *standard therapy* (*dashed*), *experimental therapy* (*two-dashed*). *Note in A all the three lines overlap*.

$M_2 < M_3$) with a margin of 0.8 or 0.7. We specify the censoring distributions (the same in each arm) to be uniform with administrative censoring at $t = 10$, and a censoring rate of either 10% or 25% in the placebo group. Also, we consider a per group sample size of 130, which is close to the smallest group size in our real data example. Additional settings including different H_1^c setup, unequal group sample sizes and higher censoring rate (up to 50%) are also considered and presented in the Supplementary Material, Section S.5.1.

6.1. *Accuracy.* We generate data using the aforementioned three sets of (M_1, M_2, M_3) and Scenario A (in which H_0 is satisfied; see the left panel of Figure 1). Our two-step NPLR procedures based on K_n and I_n are compared with the combined-pairwise NPLR test in Section 3.3, as well as a version of the Wald-type Cox model approach [Kombrink, Munk and Friede (2013)]. Specifically, the latter approach involves fitting a single Cox model, with the placebo as the reference group and indicator covariates for the standard and experimental groups. Denote the corresponding regression coefficients as β_2 and β_3, respectively. The test rejects the overall alternative $H_1^{S,R,N} = H_1^{S,R} \cap H_1^N$ if both $H_0^{S,R}$: $\exp(\beta_2) \le M_1/M_2$ and H_0^N: $\exp(\beta_3) \le (M_2/M_3)\exp(\beta_2)$ are rejected (i.e., an intersection-union test). Since ordering in hazard functions implies ordering in survival functions, $H_1^{S,R}$ implies H_1^1, H_1^N implies H_1^2, and $H_1^{S,R,N}$ implies H_1. Note that their original H_0^N is formulated in terms of β_2/β_3 expressing a retention-of-effect; changing the comparison to a ratio of hazards [i.e., $\exp(\beta_2 - \beta_3)$], H_0^N corresponds to $M_3 h_3 \prec M_2 h_2$, thus providing an absolute-margin analysis. We have also modified their original $H_0^{S,R}$ to allow for incorporation of the margins M_1 and M_2; when $M_1 = M_2 = 1$ this coincides with their original $H_0^{S,R}$.

The empirical levels of these tests are given in Table 1. Our proposed two-step NPLR procedures based on K_n and I_n are both close to the nominal level, whereas the competing methods based on combining pairwise tests are extremely

TABLE 1

Empirical significance levels at $\alpha = 0.05$ *based on* 1000 *replications, for* $n_j = 130$ *and* $k = 3$. *Calibrations for the NPLR tests are based on* 1000 *bootstrap samples. Cox denotes the combined Cox model test described in Section* 6.1, *and PW denotes the combined-pairwise NPLR test*

		10% censoring				**25% censoring**			
Scenario	(M_1, M_2, M_3)	K_n	I_n	**Cox**	**PW**	K_n	I_n	**Cox**	**PW**
A	(1, 1, 10/8)	0.052	0.052	0	0	0.041	0.049	0	0
	(1, 1, 10/7)	0.048	0.051	0	0	0.049	0.047	0	0
	(1.1, 1, 10/8)	0.055	0.051	0	0	0.044	0.044	0	0

TABLE 2

Empirical significance levels at $\alpha = 0.025$ *based on* 1000 *replications, for* $n_j = 130$ *and* $k = 3$. *Calibrations for the NPLR tests are based on* 1000 *bootstrap samples. Cox denotes a pairwise Cox model test (Section 6.1), and PW denotes a pairwise NPLR test. The alternative hypotheses are indicated on the top row*

		$H_1^{S,R}$		H_1^1		H_1^N		H_1^2		$H_1^{S,R,N}$		H_1	
		Cox		**PW**		**Cox**		**PW**		**Cox**		**PW**	
Scenario	(M_1, M_2, M_3)	**10%**	**25%**	**10%**	**25%**	**10%**	**25%**	**10%**	**25%**	**10%**	**25%**	**10%**	**25%**
A	(1, 1, 10/8)	0.022	0.023	0.023	0.021	0.031	0.026	0.026	0.027	0	0	0	0
	(1, 1, 10/7)	0.022	0.022	0.024	0.022	0.028	0.025	0.028	0.025	0	0	0	0
	(1.1, 1, 10/8)	0.024	0.025	0.024	0.026	0.023	0.023	0.026	0.026	0	0	0	0

conservative. We have found that the individual pairwise tests have accurate empirical levels (results not shown), so the use of the intersection-union principle in the construction of the combined-pairwise tests appears to be the source of their conservativeness. We conclude that the proposed two-step NPLR tests are more accurate than the combined-pairwise NPLR test and Cox model test in moderate samples.

Next, we consider testing the individual pairwise alternatives using a Bonferroni adjustment. We compare our pairwise NPLR tests with the pairwise Cox model tests described above. The empirical levels of these tests are given in Table 2. Both the NPLR and Cox model tests for the individual alternatives have accurate levels, but those for the overall alternatives are extremely conservative. Kombrink, Munk and Friede [(2013), Section 5.1], noted that such overall levels can be controlled in their retention-of-effect setting. However, the drawback of using pairwise tests is that they are too conservative overall, even though they have the benefit of addressing the individual alternatives.

6.2. *Power comparisons*. In this section, we compare the empirical powers of the proposed NPLR tests with the combined-pairwise NPLR test and Cox model test described in Section 6.1. We generate data using the aforementioned three sets of (M_1, M_2, M_3) and Scenarios B and C in which H_1 is satisfied; see the middle and right panels of Figure 1, respectively. The $S_j^{M_j}$ have proportional hazards in Scenario B, but crossing hazards in Scenario C.

The empirical powers are reported in Table 3. Our proposed two-step NPLR procedures based on K_n and I_n outperform the other two tests in all the cases considered, with an increase in power ranging from 10–80%, even under Scenario B. The combined Cox model test performs better than the combined NPLR test under proportional hazards in Scenario B, but much worse under crossing hazards in Scenario C. We conclude that the proposed two-step NPLR tests are much more powerful than any of the competing methods.

TABLE 3
Empirical powers in the setting of Table 1 except under the alternatives Scenario B (proportional hazards) and C (crossing hazards)

		10% censoring				25% censoring			
Scenario	(M_1, M_2, M_3)	K_n	I_n	**Cox**	**PW**	K_n	I_n	**Cox**	**PW**
B	(1, 1, 10/8)	0.713	0.723	0.374	0.276	0.778	0.806	0.291	0.236
	(1, 1, 10/7)	0.648	0.659	0.373	0.263	0.710	0.739	0.280	0.215
	(1.1, 1, 10/8)	0.703	0.709	0.391	0.273	0.765	0.796	0.312	0.237
C	(1, 1, 10/8)	0.933	0.933	0.135	0.665	0.938	0.938	0.154	0.608
	(1, 1, 10/7)	0.936	0.936	0.099	0.636	0.948	0.949	0.138	0.588
	(1.1, 1, 10/8)	0.926	0.926	0.113	0.652	0.940	0.940	0.134	0.592

The empirical powers of the multiplicity-adjusted pairwise NPLR tests and pairwise Cox model tests (see Section 6.1) are given in Table 4. Again the Cox approach performs better than the pairwise NPLR tests under proportional hazards in Scenario B, but much worse under crossing hazards in Scenario C (comparing the Cox and PW columns).

7. Application to a noninferiority trial. We analyze data from a three-arm noninferiority clinical trial involving treatments for major depression [Mielke, Munk and Schacht (2008)]. The data are obtained by digitizing the published KM curves and reconstructing survival and censoring information using the algorithm developed by Guyot et al. (2012). The endpoint is time (in days) to first remission, where remission is defined as the attainment of a low score (of ≤ 7) on the 17-item Hamilton Depression Rating Scale. A shorter time to first remission is desirable.

TABLE 4
Empirical powers in the setting of Table 2 except under the alternatives Scenario B (proportional hazards) and C (crossing hazards)

		$H_1^{S,R}$		H_1^1		H_1^N		H_1^2		$H_1^{S,R,N}$		H_1	
		Cox		**PW**		**Cox**		**PW**		**Cox**		**PW**	
	(M_1, M_2, M_3)	**10%**	**25%**	**10%**	**25%**	**10%**	**25%**	**10%**	**25%**	**10%**	**25%**	**10%**	**25%**
B	(1, 1, 10/8)	0.482	0.417	0.396	0.361	0.625	0.563	0.548	0.492	0.227	0.156	0.147	0.107
	(1, 1, 10/7)	0.479	0.412	0.388	0.364	0.612	0.526	0.539	0.455	0.211	0.139	0.137	0.092
	(1.1, 1, 10/8)	0.482	0.409	0.392	0.346	0.627	0.581	0.540	0.514	0.222	0.164	0.139	0.108
C	(1, 1, 10/8)	0.264	0.286	0.661	0.642	0.399	0.434	0.787	0.785	0.036	0.060	0.475	0.457
	(1, 1, 10/7)	0.259	0.278	0.657	0.635	0.340	0.376	0.756	0.746	0.025	0.047	0.445	0.421
	(1.1, 1, 10/8)	0.222	0.250	0.643	0.625	0.392	0.431	0.786	0.781	0.031	0.044	0.463	0.443

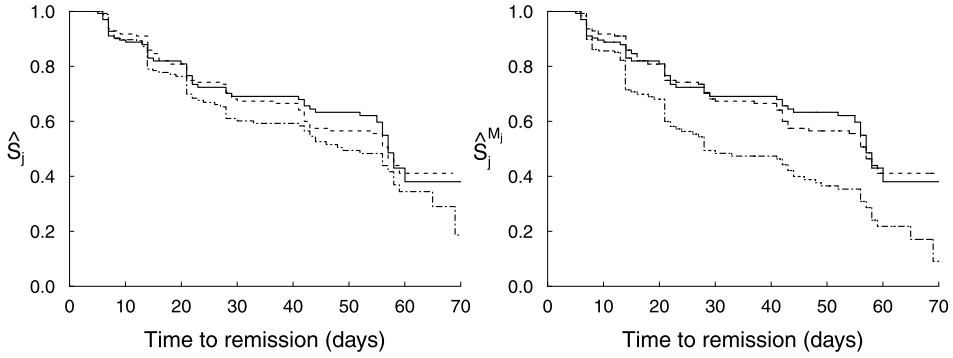

FIG. 2. *KM curves* $\hat{S}_j$ *(left) and* $\hat{S}_j^{M_j}$ *(right) from a noninferiority trial comparing treatments of major depression*: *placebo* (*solid*), *standard treatment* (*dashed*) *and experimental treatment* (*two–dashed*).

A total of 664 patients were randomized to placebo ($n_1 = 135$), standard treatment ($n_2 = 267$) or experimental treatment ($n_3 = 262$); see Figure 2 for the KM and $\hat{S}_j^{M_j}$ curves.

Using $(M_1, M_2, M_3) = (1, 1, 10/7)$, our two-step NPLR test based on K_n ($p < 0.01$) indicates noninferiority of the experimental treatment over the standard treatment and assay sensitivity, whereas I_n fails to provide such evidence ($p = 0.09$). Judging from the $\hat{S}_j^{M_j}$ curves (Figure 2, right), the reason appears to be that the maximal difference (measured by K_n and appearing at around 50 days) in the survival curves between placebo and standard treatment is large enough to establish assay sensitivity, whereas the cumulative difference (measured by I_n) is not large enough. The conclusion from I_n is supported by both the combined-pairwise Cox (see Section 6.1) and NPLR procedures. For the individual pairwise alternatives at $\alpha = 0.025$, the pairwise Cox model tests show noninferiority ($p < 0.01$) but cannot establish assay sensitivity ($p = 0.37$). The pairwise NPLR tests, on the other hand, cannot conclude assay sensitivity ($p = 0.66$) or noninferiority ($p = 0.037$ for absolute-margin and $p = 0.038$ for retention-of-effect).

It is often of interest to carry out power calculations, determine sample sizes and find optimal allocation ratios prior to a clinical trial. To do this, we suggest a Monte Carlo simulation-based approach; we have provided an example of such an analysis in the Supplementary Material, Section S.5.2 [Chang and McKeague (2019)], as well as included the corresponding R code (see Supplementary Material, Section S.6) for the power calculations.

8. Discussion. In this paper, we have developed nonparametric likelihood ratio tests for a novel class of orderings of $k \geq 2$ survival functions based on right-censored data. The proposed test statistics are constructed as maximally-selected and integrated local NPLR statistics. We show that the limiting distributions of

these test statistics can be expressed in terms of weighted sums of squares involving projections of certain Gaussian processes onto the given ordered alternative. The limiting distributions are complex, so we propose calibrating the tests via the multiplier bootstrap approach. We carried out a simulation study for the case of simple linear ordering, and found that the new tests have more accurate type I error and are more powerful than a competing combined pairwise approach based on either NPLR or a comparable Cox model test for an absolute-margin analysis. In addition, a multiplicity-adjusted version of our pairwise NPLR tests is shown to be more powerful than the comparable Cox model approach under crossing hazards. We applied the proposed tests to data from a three-arm noninferiority trial and obtained a more significant result than the alternative Cox model procedure. While our focus is on ordered alternative hypotheses, a by-product of our methodology (when $M_j \equiv 1$) is a new nonparametric likelihood ratio approach to the classical k-sample problem of testing equality of multiple survival functions.

All k-sample comparison problems naturally involve weights corresponding to each group, often selected in an ad hoc fashion, but in our case we use a principled (likelihood ratio) approach that implicitly provides optimal weighting. These weights emerge explicitly in the asymptotic distributions of the nonparametric likelihood ratio statistics. In finite samples, the weighting adjusts for sampling variation, the sample proportions p_j, the margins and the censoring patterns, which are not fully incorporated into, for example, previously proposed tests of Jonckheere–Terpstra-type [Gehan (1965), Liu and Tsai (1999)] or Bonferroni correction of pairwise comparisons.

We have provided a complete solution to the problem of testing for order-restrictions among survival functions under right censorship via nonparametric likelihood ratio statistics. Our solution hinges on introducing a novel type of PAVA and establishing its asymptotic equivalence to the projection of Gaussian processes mentioned in the previous paragraph. The proposed procedure is formulated for simple linear orderings that are relevant to noninferiority testing, and naturally extends to more general tests for partial orderings. The only change in the limiting distribution in Theorem 1 (and the associated bootstrap calibration) is to use the projection corresponding to the particular partial order. Possible applications of this generalization include comparing multiple experimental treatments with various standard therapies. Another extension is to provide equivalence testing, which involves two-sided ordered alternatives for the survival functions [see, e.g., Romano (2005), for parametric settings]. This could be done within our framework by testing H_1, along with a reversed version of H_1 with different margins, based on the intersection-union principle.

APPENDIX A: CHARACTERIZATION OF NPLR VIA PAVA

Since the inequality constrained maximizer of the denominator of the local NPLR characterized by (2.5) and (2.6) is unique (see Section 4.1), we only need to

find one solution (e.g., the one obtained by PAVA). Let q index the sweeps in our version of PAVA (as discussed in Section 2.3): $q = 1, \ldots, Q$, where the algorithm terminates after sweep Q. Denote the set of blocks resulting from sweep q by $\mathcal{B}^{(q)}$ and recall that the algorithm is initialized by $\mathcal{B}^{(0)} = \{\{1\}, \ldots, \{k\}\}$. Define, for a given block B, $\bar{S}_B(t) = \prod_{i \le N_j(t)} (1 - \check{h}_{ij})^{M_j}$, where $\check{h}_{ij}$ is given in (2.8), and j can be any given element of B because of the equality constraint for all $j \in B$; note that $\bar{S}_B(t)$ does not depend on q once we fix a $B \in \mathcal{B}^{(q)}$. Since $\check{h}_{ij}$ takes the same form as (2.3), (2.5) is satisfied if we can show that the Lagrange multipliers under the equality constraints (2.4) for $j \in B$ and any $B \in \mathcal{B}^{(Q)}$ satisfy (2.6a)–(2.6c). Indeed, (2.6a) is satisfied since our PAVA terminates with $\bar{S}_B(t) \ge \bar{S}_{B'}(t)$ for $B, B' \in \mathcal{B}^{(Q)}$ if B precedes B'. As for (2.6b) and (2.6c), suppose the final sweep results in more than one block and let $B \in \mathcal{B}^{(Q)}$ (the proof is similar when it results in only one block). Denote the maximal element of B as l. Then $\tilde{\lambda}_l = 0$ (by definition when B is the last block), because there is no pooling between B and the next block when B is not the last block. This means (2.6b) and (2.6c) are satisfied when $j = l$. On the other hand, for $j' \in B \setminus \{l\}$, because we impose equality constraints when pooling within a block, (2.6c) holds when $j = j'$.

The challenging part of the argument is to show that (2.6b) holds when $j = j'$ for $j' \in B \setminus \{l\}$. To this end, we use induction on the sweep index q. In contrast to the previous argument that focuses on $\mathcal{B}^{(Q)}$, we instead take $B \in \mathcal{B}^{(q)}$ and show

$$\lambda_j^{(q)} > 0 \qquad \text{for all } j \in B \setminus \{l\}, \tag{A.1}$$

where $\lambda_j^{(q)}$ is jth Lagrange multiplier after sweep q and l again denotes the maximal element of B. The induction proof assumes the result for sweep $q-1$ and then proves (A.1). This is done by comparing the Lagrange multipliers before and after pooling at sweep q in which a new block $B = B_1 \cup B_2$ is produced from adjacent violators $B_1, B_2 \in \mathcal{B}^{(q-1)}$.

For the initial induction step $q = 1$, let the first two adjacent violators be $\{\iota\}, \{\iota + 1\} \in \mathcal{B}^{(0)}$ [i.e., $\bar{S}_{\{\iota\}}(t) < \bar{S}_{\{\iota+1\}}(t)$], which are being pooled as $B = \{\iota\} \cup \{\iota + 1\} \in \mathcal{B}^{(1)}$. In connection with (2.3) and (2.4) in the characterization of the local NPLR, define the function $a_j(x) = [\prod_{i \le N_j(t)} \{1 - d_{ij}/(r_{ij} + x)\}]^{M_j}$ for $x \in (\mu_j, \infty)$ and $j = 1, \ldots, k$, where we suppress the dependence of a_j on t. Since $a_\iota(M_\iota \lambda)/a_{\iota+1}(-M_{\iota+1}\lambda)$ is strictly increasing and continuous in λ, tends to ∞, and is less than 1 when $\lambda = 0$ [by $\bar{S}_{\{\iota\}}(t) < \bar{S}_{\{\iota+1\}}(t)$], it equals 1 at a unique $\lambda_\iota^{(1)} > 0$. So we have shown that (A.1) holds for $q = 1$ and $B \in \mathcal{B}^{(1)}$ ($l = \iota + 1$ here). For the main inductive step, now suppose (A.1) holds at sweep $q - 1$ ($q \ge 2$). Then for sweep q, let the first two adjacent violators be $B_1, B_2 \in \mathcal{B}^{(q-1)}$, which are pooled as $B = B_1 \cup B_2 \in \mathcal{B}^{(q)}$. Suppose B_1 precedes B_2, so $\bar{S}_{B_1}(t) < \bar{S}_{B_2}(t)$. The situation is more complex than $q = 1$, because instead of dealing with one function $a_j(\cdot)$ in each block, we have to deal with possibly many $a_j(\cdot)$ satisfying a string of

equalities coming from the earlier pooling steps that resulted in B_1 and B_2:

$$\begin{aligned}\bar{S}_{B_1}(t) &= a_\iota\bigl(M_\iota\lambda_\iota^{(q-1)}\bigr) = a_{\iota+1}\bigl(M_{\iota+1}\bigl(\lambda_{\iota+1}^{(q-1)} - \lambda_\iota^{(q-1)}\bigr)\bigr) = \cdots \\ &= a_\jmath\bigl(-M_\jmath\lambda_{\jmath-1}^{(q-1)}\bigr)\end{aligned} \tag{A.2}$$

and

$$\bar{S}_{B_2}(t) = a_{\jmath+1}\bigl(M_{\jmath+1}\lambda_{\jmath+1}^{(q-1)}\bigr) = \cdots = a_l\bigl(-M_l\lambda_{l-1}^{(q-1)}\bigr), \tag{A.3}$$

where ι and $\jmath$ denote the minimal element and maximal element of B_1, respectively. Using (A.2), (A.3) and the fact that $\bar{S}_{B_1}(t) < \bar{S}_{B_2}(t)$, we next show that $\bar{S}_B(t)$ is strictly sandwiched between $\bar{S}_{B_1}(t)$ and $\bar{S}_{B_2}(t)$. The idea is to compare (A.2) and (A.3) with $\bar{S}_B(t)$. After pooling at sweep q, $\bar{S}_B(t)$ satisfies a similar string of equalities to (A.2) except with $q-1$ replaced by q and the last term being $a_\jmath(M_\jmath(\lambda_\jmath^{(q)} - \lambda_{\jmath-1}^{(q)}))$. Also, $\bar{S}_B(t)$ satisfies a similar string of equalities to (A.3) except with $q-1$ replaced by q and the first term being $a_{\jmath+1}(M_{\jmath+1}(\lambda_{\jmath+1}^{(q)} - \lambda_\jmath^{(q)}))$. First, we show that $\bar{S}_{B_1}(t) < \bar{S}_B(t)$, arguing by contradiction. Suppose $\bar{S}_{B_1}(t) \geq \bar{S}_B(t)$. Then because $a_j(\cdot)$ is strictly increasing, a comparison of (A.2) and the string of equalities satisfied by $\bar{S}_B(t)$ leads to

$$\lambda_\iota^{(q)} \leq \lambda_\iota^{(q-1)}, \tag{A.4}$$

$$\lambda_{j+1}^{(q)} - \lambda_j^{(q)} \leq \lambda_{j+1}^{(q-1)} - \lambda_j^{(q-1)}, \qquad j = \iota, \ldots, \jmath-2, \tag{A.5}$$

$$\lambda_\jmath^{(q)} - \lambda_{\jmath-1}^{(q)} \leq -\lambda_{\jmath-1}^{(q-1)}. \tag{A.6}$$

By (A.4) and (A.5), $\lambda_j^{(q)} \leq \lambda_j^{(q-1)}$ for $j = \iota+1, \ldots, \jmath-1$. This and (A.6) imply $\lambda_\jmath^{(q)} \leq 0$. Also, $\bar{S}_{B_1}(t) \geq \bar{S}_B(t)$ implies $\bar{S}_{B_2}(t) > \bar{S}_B(t)$, so that a comparison of (A.3) and the string of equalities satisfied by $\bar{S}_B(t)$ gives

$$\lambda_{l-1}^{(q-1)} < \lambda_{l-1}^{(q)}, \tag{A.7}$$

$$\lambda_j^{(q)} - \lambda_{j-1}^{(q)} < \lambda_j^{(q-1)} - \lambda_{j-1}^{(q-1)}, \qquad j = l-1, \ldots, \jmath+2, \tag{A.8}$$

$$\lambda_{\jmath+1}^{(q)} - \lambda_\jmath^{(q)} < \lambda_{\jmath+1}^{(q-1)}. \tag{A.9}$$

By (A.7) and (A.8), we have $\lambda_j^{(q-1)} < \lambda_j^{(q)}$ for $j = \jmath+1, \ldots, l-1$. The last inequality and (A.9) then gives $0 > -\lambda_\jmath^{(q)}$, which contradicts $\lambda_\jmath^{(q)} \leq 0$. Thus we have $\bar{S}_{B_1}(t) < \bar{S}_B(t)$ instead, which implies $\lambda_\jmath^{(q)} > 0$ by (A.4)–(A.6) with $\leq$ replaced by $>$. Next, suppose $\bar{S}_B(t) \geq \bar{S}_{B_2}(t)$. Then by (A.7)–(A.9) with $>$ replaced by $\leq$, we can show $0 \leq \lambda_\jmath^{(q)}$, which contradicts with $\lambda_\jmath^{(q)} > 0$ obtained from $\bar{S}_{B_1}(t) < \bar{S}_B(t)$. Thus we have $\bar{S}_B(t) < \bar{S}_{B_2}(t)$.

Since $\bar{S}_{B_1}(t) < \bar{S}_B(t)$, we have $\lambda_j^{(q)} > \lambda_j^{(q-1)}$ for $j = \iota, \ldots, \jmath-1$ and $\lambda_\jmath^{(q)} > 0$, using a similar argument as before but replacing $\leq$ in (A.4)–(A.6) with $>$. Also,

$\bar{S}_B(t) < \bar{S}_{B_2}(t)$ implies $\lambda_j^{(q)} > \lambda_j^{(q-1)}$ for $j = \jmath + 1, \ldots, l-1$, as shown in the previous paragraph. This and the fact that $\lambda_j^{(q-1)} > 0$ for $j \in B \setminus \{\jmath, l\}$ by sweep $q-1$ lead to $\lambda_j^{(q)} > 0$ for $j \in B \setminus \{\jmath, l\}$. This and $\lambda_\jmath^{(q)} > 0$ complete the proof of (A.1).

APPENDIX B: PROOF OF THEOREM 1

As mentioned in Section 4.2, we use Lemmas 3 and 4 to prove Theorem 1; the proof of Lemma 3 is in the Supplementary Material, Section S.1.1 [Chang and McKeague (2019)]. Then the first assertion in Theorem 1 regarding asymptotic null distribution of K_n can be established using the continuous mapping theorem [van der Vaart (2000), Theorem 18.11]. Furthermore, by the uniform consistency $\hat{F}_0 \xrightarrow{P} F_0$ in $\ell^\infty([t_1, t_2])$ (see first paragraph in Section 3.1) we have weak convergence of $[-2\log\mathcal{R}, \hat{F}_0]^T$ to $[\mathrm{SSB}, F_0]^T$ in $(\ell^\infty([t_1, t_2]))^2$. Then by a remark following Proposition II.8.6 of Andersen et al. (1993) and the continuous mapping theorem, we obtain the second assertion in Theorem 1 regarding asymptotic null distribution of I_n.

It remains to prove Lemma 4. As a first step toward this, we establish the joint limiting behavior of $\bar{\Delta}_j(t)$. Here and in the sequel, all o_p and O_p terms are understood to hold uniformly over $t \in [t_1, t_2]$.

B.1. Joint limiting behavior of $\bar{\boldsymbol{\Delta}}_j(t)$. We first show that $\bar{\boldsymbol{\Delta}}(t) \equiv [\bar{\Delta}_1(t), \ldots, \bar{\Delta}_k(t)]^T$ is asymptotically equivalent to a time-dependent random linear transformation of the processes $\hat{U}_j(t)$, $j = 1, \ldots, k$. Then we use this result to characterize the limiting behavior of the $\bar{\Delta}_j^2(t)\hat{\sigma}_j^2(t)/n_j$ term in the asymptotic equivalence claim in Lemma 4.

We start by examining the estimating equations in (2.4). Since $\bar{\Delta}_j(t) = O_p(\sqrt{n})$ (by the Supplementary Material, Section S.2), a Taylor expansion gives

$$\begin{aligned}\sum_{i \le N_j(t)} &\log\{1 - d_{ij}/(r_{ij} + \bar{\Delta}_j(t))\} \\ &= \log \hat{S}_j(t) + \bar{\Delta}_j(t)\hat{\sigma}_j^2(t)/n_j + O_p(1/n).\end{aligned} \tag{B.1}$$

This and (2.4) give

$$\hat{\psi}_j(t) + M_j\bar{\Delta}_j(t)\frac{\hat{\sigma}_j^2(t)}{n_j} - M_{j+1}\bar{\Delta}_{j+1}(t)\frac{\hat{\sigma}_{j+1}^2(t)}{n_{j+1}} = O_p(1/n), \tag{B.2}$$

where recall from Section 3.2 that $\hat{\psi}_j(t) = M_j \log \hat{S}_j(t) - M_{j+1}\log \hat{S}_{j+1}(t)$, $j = 1, \ldots, k-1$. An equivalent matrix formulation of the above is

$$\boldsymbol{T}(t)\bar{\boldsymbol{\lambda}}(t)/n = -\hat{\boldsymbol{\psi}}(t) + O_p(1/n), \tag{B.3}$$

where $\hat{\boldsymbol{\psi}}(t) \equiv [\hat{\psi}_1(t), \ldots, \hat{\psi}_{k-1}(t)]^T$, $\bar{\boldsymbol{\lambda}}(t) \equiv [\bar{\lambda}_1(t), \ldots, \bar{\lambda}_{k-1}(t)]^T$, and $\boldsymbol{T}(t)$ is the $(k-1) \times (k-1)$ nonsingular tridiagonal matrix (with nonzero elements only on the main diagonal, and the first diagonals below and above the main diagonal) given in (S.8) in the Supplementary Material [Chang and McKeague (2019)]. A recursive algorithm for inverting tridiagonal matrices is available [Usmani (1994)], but in our case we can find an explicit expression for $\boldsymbol{T}^{-1}(t)$ (see Supplementary Material, Section S.1.2 for details). Multiplying both sides of $\bar{\boldsymbol{\lambda}}(t) = -n\boldsymbol{T}^{-1}(t)\{\hat{\boldsymbol{\psi}}(t) + O_p(1/n)\}$ by the $k \times (k-1)$ matrix $\boldsymbol{D}_M$ (which just depends on the M_j) in (S.13) in the Supplementary Material [Chang and McKeague (2019)], we obtain

$$\bar{\boldsymbol{\Delta}}(t) = \boldsymbol{D}_M \bar{\boldsymbol{\lambda}}(t) = -n\boldsymbol{D}_M \boldsymbol{T}^{-1}(t)\big\{\boldsymbol{D}_M^T \boldsymbol{A}(t) + O_p(1/n)\big\}, \tag{B.4}$$

where $\boldsymbol{A}(t) = [\log \hat{S}_1(t) - \log S_1(t), \ldots, \log \hat{S}_k(t) - \log S_k(t)]^T = [\sqrt{\hat{\theta}_1(t)}\hat{U}_1(t)/M_1, \ldots, \sqrt{\hat{\theta}_k(t)}\hat{U}_k(t)/M_k]^T/\sqrt{n}$. Then, as we show in Supplementary Material, Section S.1.2 [Chang and McKeague (2019)], by inserting the explicit expression for $\boldsymbol{T}^{-1}(t)$ we obtain

$$\big[\boldsymbol{D}_M \boldsymbol{T}^{-1}(t)\boldsymbol{D}_M^T\big]_{ij}(t) = \begin{cases} \dfrac{M_i^2}{\hat{\theta}_i(t)\hat{\phi}(t)} \displaystyle\sum_{l \neq i} \prod_{g \in E_l} \hat{\theta}_g(t), & i = j, \\ -\dfrac{M_i M_j}{\hat{\theta}_i(t)\hat{\phi}(t)} \displaystyle\prod_{g \in E_j} \hat{\theta}_g(t), & i \neq j, \end{cases} \tag{B.5}$$

where $\hat{\phi}(t) = \sum_{l=1}^k \prod_{g \in E_l} \hat{\theta}_g(t)$ is a consistent estimate of $\phi(t)$, which is defined along with E_l in Section 2.1. Inserting (B.5) into the previous display, and recalling that $\hat{\theta}_j(t) = nM_j^2\hat{\sigma}_j^2(t)/n_j$, algebraic manipulation then leads to

$$\bar{\Delta}_j^2(t)\frac{\hat{\sigma}_j^2(t)}{n_j} = \left(\hat{U}_j(t) - \frac{1}{\phi(t)} \sum_{l=1}^k \prod_{g \in E_l} \theta_g(t)\hat{U}_l(t)\frac{\sqrt{\theta_l(t)}}{\sqrt{\theta_j(t)}}\right)^2 + o_p(1) \tag{B.6}$$

for $j = 1, \ldots, k$. This allows us to express the leading term above in terms of the weights $w_j(t)$ and the $\hat{U}_j(t)$ processes as

$$w_j(t)\left(\frac{\hat{U}_j(t)}{\sqrt{w_j(t)}} - \check{U}(t)\right)^2. \tag{B.7}$$

Now we have handled the first part of the asymptotic equivalence claim in Lemma 4. The second part is handled in the next section.

B.2. Joint limiting behavior of $\tilde{\boldsymbol{\Delta}}_j(t)$. The argument is analogous to the equality constrained results in Appendix B.1, because our PAVA (see Appendix A) shows that $\tilde{\Delta}_j(t)$ for $j \in B$ is the equality constrained solution within the block

$B \in \mathcal{B}(t)$. The only difference is that we need to normalize the weight within each block, as we now show.

The analogue of (B.6) can be shown to be

$$\tilde{\Delta}_j^2(t)\frac{\hat{\sigma}_j^2(t)}{n_j} = \left(\hat{U}_j(t) - \frac{1}{\phi_B(t)}\sum_{l\in B}\prod_{g\in E_l^B}\theta_g(t)\hat{U}_l(t)\frac{\sqrt{\theta_l(t)}}{\sqrt{\theta_j(t)}}\right)^2 + o_p(1), \tag{B.8}$$

where $\phi_B(t) = \sum_{l\in B}\prod_{g\in E_l^B}\theta_g(t)$ and $E_l^B = B \setminus \{l\}$. When B contains only one element j, we have $\tilde{\Delta}_j(t) = 0$; in this case we define $\prod_{g\in E_l^B}\theta_g = \prod_{g\in\varnothing}\theta_g \equiv 1$ so that the leading term in (B.8) vanishes as it should. We can further organize the leading term of (B.8) into a weighted form analogous to (B.7):

$$w_j(t)\left(\frac{\hat{U}_j(t)}{\sqrt{w_j(t)}} - \hat{U}_B(t)\right)^2. \tag{B.9}$$

We have now shown the asymptotic equivalence claim in Lemma 4. The remaining claims are handled in the Supplementary Material, Section S.1.3 [Chang and McKeague (2019)].

Acknowledgments. The authors thank Hammou El Barmi and Shih-Hao Huang for helpful comments, and Pei-Yuan Tsai for computational support.

SUPPLEMENTARY MATERIAL

Supplement to "Nonparametric testing for multiple survival functions with noninferiority margins" (DOI: 10.1214/18-AOS1686SUPP; .pdf). The supplementary material provides remaining technical details for the proof of Theorem 1. The R code used for the simulations and the application are also provided.

REFERENCES

ANDERSEN, P. K., BORGAN, Ø., GILL, R. D. and KEIDING, N. (1993). *Statistical Models Based on Counting Processes*. Springer, New York. MR1198884

BERGER, R. L. and HSU, J. C. (1996). Bioequivalence trials, intersection-union tests and equivalence confidence sets. *Statist. Sci.* **11** 283–302. MR1445984

BORGAN, Ø. and LIESTØL, K. (1990). A note on confidence intervals and bands for the survival function based on transformations. *Scand. J. Stat.* **17** 35–41. MR1062844

BOYD, S. and VANDENBERGHE, L. (2004). *Convex Optimization*. Cambridge Univ. Press, Cambridge. MR2061575

CHANG, H., EL BARMI, H. and MCKEAGUE, I. W. (2016). Tests for stochastic ordering under biased sampling. *J. Nonparametr. Stat.* **28** 659–682. MR3555451

CHANG, H. and MCKEAGUE, I. W. (2016). Empirical likelihood based tests for stochastic ordering under right censorship. *Electron. J. Stat.* **10** 2511–2536. MR3545467

CHANG, H. and MCKEAGUE, I. W. (2019). Supplement to "Nonparametric testing for multiple survival functions with noninferiority margins." DOI:10.1214/18-AOS1686SUPP.

CHI, Y. (2002). Ordered tests for right-censored survival data. *Aust. N. Z. J. Stat.* **44** 367–380. MR1919199

DICKMAN, P. W., SLOGGETT, A., HILLS, M. and HAKULINEN, T. (2004). Regression models for relative survival. *Stat. Med.* **23** 51–64.

DIGNAM, J. J., HUANG, L., RIES, L., REICHMAN, M., MARIOTTO, A. and FEUER, E. (2009). Estimating breast cancer-specific and other-cause mortality in clinical trial and population-based cancer registry cohorts. *Cancer* **115** 5272–5283.

DYKSTRA, R. L. and FELTZ, C. J. (1989). Nonparametric maximum likelihood estimation of survival functions with a general stochastic ordering and its dual. *Biometrika* **76** 331–341. MR1016024

EL BARMI, H. and MCKEAGUE, I. W. (2013). Empirical likelihood-based tests for stochastic ordering. *Bernoulli* **19** 295–307. MR3019496

EL BARMI, H. and MUKERJEE, H. (2005). Inferences under a stochastic ordering constraint: The k-sample case. *J. Amer. Statist. Assoc.* **100** 252–261. MR2156835

FINNER, H. and STRASSBURGER, K. (2002). The partitioning principle: A powerful tool in multiple decision theory. *Ann. Statist.* **30** 1194–1213. MR1926174

FREITAG, G., LANGE, S. and MUNK, A. (2006). Non-parametric assessment of non-inferiority with censored data. *Stat. Med.* **25** 1201–1217. MR2225587

GEHAN, E. A. (1965). A generalized Wilcoxon test for comparing arbitrarily singly-censored samples. *Biometrika* **52** 203–223. MR0207130

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 Med. Res. Methodol.* **12** 1–13.

HAUSCHKE, D. and PIGEOT, I. (2005). Establishing efficacy of a new experimental treatment in the 'gold standard' design. *Biom. J.* **47** 782–786. MR2221912

HIDA, E. and TANGO, T. (2011). On the three-arm non-inferiority trial including a placebo with a prespecified margin. *Stat. Med.* **30** 224–231. MR2758874

KITAMURA, Y., SANTOS, A. and SHAIKH, A. M. (2012). On the asymptotic optimality of empirical likelihood for testing moment restrictions. *Econometrica* **80** 413–423. MR2920761

KOMBRINK, K., MUNK, A. and FRIEDE, T. (2013). Design and semiparametric analysis of non-inferiority trials with active and placebo control for censored time-to-event data. *Stat. Med.* **32** 3055–3066. MR3073783

LI, G. (1995). On nonparametric likelihood ratio estimation of survival probabilities for censored data. *Statist. Probab. Lett.* **25** 95–104. MR1365025

LIU, P. Y. and TSAI, W. Y. (1999). A modified logrank test for censored survival data under order restrictions. *Statist. Probab. Lett.* **41** 57–63. MR1659237

MCDANIEL, L. S., YU, M. and CHAPPELL, R. (2016). Sample size under the additive hazards model. *Clin. Trials* **13** 188–198.

MCKEAGUE, I. W. and ZHAO, Y. (2002). Simultaneous confidence bands for ratios of survival functions via empirical likelihood. *Statist. Probab. Lett.* **60** 405–415. MR1947180

MIELKE, M., MUNK, A. and SCHACHT, A. (2008). The assessment of non-inferiority in a gold standard design with censored, exponentially distributed endpoints. *Stat. Med.* **27** 5093–5110. MR2516746

MÜTZE, T., KONIETSCHKE, F., MUNK, A. and FRIEDE, T. (2017). A studentized permutation test for three-arm trials in the 'gold standard' design. *Stat. Med.* **36** 883–898. MR3606651

OWEN, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* **75** 237–249. MR0946049

OWEN, A. B. (2001). *Empirical Likelihood*. Chapman & Hall/CRC, Boca Raton, FL.

PARZEN, M. I., WEI, L. J. and YING, Z. (1997). Simultaneous confidence intervals for the difference of two survival functions. *Scand. J. Stat.* **24** 309–314. MR1481417

ROMANO, J. P. (2005). Optimal testing of equivalence hypotheses. *Ann. Statist.* **33** 1036–1047. MR2195627

ROTHMANN, M. D., WIENS, B. L. and CHAN, I. S. F. (2011). *Design and Analysis of Non-inferiority Trials*. CRC Press, Boca Raton, FL. MR2760027

THOMAS, D. R. and GRUNKEMEIER, G. L. (1975). Confidence interval estimation of survival probabilities for censored data. *J. Amer. Statist. Assoc.* **70** 865–871. MR0405766

USMANI, R. A. (1994). Inversion of a tridiagonal Jacobi matrix. *Linear Algebra Appl.* **212/213** 413–414. MR1306990

VAN DER VAART, A. W. (2000). *Asymptotic Statistics*. Cambridge Univ. Press, Cambridge. MR1652247

WELLEK, S. (2010). *Testing Statistical Hypotheses of Equivalence and Noninferiority*, 2nd ed. CRC Press, Boca Raton, FL. MR2676002

WHITE, W. B., WEBER, M. A., SICA, D., BAKRIS, G. L., PEREZ, A., CAO, C. and KUPFER, S. (2011). Effects of the angiotensin receptor blocker azilsartan medoxomil versus olmesartan and valsartan on ambulatory and clinic blood pressure in patients with stages 1 and 2 hypertension. *Hypertension* **57** 413–420.

INSTITUTE OF STATISTICAL SCIENCE
ACADEMIA SINICA
128 ACADEMIA ROAD, SECTION 2, NANKANG
TAIPEI 11529
TAIWAN (R.O.C.)
E-MAIL: hwchang@stat.sinica.edu.tw

DEPARTMENT OF BIOSTATISTICS
COLUMBIA UNIVERSITY
722 WEST 168TH STREET
NEW YORK, NEW YORK 10032
USA
E-MAIL: im2131@columbia.edu