A case study of non-inferiority testing with survival outcomes

Hsin-wen Chang
Academia Sinica, Taiwan
Ian W. McKeague
Columbia University, USA
Yu-Ju Wang
Academia Sinica, Taiwan

This is a case study for a new class of nonparametric tests designed to assess evidence of non-inferiority ordering among multiple survival functions. The tests are devised for tree-structured orderings, as needed for the comparison of an experimental treatment to one or more alternative treatments. Applications to data from two non-inferiority trials are developed: 1) a two-armed trial for the treatment of liver cancer in which we find strong evidence of the non-inferiority of an experimental treatment (lenvatinib) to a standard treatment (sorafenib), and 2) a three-armed trial for the treatment of major depression in which we find strong evidence that an experimental treatment is both superior to placebo and non-inferior to a standard treatment. We implement the approach in R*, and explain in detail how to carry out the analyses for 1) and 2).*

Keywords *: Censored data, Empirical likelihood, Stochastic ordering.*

1. Introduction

Non-inferiority trials have the goal of showing that an experimental treatment is no worse than a standard treatment up to an acceptably small margin. Assessing non-inferiority (or NI), however, is more complex than assessing superiority. For a recent discussion of the challenges in the design and interpretation of non-inferiority trials, see Mauri and D'Agostino (2017). Currently more than 600 NI trials are listed in MEDLINE. Key aspects in the design of NI trials are: 1) the need to have prior randomized clinical trials (RCTs) evaluating superiority of standard treatment to placebo, 2) the study must be able to distinguish between effective and ineffective therapies ("assay sensitivity"), 3) a NI margin must be defined during the design phase (based on effect of the standard treatment in previous studies), and 4) the NI trial should preserve the conditions of the study in which the standard treatment was shown to be effective ("constancy assumption"). NI trials typically deal with new therapies that are not anticipated to have efficacy *superior* to standard treatment, because they may have other benefits, such as being more economical, less extensive, or less toxic (Wellek,

2010; Rothmann et al., 2011).

In the present paper we provide a case study for the analysis of data from NI trials involving survival endpoints that are possibly right-censored. Our approach is based on the type of nonparametric likelihood ratio (NPLR) statistics used in empirical likelihood (Owen, 2001), and extends procedures developed by Chang and McKeague (2016, 2019) and El Barmi and McKeague (2013).

The classical notion of stochastic ordering plays a basic role in framing the notion of NI for survival outcomes. A survival function S_1 is said to be *stochastically larger* than another survival function S_2 if $S_1(t) \geq S_2(t)$ for all $t \geq 0$, and the inequality is strict for at least one t; then we write $S_1 \succ S_2$. Superiority testing involves distinguishing between $S_1 = S_2$ versus $S_1 \succ S_2$, after initially eliminating the possibility of *crossings* of S_1 and S_2, see Chang and McKeague (2016) for more details.

Chang and McKeague (2019) developed an empirical likelihood (EL) approach for analyzing NI trials with $k \geq 2$ arms, for possibly right-censored data, by establishing linear orderings among the transformed survival curves $S_1^{M_1} \succ S_2^{M_2} \succ \ldots \succ S_k^{M_k}$, where $M_1, \ldots, M_k > 0$ are pre-specified NI margins. Two test statistics were introduced: a supremum-type statistic K_n and an integral-type statistic I_n, obtained by either maximizing or integrating the log-NPLR statistic over the follow-up period, respectively. These tests can detect either local or cumulative differences among the transformed survival curves.

The contribution of the present paper is to study this approach for *tree-structured* orderings: $S_j^{M_j} \succ S_k^{M_k}, j = 1, \ldots, k-1$. These provide another way of establishing NI that is appropriate, for example, when the experimental treatment is compared separately to a standard therapy and to a placebo. We again use the test statistics K_n and I_n, except now the NPLR on which they are based is defined in terms of the tree-structured ordering.

We illustrate our new approach in a case study of two NI trials (Kudo et al., 2018; Mielke et al., 2008). The data are obtained from the published articles by digitizing Kaplan–Meier (KM) curves, and reconstructing survival and censoring information using the algorithm developed by Guyot et al. (2012).

The paper is organized as follows. In Section 2 we provide background concerning the two NI trials, and explain how the data sets were digitized from the KM curves provided in the published papers. In each case, we discuss in non-technical terms the motivation for the NI testing that is needed, and illustrate our proposed approach by comparing plots of (power-transformed) KM curves for each treatment group. Analytical methods for NI testing are discussed in detail in Section 3, and the results of applying the various procedures to the two NI trials are presented in Section 4. Concluding remarks are provided in Section 5. The R functions implementing our approach are described in Appendix A, and we provide a short proof regarding the gap between survival functions under proportional hazards in Appendix B.

2. Data sets

2.1. *Two-armed trial comparing treatments for liver cancer*

Sorafenib has been the only first-line treatment of hepatocellular carcinoma, the most common type of liver cancer. However, only 2% of patients respond to sorafenib. To expand patients' treatment options, medical researchers considered lenvatinib (an oral multikinase inhibitor like sorafenib), which has been approved for treating other types of cancers. After a phase 2 study showed efficacy and safety in patients with advanced hepatocellular carcinoma, lenvatinib was compared with sorafenib as a first-line treatment for unresectable hepatocellular carcinoma in a phase 3 randomized non-inferiority study. This study was conducted at 154 sites in 20 countries throughout the Asia-Pacific, North American, and European regions (including France).

The data were analyzed by Kudo et al. (2018)

to answer the following questions of interest: whether lenvatinib is non-inferior to sorafenib, and whether lenvatinib is superior to sorafenib. The primary endpoint is overall survival, the time from randomization until death from any cause. Kudo et al. (2018) obtained a significant non-inferiority result based on confidence intervals for the hazard ratio (assuming a Cox model), but superiority of lenvatinib over sorafenib was not established.

A total of 1492 patients were recruited for the trial, of whom 954 eligible patients were randomly assigned in a 1:1 ratio to receive either lenvatinib (n_1 = 478) or sorafenib (n_2 = 476) between March 1, 2013 and July 30, 2015. The total numbers of uncensored events in the lenvatinib and sorafenib arms were 351 and 350, respectively.

To reconstruct the survival and censoring information, we need to provide two inputs to the R code developed by Guyot et al. (2012). The first input is obtained using the open source software GetData Graph Digitizer to extract information from a copy of Figure 2 in Kudo et al. (2018), showing KM estimates of overall survival by treatment group, and numbers at risk every 3-months during follow-up (from 0 to 42 months). Figure 1 is a screenshot of GetData Graph Digitizer giving an approximation of the sorafenib survival curve (in red). The columns on the right show the coordinates of points (in yellow) selected by mouse clicks at each jump in the KM curve, including the initial point $(0, 1)$. This list of coordinates constitutes the first input.

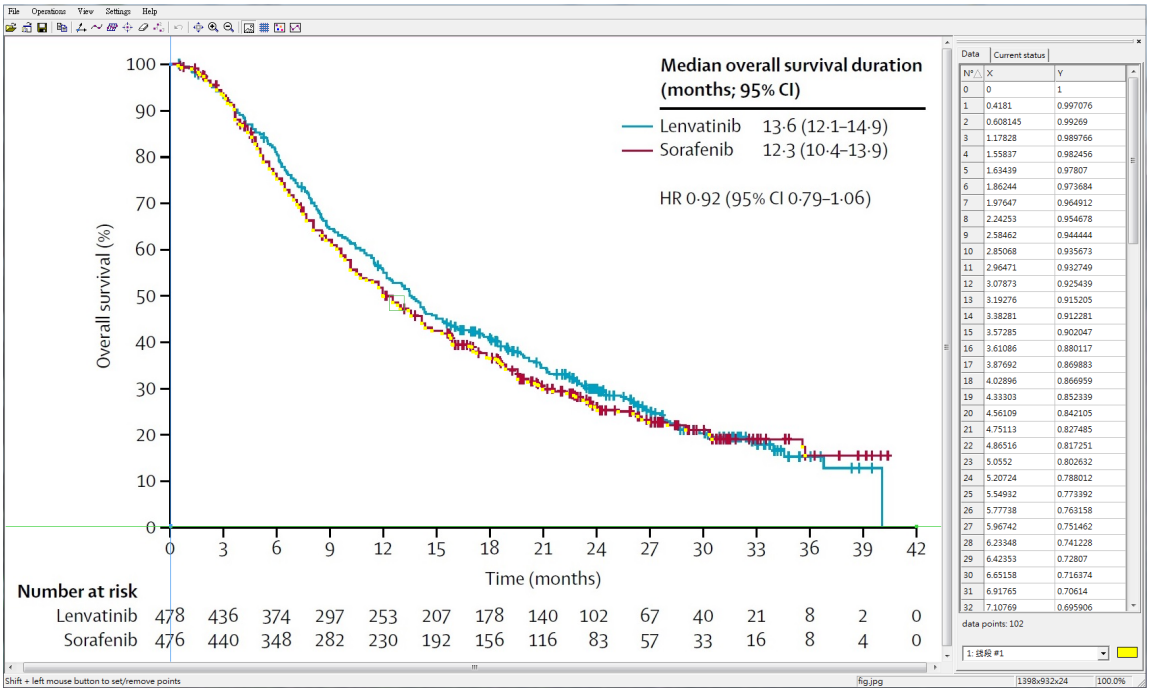

Figure 1: Screenshot showing use of the open source software GetData Graph Digitizer to reconstruct the survival data for the subjects treated with sorafenib.

Table 1: Counts of mouse clicks on the KM curve and subjects at risk during follow-up (sorafenib).

Interval	Time (months)	Lower	Upper	Size of risk set
1	0	1	12	476
2	3	13	28	440
3	6	29	39	348
4	9	40	51	282
5	12	52	59	230
⋮	⋮	⋮	⋮	⋮

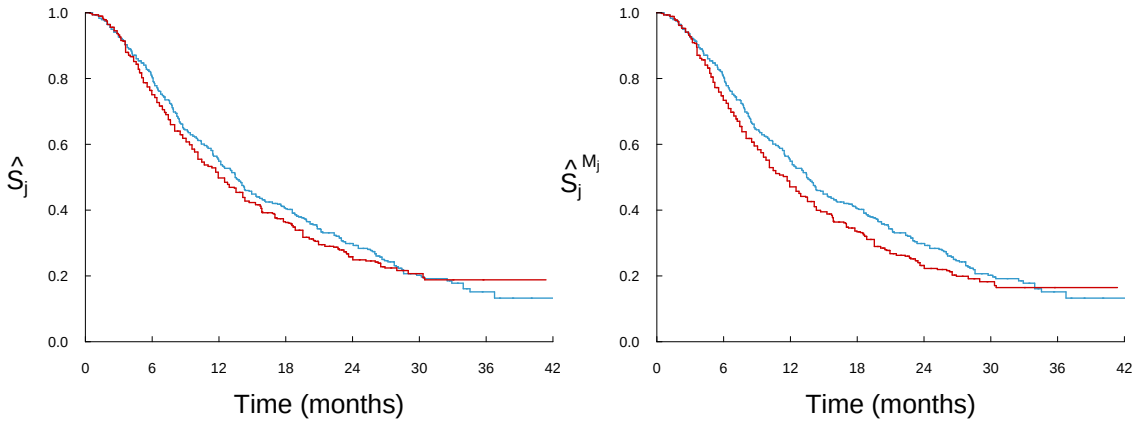

Figure 2: KM curves $\hat{S}_j$ (left) and $\hat{S}_j^{M_j}$ (right) from a RCT comparing two treatments for liver cancer: lenvatinib (cyan) and sorafenib (red).

The second input is contained in Table 1, where the first column indexes the time-intervals $[0,3), [3,6), \ldots$, the second column shows the time (in months) at the start of the interval, the third column shows the cumulative number of mouse clicks by the start of the interval, the fourth column shows the number of mouse clicks by the end of the interval, and the last column is the size of the risk set just before the start of the interval.

Based on the survival and censoring data reconstructed in this manner, the left panel of Figure 2 shows the KM curves for lenvatinib ($\hat{S}_1$) and sorafenib ($\hat{S}_2$). The reconstructions can have inaccuracies: comparing the left panel of Figure 2 with the original curves in Figure 1, there are some slight differences in the tail region. There is remarkably good agreement, however, between the hazard ratio based on the reconstructed data (0.93) and the original data (0.92).

It appears that the sorafenib survival curve lies below that of lenvatinib, possibly indicating superiority of lenvatinib over sorafenib. It is more challenging, however, to visually assess non-inferiority in terms of the KM curves. To this end, for reasons to be explained below, we need to use a transformed version of the sorafenib survival curve, $\hat{S}_2^{M_2}$, in place of $\hat{S}_2$, as displayed in the right panel of Figure 2.

The idea is to use the equivalence between the notion "treatment A is non-inferior to treatment B" and the notion "A is superior to a *worsened* version of treatment B." The right panel

of Figure 2 shows that the transformed survival curve for sorafenib (red line) has become lower (i.e., worsened), so the non-inferiority of lenvatinib can be assessed by whether its survival curve (with $M_1 = 1$ in order to keep the transformed $\hat{S}_1^{M_1}$ the same as the original) lies above the transformed sorafenib survival curve (using $M_2 = 1.08$ in $\hat{S}_2^{M_2}$). Here $M_2 = 1.08$ is the pre-specified non-inferiority margin used by Kudo et al. (2018) for the hazard ratio between lenvatinib and sorafenib. Note the actual values of the M_j do not matter, only their relative magnitudes $M_1 : M_2 = 1 : 1.08$.

The above transformation approach not only provides visual insight, but also enables us to translate the questions of interest into the language of survival functions: both the superiority and non-inferiority tests take the form of a test of whether the lenvatinib survival curve S_1 falls above a transformed version of the sorafenib survival curve $S_2^{M_2}$.

2.2. *Three-armed trial comparing treatments for major depression*

In this example, we consider data from a three-armed non-inferiority clinical trial involving treatments for major depression (Mielke et al., 2008). A similar digitization is conducted as in Section 2.1. The endpoint is time (in days) to first remission. We previously analyzed these data in Chang and McKeague (2019) to assess whether the experimental treatment group (n_1 = 262) is non-inferior to the standard treatment group (n_2 = 267), and if the standard treatment is superior to the placebo group (n_3 = 135). A significant result was obtained using the two-step NPLR test designed to detect local differences (K_n, $p < 0.01$), but the test based on the integral-type statistic I_n did not support the superiority of the standard treatment to the placebo.

Here we address another question that is of interest in practice (Hauschke and Pigeot, 2005; Kombrink et al., 2013): is the experimental treatment superior to the placebo as well as non-inferior to the standard therapy? While the comparison of the standard treatment to the placebo in Chang and McKeague (2019) is important to ensure the quality of the whole study, there are situations where it is known to be difficult to distinguish between the placebo and the standard therapy. The question in such studies is whether we can identify promising experimental treatments even if the standard treatment cannot be established as superior to placebo.

The left panel of Figure 3 shows the KM curves $\hat{S}_j$, $j = 1, 2, 3$, corresponding to the placebo, standard and experimental treatments, respectively. Because a shorter time to first remission is desirable, a lower survival function indicates a more effective treatment. It seems that the survival curve of the experimental treatment lies below that of the placebo, suggesting superiority of the experimental treatment over the placebo.

To address the question of non-inferiority, the idea is again to use the transformation approach discussed in the previous section. The right panel of Figure 3 suggests non-inferiority of the experimental treatment to the standard treatment because the survival curve of the experimental treatment (with $M_3 = 1$, keeping the transformed $\hat{S}_3^{M_3}$ the same as the original one) lies below the transformed survival curve of the standard treatment (with $M_2 = 0.7$). Here $M_2 = 0.7$ corresponds to a non-inferiority margin of the hazard ratio between the experimental and standard treatments, as explained in more detail in Section 3.

As before, translating our question of interest into the language of survival functions, we want to test whether the survival curve of the experimental treatment (S_3) lies below both the survival curve of the placebo (S_1), and the transformed survival curve $S_2^{M_2}$ of the standard treatment. In general, the approach is to first set the pair(s) of groups involved in the superiority hypothesis to have $M_j = 1$, and then set the remaining M_j according to the pre-specified non-inferiority margins.

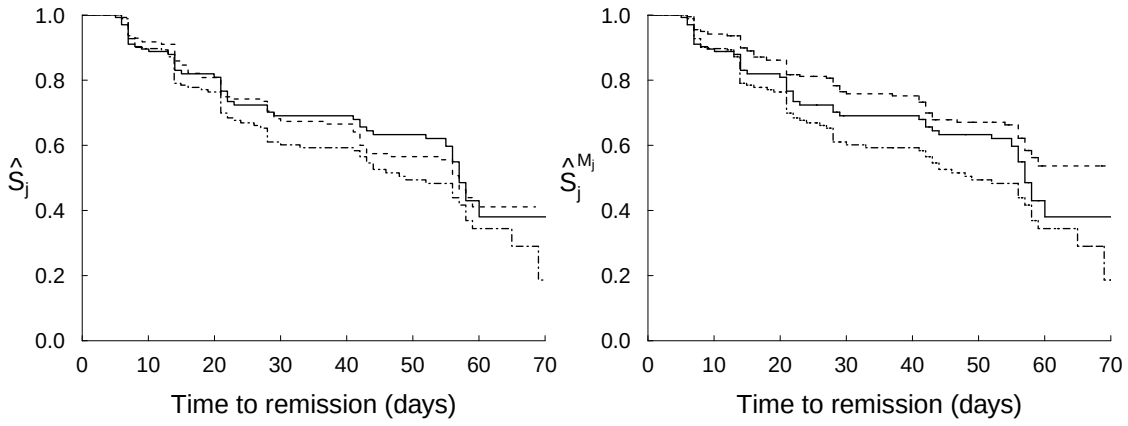

Figure 3: KM curves $\hat{S}_j$ (left) and $\hat{S}_j^{M_j}$ (right) from a non-inferiority trial comparing treatments of major depression: placebo (solid), standard treatment (dashed) and experimental treatment (two-dashed).

3. Methods

Let $S_1, \ldots, S_k$ be unknown survival functions corresponding to $k \geq 2$ treatments. A general framework for the primary research questions raised in Section 2 is to establish a tree-structured ordering of the form

$$H_1 \colon S_j^{M_j} \succ S_k^{M_k} \text{ for all } j = 1, \ldots, k-1, \quad (1)$$

where $M_1, \ldots, M_k > 0$ represent the pre-specified margins (which are informed by regulatory guidelines and previous clinical studies). As in the notion of stochastic ordering mentioned in the Introduction, for functions $f(t)$ and $g(t)$ of t over a given follow-up period $[t_1, t_2]$, we define $f \succ g$ to mean $f(t) \geq g(t)$ for all t with a strict inequality for some t. This means that the transformed survival function f lies above the transformed survival function g. The pairwise orderings can be interpreted using hazard ratios, as in Chang and McKeague (2019).

If the ordering $S_1^{M_1} \succ S_2^{M_2}$ holds, when longer survival is desirable, $M_1 \geq M_2$ represents superiority of treatment 1 over 2, whereas $M_1 < M_2$ corresponds to non-inferiority of 1 over 2. This is the case in Section 2.1, in which we want to show $S_1^{M_1} \succ S_2^{M_2}$, where S_1 represents the experimental therapy, and S_2 the standard therapy. Here $M_1 = 1$, and $M_2 = 1.08$ is a tolerable margin for the hazard ratio between the experimental and standard treatments.

On the other hand, if the ordering $S_j^{M_j} \succ S_3^{M_3}$ holds for $j = 1, 2$, when shorter survival is desirable, $M_j \geq M_3$ represents superiority of treatment 3 over j, whereas $M_j < M_3$ corresponds to non-inferiority of 3 over j. This is the case in Section 2.2, in which we want to show $S_1^{M_1} \succ S_3^{M_3}$ and $S_2^{M_2} \succ S_3^{M_3}$, where S_1 represents the placebo, S_2 the standard therapy, and S_3 the experimental therapy; $M_1 = M_3 = 1$, and $M_2 = 0.7$ is a tolerable margin for the hazard ratio between the experimental and standard treatments.

Our method for establishing (1) will be based on a similar two-step method described in Chang and McKeague (2019), but with modifications to accommodate the tree ordering in H_1. It 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

$$H_0 \colon S_1^{M_1} = \ldots = S_k^{M_k}. \quad (2)$$

The two-step procedure tests 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 first test is needed to eliminate departures from $H_0 \cup H_1$ (e.g., crossings or alternative orderings), and is based on a simultaneous confidence tube for the functions $M_j \log S_j(t) - M_k \log S_k(t)$, $j = 1, \ldots, k-1$.

The second test is needed to distinguish between H_0 and H_1 given that the first test rejects. Modifying the NPLR given in (2.2) of Chang and McKeague (2019) so the denominator is subject to the constraint $S_j(t) \leq S_k(t)$ for all $j = 1, \ldots, k-1$, the only change in the limiting distribution of the test statistics K_n and I_n (and their bootstrap calibration) is to use the projection corresponding to the tree-ordering instead of the linear-ordering. As before, we can show that the family-wise error rate of this two-step procedure can be controlled at the same alpha-level as the individual tests.

A competing method of testing H_1 is to test all the pairwise hypotheses of the form

$$H_0^j : S_j^{M_j} \preceq S_k^{M_k} \text{ versus } H_1^j : S_j^{M_j} \succ S_k^{M_k}, \quad (3)$$

for $j = 1, \ldots, k-1$. The combined procedure is to test whether at least one of the H_0^j holds versus the alternative that all of the H_1^j hold (i.e., H_1). This can be done using the intersection-union principle, which does not require a multiplicity adjustment (see, e.g., Berger and Hsu, 1996). Specifically, all of H_0^j need to be rejected at the prescribed overall α-level in order to have evidence in support of H_1. Possible choices of each pairwise test of H_0^j include our NPLR tests, or a Wald-type Cox model test, cf., Kombrink et al. (2013).

To analyze the data in Section 2.1, the latter approach involves fitting a single Cox model, with sorafenib as the reference group and an indicator covariate for treatment by lenvatinib. Denote the corresponding Cox regression coefficient as β_1. We then test the alternative $H_1^{NS} : \exp(\beta_1) < (M_2/M_1)$, where the superscript NS indicates that the hypothesis is non-inferiority or superiority. Since ordering in hazard functions implies ordering in survival functions, H_1^{NS} implies H_1 for $k = 2$.

To analyze the data in Section 2.2, a single Cox model is again used, with the placebo as the reference group and indicator covariates for the standard and experimental treatments. Denote the corresponding regression coefficients as β_2 and β_3, respectively. The combined pairwise Cox model test accepts the overall alternative $H_1^{S,E,N} = H_1^{S,E} \cap H_1^N$ if both $H_0^{S,E} : \exp(\beta_3) \leq M_1/M_3$ and $H_0^N : \exp(\beta_3) \leq (M_2/M_3)\exp(\beta_2)$ are rejected, where the superscript N indicates that the hypothesis is non-inferiority, and the superscripts S and E indicate that the hypothesis involves the standard and experimental treatments. Since ordering in hazard functions implies ordering in survival functions, $H_1^{S,E}$ implies H_1^1, H_1^N implies H_1^2, and $H_1^{S,E,N}$ implies H_1.

We also consider testing the individual pairwise alternatives $H_1^j, j = 1, \ldots, k-1$, using a Bonferroni adjustment. We compare our pairwise NPLR tests with the pairwise Cox model tests described above. Recall the drawback of using pairwise tests is that they are too conservative overall, even though they have the benefit of addressing the individual alternatives.

4. Results

4.1. *Two-armed trial*

In this section we apply the methods described in Section 3 to the liver cancer survival data discussed in Section 2.1.

Using $(M_1, M_2) = (1, 1.08)$, our two-step NPLR test based on the supremum-type statistic K_n ($p = 0.036$) and the integral-type statistic I_n ($p = 0.008$) both provide evidence of non-inferiority of lenvatinib over sorafenib, the evidence being particularly strong using I_n.

On the other hand, with $(M_1, M_2) = (1, 1)$, K_n ($p = 0.128$) and I_n ($p = 0.082$) both fail to provide evidence of superiority of lenvatinib over sorafenib. Intuitively, these results mean that in the right panel of Figure 2, both the maximal difference (as related to K_n) and the cumulative difference (as related to I_n) in the survival curves between lenvatinib and sorafenib are sufficiently large, with a particularly strong

cumulative difference; but in the left panel of Figure 2, both the maximal difference and cumulative difference in the survival curves between lenvatinib and sorafenib are not large enough.

Next we compare the result of the two-step NPLR test based on I_n (given above) with the Cox model approach, as both of them measure cumulative difference. The Cox model tests show non-inferiority of lenvatinib over sorafenib ($p = 0.024$, compared with $p = 0.008$ using the NPLR approach), but provide insufficient evidence of superiority of lenvatinib over sorafenib ($p = 0.169$, compared with $p = 0.082$ using the NPLR approach). We can see that in this data set, the NPLR approach gives more significant results compared with the Cox model approach. A possible explanation is that the (transformed) survival curves do not satisfy the proportional hazards assumption of the Cox model. An exploratory analysis of the right panel of Figure 2 (not a formal test) shows that the transformed survival curves cross, whereas there can be no crossings in survival curves that satisfy proportional hazards; see Appendix B for further explanation.

Table 2: Part of the data set `twoarm`. The variable names time, censor and group are, respectively, the survival time (in months), the indicator of non-censorship, and the treatment group label (1=lenvatinib, 2=sorafenib). In this display and the subsequent one, the survival time is rounded to three decimal places.

Subject	Time	Censor	Group
1	0.323	0	1
2	0.646	1	1
3	0.646	1	1
4	0.646	1	1
5	0.836	0	1
6	1.026	1	1
⋮	⋮	⋮	⋮
479	0.418	1	2
480	0.608	1	2
481	0.608	1	2
482	0.893	0	2
483	1.178	1	2
484	1.178	1	2
⋮	⋮	⋮	⋮

Using the R functions described in Appendix A, the steps to obtain the above results are as follows. Inspecting the data,

```
R> twoarm
```

gives output as in Table 2.

For the superiority test using the two-step NPLR test, the margins are specified by setting `M_vec=(1,1)`. The NI test (not shown) is carried out in the same way except with `M_vec =(1,1.08)`. First we run

```
R> nocrossings(twoarm,M_vec=c
    (1,1),group_k=2)
```

which gives

```
$decision
[1] 1
```

indicating that H_{01}^c is rejected, so we conclude there is no (significant) crossing or alternative ordering of the two survival functions (see Section 3). We can then proceed to the second part of the two-step test. For the test based on K_n, running

```
R> supELtest(twoarm,M_vec=c(1,1)
    ,group_k=2)
```

gives the value of the test statistic

```
$teststat
[1] 5.650466
```

the critical value based on bootstrap calibration

```
$critval
[1] 7.341974
```

and the p-value

```
$pvalue
[1] 0.128
```

Similarly, for the test based on I_n, running

```
R> intELtest(twoarm,M_vec=c(1,1)
    ,group_k=2)
```

gives

```
$teststat
[1] 1.301929
$critval
[1] 1.565695
$pvalue
[1] 0.082
```

For the Cox model-based Wald-type superiority test, run

```
R> M_vec=c(1,1)
R> dat=Surv(twoarm[,1],twoarm
   [,2])
R> Cox_model=coxph(dat~as.
   numeric(twoarm[,3]==1))
R> beta_S=Cox_model$coefficients
   [1]-log(M_vec[2]/M_vec[1])
R> sd_S=sqrt(diag(Cox_model$var)
   )[1]
R> T_S=beta_S/sd_S
R> pnorm(T_S)
```

resulting in the p-value `0.168538`.

4.2. *Three-armed trial*

In this section we apply the methods in Section 3 to the major depression data in Section 2.2. Our two-step NPLR tests based on K_n ($p < 0.001$) and I_n ($p < 0.001$) each establish that the experimental treatment is superior to the placebo, and non-inferior to the standard treatment. In the right panel of Figure 3, both the maximal and cumulative difference in the (transformed) survival curves between placebo and experimental treatment, and between standard and experimental treatment, are therefore deemed sufficiently large to give this conclusion. This example illustrates how evidence for the beneficial effect of an experimental treatment can arise, even when the standard treatment cannot be established as better than placebo (as seen in Section 2.2).

The above conclusion is supported by both the combined-pairwise Cox model test and the combined-pairwise NPLR test based on I_n. However, looking at the individual pairwise alternatives at the Bonferroni-corrected $\alpha = 0.025$, the pairwise Cox model tests show non-inferiority of the experimental treatment over the standard treatment ($p < 0.001$) but cannot establish superiority of the experimental treatment over the placebo ($p = 0.029 > \alpha$). The pairwise NPLR tests based on I_n give similar results, with a slightly larger p-value of 0.036 for the superiority test. These results reflect the fact that individual pairwise testing with multiplicity adjustment is more conservative than the combined-pairwise and the two-step NPLR approaches.

Turning now to the R code, inspecting the data

```
R> threearm
```

gives output as in Table 3.

Table 3: Part of the data set `threearm` with group labels (1=placebo, 2=standard, 3=experimental).

Subject	Time	Censor	Group
1	2.516	0	1
2	5.033	1	1
3	5.948	1	1
4	5.948	1	1
5	5.948	1	1
⋮	⋮	⋮	⋮
91	12.490	0	2
92	12.490	0	2
93	12.490	0	2
94	12.490	0	2
95	12.490	0	2
⋮	⋮	⋮	⋮
150	12.993	1	3
151	12.993	1	3
152	12.993	1	3
153	13.451	0	3
154	13.451	0	3
⋮	⋮	⋮	⋮

The NI margins are specified as `M_vec =(1,0.7,1)`. For the first step of the two-step NPLR test,

```
R> nocrossings(threearm,M_vec=c
   (1,0.7,1),group_k=3)
```

gives

```
$decision
[1] 1
```

indicating that there is no evidence of a crossing or an alternative ordering of the three survival functions. We then run the NPLR test based on K_n:

```
R> supELtest(threearm,M_vec=c
   (1,0.7,1),group_k=3)
$teststat
[1] 28.31915
```

```
$critval
[1] 8.160517
$pvalue
[1] < 0.001
```

The p-value is reported (as $p < 0.001$) up to the precision allowed by the 1000 bootstrap replications. For the NPLR test based on I_n:

```
R> intELtest(threearm,M_vec=c
   (1,0.7,1),group_k=3)
$teststat
[1] 6.505531
$critval
[1] 1.734457
$pvalue
[1] < 0.001
```

The pairwise NPLR tests based on I_n are run as follows:

```
R> threearm13=threearm[threearm
   [,3]!=2,]
R> threearm23=threearm[threearm
   [,3]!=1,]
R> p_In_13 = intELtest(
   threearm13,M_vec=c(1,1),
   group_k=3)$pvalue_numeric
$teststat
[1] 1.41794
$critval
[1] 1.220444
$pvalue
[1] 0.036
R> p_In_23 = intELtest(
   threearm23, M_vec = c(0.7, 1)
   , group_k = 3)$pvalue_numeric
$teststat
[1] 6.683489
$critval
[1] 1.013017
$pvalue
[1] < 0.001
```

The decision for the combined-pairwise NPLR test is

```
R> (combined_pairwise_In_rej_H0
   = p_In_13 < 0.05 & p_In_23 <
   0.05)
[1] TRUE
```

The decisions for the individual pairwise NPLR tests with Bonferroni adjustment are

```
R> (
   individual_pairwise_In_rej_H0j
    = c(p_In_13 < 0.025, p_In_23
    < 0.025))
[1] FALSE TRUE
```

for superiority and non-inferiority, respectively. For the pairwise Cox model tests, run

```
R> M_vec=c(1,0.7,1)
R> dat=Surv(threearm[,1],
   threearm[,2])
R> Cox_model=coxph(dat~as.
   numeric(threearm[,3]==2)+as.
   numeric(threearm[,3]==3))
R> Delta_Cox=M_vec[2]/M_vec[3]
R> beta_SE=
   Cox_model$coefficients[2]-log
   (M_vec[1]/M_vec[3])
R> sd_SE=sqrt(diag(Cox_model$var
   ))[2]
R> T_SE=beta_SE/sd_SE
R> beta_N=Cox_model$coefficients
   [2]-Cox_model$coefficients
   [1]-log(Delta_Cox)
R> sd_N=sqrt(diag(Cox_model$var)
   [2]-2*Cox_model$var[1,2]+diag
   (Cox_model$var)[1])
R> T_N=beta_N/sd_N
```

resulting in the p-values

```
R> (p_SE_Cox = 1 - pnorm(T_SE))
[1] 0.02900614
```

and

```
R> (p_N_Cox = 1 - pnorm(T_N))
[1] 4.047831e-06
```

The decision for the combined-pairwise Cox model test is

```
R> (combined_pairwise_Cox_rej_H0
    = p_SE_Cox < 0.05 & p_N_Cox
   < 0.05)
[1] TRUE
```

The decisions for the individual pairwise Cox model tests with Bonferroni adjustment are

```
R> (
    individual_pairwise_Cox_rej_H0j
     = c(p_SE_Cox < 0.025,
    p_N_Cox < 0.025))
[1] FALSE TRUE
```

for superiority and non-inferiority, respectively.

5. Conclusion

Non-inferiority testing for survival outcomes has previously been formulated in terms of Cox models, but a more flexible and powerful approach is to utilize nonparametric likelihood ratio techniques. In this paper we have provided a case study contrasting these two competing approaches using data from two clinical trials, along with R code implementation.

In the two-armed trial, we are able to demonstrate a much more significant non-inferiority result than the Cox model approach. In the three-armed trial, individual pairwise testing (with Bonferroni adjustment) is conservative; all the other tests show that the experimental treatment is both superior to placebo and non-inferior to the standard treatment.

The novelty of the proposed approach arises from framing the non-inferiority test in terms of tree-structured (rather than linear) ordering; this is of interest when comparing an experimental treatment to standard treatment(s) and separately to a placebo.

In future work it would be of interest to construct NPLR-based *confidence bands* comparing multiple treatment groups in NI trials, as often desired for visualization purposes (Althunian et al., 2017). Currently available confidence bands for comparing multiple survival functions are constructed based on pairwise contrasts, and are typically too wide to show NI due to the same inefficiency we have demonstrated for the individual pairwise tests. (To our knowledge, there is no confidence band corresponding to the combined-pairwise test.) NPLR-based confidence bands for ratios of survival functions developed by McKeague and Zhao (2002) apply to superiority trials, but to our knowledge no NPLR-based confidence bands have been developed for the study of non-inferiority ordering.

Acknowledgements

The research of Hsin-wen Chang was partially supported by Ministry of Science and Technology of Taiwan under Grant 106-2118-M-001-015-MY3. The research of Ian McKeague was partially supported by NIH Grants R01 GM095722 and R01 AG062401. The authors thank Shih-Hao Huang for helpful comments and the referee for many constructive suggestions.

References

Althunian, T. A., de Boer, A., Groenwold, R. H. H., and Klungel, O. H. (2017). Defining the noninferiority margin and analysing noninferiority: An overview. *British Journal of Clinical Pharmacology*, 83(8):1636–1642.

Berger, R. L. and Hsu, J. C. (1996). Bioequivalence trials, intersection-union tests and equivalence confidence sets. *Statistical Science*, 11(4):283–302.

Chang, H.-w. and McKeague, I. W. (2016). Empirical likelihood based tests for stochastic ordering under right censorship. *Electronic Journal of Statistics*, 10(2):2511–2536.

Chang, H.-w. and McKeague, I. W. (2019). Nonparametric testing for multiple survival functions with non-inferiority margins. *The Annals of Statistics*, 47(1):205–232.

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

Guyot, P., Ades, A. E., Ouwens, M. J. N. M., and Welton, N. J. (2012). Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan–Meier survival curves. *BMC Medical Research Methodology*, 12(1):1–13.

Hauschke, D. and Pigeot, I. (2005). Establishing efficacy of a new experimental treatment in the 'gold standard' design. *Biometrical Journal*, 47(6):782–786.

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. *Statistics in Medicine*, 32(18):3055–3066.

Kudo, M., S Finn, R., Qin, S., Han, K.-H., Ikeda, K., Piscaglia, F., Baron, A., Park, J., Han, G., Jassem, J., Frederic Blanc, J., Vogel, A., Komov, D., R Jeffry Evans, T., Lopez, C., Dutcus, C., Guo, M., Saito, K., Kraljevic, S., and Cheng, A.-L. (2018). Lenvatinib versus sorafenib in first-line treatment of patients with unresectable hepatocellular carcinoma: a randomised phase 3 non-inferiority trial. *Lancet (London, England)*, 391.

Mauri, L. and D'Agostino, R. B. (2017). Challenges in the design and interpretation of noninferiority trials. *New England Journal of Medicine*, 377(14):1357–1367.

McKeague, I. W. and Zhao, Y. (2002). Simultaneous confidence bands for ratios of survival functions via empirical likelihood. *Statistics & Probability Letters*, 60(4):405–415.

Mielke, M., Munk, A., and Schacht, A. (2008). The assessment of non-inferiority in a gold standard design with censored, exponentially distributed endpoints. *Statistics in Medicine*, 27(25):5093–5110.

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

Rothmann, M., Wiens, B., and Chan, I. (2011). *Design and Analysis of Non-Inferiority Trials*. Chapman & Hall/CRC Biostatistics Series. Taylor & Francis.

Wellek, S. (2010). *Testing Statistical Hypotheses of Equivalence and Noninferiority*. Chapman and Hall/CRC. Taylor & Francis.

Correspondence: im2131@columbia.edu.

Appendix A: R functions

Here we describe the arguments used in the three R functions we have written to implement the proposed tests. Only `data` and `M_vec` need to be modified depending on the application. The other arguments can be left unchanged.

```
R> intELtest(data,M_vec,group_k=max(data[,3]),t1=0,t2=Inf,nboot
   =1000,alpha=0.05,seed=1011)

R> supELtest(data,M_vec,group_k=max(data[,3]),t1=0,t2=Inf,nboot
   =1000,alpha=0.05,seed=1011)

R> nocrossings(data,M_vec,group_k=max(data[,3]),t1=0,t2=Inf,nboot
   =1000,alpha=0.05,seed=1011)
```

- `data`: a data frame/matrix with 3 columns: column 1 contains the observed survival/censoring times, column 2 the indicators of non-censorship, and column 3 the group labels.
- `M_vec`: vector of pre-specified margins $(M_1, M_2, \ldots, M_k)$.
- `group_k`: the label of the group hypothesized to have the smallest survival rate under the tree-structured hypothesis of the form H_1. The default value is the largest group label.
- `t1`: the lower endpoint of the pre-specified time interval over which comparison between the survival functions is carried out. The default value is 0.
- `t2`: the upper endpoint of the pre-specified time interval. The default value is ∞.
- `nboot`: the number of bootstrap replications. The default value is 1000.
- `alpha`: the pre-specified significance level of the tests. The default value is 0.05.
- `seed`: random number generator seed for the bootstrap replications. The default is 1011.

Appendix B: gaps between survival curves under proportional hazards

Let two survival curves S_1 and S_2 (that are possibly transformed from other survival curves) satisfy proportional hazards, in the sense that $S_1(t) = S_2^c(t)$ for some $c > 0$. Without loss of generality, suppose $c < 1$ (which means $S_1(t) \geq S_2(t)$), and $S_1(t) < 1$ for $t > 0$. Then the gap between the two survival curves is

$$S_1(t) - S_2(t) = S_2^c(t) - S_2(t) = S_2^c(t) \left\{ 1 - S_2^{1-c}(t) \right\}.$$

The gap vanishes if and only if either $S_2(t) = 0$ (implying $S_1(t) = 0$) or $S_2(t) = 1$ (implying $S_1(t) = 1$). The latter case arises only at $t = 0$. We conclude that there is a gap in the survival curves when $t > 0$ and $S_2(t) > 0$.