Nonparametric Predictive Multiple Comparisons of Lifetime Data

Tahani Coolen-Maturi*, Pauline Coolen-Schrijner†, Frank P.A. Coolen‡

Dept of Mathematical Sciences, Durham University, Durham DH1 3LE, UK.

Abstract

We consider lifetime experiments to compare units from different groups, where the units' lifetimes may be right-censored. Nonparametric predictive inference for comparison of multiple groups is presented, in particular lower and upper probabilities for the event that a specific group will provide the largest next lifetime. We include the practically relevant consideration that the overall lifetime experiment may be terminated at an early stage, leading to simultaneous right-censoring of all units still in the experiment.

Keywords: Crossing survival curves, early termination, lifetime data, lower and upper probabilities, nonparametric predictive inference, precedence testing, right-censoring.

1 Introduction

We consider comparison of lifetimes of units from multiple groups, simultaneously placed on an experiment. For each unit in the experiment, an event time will be observed, where 'event' is either its failure (or a related event of main interest in the study) or right-censoring, with the censoring mechanism assumed to be independent of the failure process. We include in our study explicitly the possibility that the experiment may be terminated early, which will incur a right-censored observation at the moment of termination for all units still in the experiment. This scenario is also considered in so-called 'precedence testing' (Balakrishnan and Ng, 2006).

*Email: tahani.maturi@durham.ac.uk (Corresponding author)

†Pauline died in April 2008, at that time the research presented in this paper was at an advanced stage.

‡Email: frank.coolen@durham.ac.uk

Actually, the method presented here does not require simultaneous experiments and can be used for multiple comparisons of groups of lifetime data quite generally, but the early termination aspect has been included as it is of clear practical relevance and as the way it is dealt with in our approach is attractive, as discussed in detail in the paper. We present Nonparametric Predictive Inference (NPI) (Augustin and Coolen, 2004; Coolen, 2006) for such situations, with uncertainty quantified by lower and upper probabilities for events that compare the failure times of one further unit from each group. Lower and upper probabilities generalize classical probabilities, and a lower (upper) probability for event A, denoted by $\underline{P}(A)$ ($\overline{P}(A)$), can be interpreted in several ways (Coolen, 2006): as supremum buying (infimum selling) price for a gamble on the event A, or as the maximum lower (minimum upper) bound for the probability of A that follows from the assumptions made. Informally, $\underline{P}(A)$ ($\overline{P}(A)$) can be considered to reflect the evidence in favour of (against) event A.

Suppose we have $k \geq 2$ independent groups, in classical statistics these tend to be referred to as 'populations'. We avoid the term 'populations' in NPI as we only consider one future observation and do not make use of any population distribution, even no assumptions about existence of such a distribution or about a meaningful population are made. For group j ($j = 1, \ldots, k$), n_j units are placed on a lifetime experiment, let their random times to failure be $X_{j,1}, \ldots, X_{j,n_j}$. In classical statistics, it is typically assumed that these random quantities are independent and identically distributed, with continuous distribution function F_j. Several nonparametric test methods have been proposed in the literature for comparing k groups of units placed simultaneously on a lifetime experiment.

In order to save time and cost, such a lifetime experiment may be ended before the event times of all units have been observed. Testing for differences between the lifetimes for the different groups, in such a setting, is known as 'precedence testing'. In classical statistics, there are several variations, for example with the experiment ending when a specific number of failures has been observed for a specific group. The actual stop criterion used affects the outcome of classical tests, as data that could alternatively have occurred influence the sampling distributions of the test statistics used. In NPI, the stop criterion does not explicitly affect inferences, because considerations of alternative data, that did not actually occur and that

depend on a specific stop criterion, play no role in NPI, which is similar to the likelihood principle in Bayesian statistics. However, one should take care that the basic exchangeability assumption underlying NPI is not trivially acceptable for all possible stop criteria, this is an interesting topic on foundations of statistics. It should be noted that similar considerations are also important in Bayesian statistics (de Cristofaro, 2004).

Classical precedence testing methods consider the null hypothesis $H_0 : F_1(x) = \ldots = F_k(x)$ for all x, which is tested against several alternative hypotheses, e.g. the most general alternative $H_1 : F_i(x) \neq F_j(x)$ for at least one pair of i and j and some value of x. Another alternative hypothesis that has been used is the one-sided alternative $H_2 : F_i(x) \leq F_1(x)$, with strict inequality for at least one $i = 2, 3, \ldots, k$ and some x. This is of particular use in applications where one wants to compare a control population, with distribution function F_1, to other populations, with distribution functions $F_2, \ldots, F_k$, to test if any of the other populations are better than the control population. Several tests for different alternative hypotheses are presented by Chakraborti and van der Laan (1997) and Chakraborti and Desu (1990). For given $p \in (0, 1)$, these tests typically depend on the statistics $U_{jp} = n_j \hat{F}_j \hat{F}_1^{-1}(p)$, $j = 2, 3, \ldots, k$, where $\hat{F}_j$ denotes the Kaplan-Meier estimator of $F_j(x)$ (Kaplan and Meier, 1958) and $\hat{F}_1^{-1}(u)$ is the Kaplan-Meier quantile function corresponding to $\hat{F}_1$. The asymptotic distribution of some functions of these statistics U_{jp} are given by Chakraborti and van der Laan (1997), who also present more details of such nonparametric precedence tests.

In the NPI approach presented in this paper, no null hypothesis is tested. Instead, different groups are compared by considering one further unit from each group, which is assumed to be exchangeable with those units that were actually tested for the corresponding group. One could, for example, interpret such inferences by considering a scenario in which for group j there were actually $n_j + 1$ units in the experiment, one of which had been randomly selected (all with equal probability) and for this unit no information is revealed, and it would be the 'future unit' involved in the events of interest considered in NPI. In practice, the different groups could for example relate to different treatments for one disease, in which case the NPI multiple comparisons could be interpreted as explicitly considering the effect the treatments would have on one future patient, as such it could give important guidance on choice of treatment for a patient.

The NPI approach uses lower and upper probabilities to quantify the uncertainties involved in the comparisons of such random quantities, this enables meaningful inferences without the need for further assumptions, and in applications it will often be an attractive way to compare different groups, as focus on one future unit per group enables many inferential problems to be formulated in a direct way. NPI also enables comparisons based on multiple future observations for each group, this is not considered in this paper.

A brief introduction to NPI is given in Section 2, where we also discuss the novelty of the results in this paper. The lower and upper probabilities for comparing k groups in order to select the best group, with possible early termination of the lifetime experiment, are presented in Section 3, and this approach is illustrated and discussed in three examples in Section 4, the last of which involves data with crossing empirical survival curves, a frequently discussed topic of research in the literature. Some concluding remarks are given in Section 5, including some comments on the practical use of the method introduced here.

2 Nonparametric predictive inference (NPI)

Nonparametric predictive inference (NPI) is based on Hill's assumption $A_{(n)}$ (Hill, 1968), which implies direct (lower and upper) probabilities for a future observable random quantity, based on observed values of n related random quantities (Augustin and Coolen, 2004; Coolen, 2006). NPI is suitable if there is little knowledge about random quantities of interest, other than the n observations, or if one does not want to use such information. Suppose that $X_1, \ldots, X_n, X_{n+1}$ are positive, continuous and exchangeable random quantities representing lifetimes. Let the ordered observed values of $X_1, \ldots, X_n$ be denoted by $x_1 < x_2 < \ldots < x_n < \infty$. Let $x_0 = 0$ and $x_{n+1} = \infty$ for ease of notation. We assume that no ties occur, our results can be generalised to allow ties (Hill, 1993). For positive X_{n+1}, representing a future observation, based on n observations, $A_{(n)}$ (Hill, 1968) is $P(X_{n+1} \in (x_i, x_{i+1})) = 1/(n+1)$ for $i = 0, 1, \ldots, n$.

$A_{(n)}$ does not assume anything else, and can be considered to be a post-data assumption related to exchangeability (De Finetti, 1974). Hill (1988) discusses $A_{(n)}$ in detail. Inferences based on $A_{(n)}$ are predictive and nonparametric, and can be considered suitable if there is

hardly any knowledge about the random quantity of interest, other than the n observations, or if one does not want to use such information, e.g. to study effects of additional assumptions underlying other statistical methods. $A_{(n)}$ is not sufficient to derive precise probabilities for many events of interest, but it provides bounds for probabilities via the 'fundamental theorem of probability' (De Finetti, 1974), which are lower and upper probabilities in interval probability theory (Walley, 1991; Weichselberger, 2001). Lawless and Fredette (2005) presented the concept of 'exact calibration' for frequentist inferences and, however without explicit reference to $A_{(n)}$, discussed that inferences based on $A_{(n)}$ are exactly calibrated for any value of n, so confirming the strong frequentist consistency properties of NPI.

Coolen and Yan (2004) presented rc-$A_{(n)}$ as a generalization of $A_{(n)}$ for right-censored data, using the extra assumption that, at a moment of censoring, the residual time-to-failure of a right-censored unit is exchangeable with the residual time-to-failure of all other units that have not yet failed or been censored. Their NPI method for dealing with right-censored observations can be considered as a predictive alternative to the Kaplan-Meier estimator, see Coolen and Yan (2004) for detailed discussion and examples. In addition, Coolen and Yan (2003) introduced NPI for comparison of two groups of lifetime data that contain right-censored observations, using the suitable rc-$A_{(n)}$ assumption per group. However, they did not consider situations with more than two groups. It is well known that one can generally not infer multiple comparisons from pairwise comparisons for all subsets of two groups, this is even stronger the case with statistical methods using lower and upper probabilities, for which combination rules are more complex than for precise probabilistic methods (Walley, 1991). Coolen and Yan (2003) did not study the effect of early termination of the lifetime experiment. Recently, Coolen-Schrijner et al. (2009) presented NPI for comparison of two groups of data with early termination of the experiment, say at time T_0, but they required all observations prior to T_0 to actually be failure times, so no right-censoring was possible apart from the right-censoring at T_0 of all units that had not yet failed. Maturi et al. (2010b) presented NPI for comparison of two groups of lifetime data for some specific progressive censoring scenarios, deriving closed-form expressions for these special cases. It should be emphasized that, thus far, NPI for comparison of more than two groups of lifetime data has not been presented in the literature, this is the main contribution

of the current paper. Thus, Section 3 presents an important generalization of the results by Coolen and Yan (2003) and Coolen-Schrijner et al. (2009), by developing NPI for comparison of multiple groups of lifetime data including right-censored observations, and also including possible early termination of the lifetime experiment.

3 NPI for lifetime data with early termination

In this section, we consider a life-testing experiment to compare units of $k \geq 2$ groups, which are assumed to be fully independent, with the experiment starting on all units at time 0. The experiment can be terminated before all units have failed, say at time T_0, which is assumed not to hold any information on residual time-to-failure for units that have not yet failed. We also allow right-censoring to occur before the experiment is stopped, due to a censoring process that is independent of the failure process. So we consider both right-censored observations in the original data and possible right-censoring due to stopping the experiment at T_0. For group j, $j = 1, \ldots, k$, n_j units are in the experiment, of which u_j units fail before (or at) T_0, with ordered failure times $x_{j,1} < x_{j,2} < \ldots < x_{j,u_j} \leq T_0$, and $c_{j,1} < c_{j,2} < \ldots < c_{j,v_j} < T_0$ are right-censoring times (we assume no tied observations for ease of notation, generalization is straightforward by considering limits, and the examples in Section 4 include ties). Let $x_{j,0} = 0$ and $x_{j,u_j+1} = \infty$ $(j = 1, \ldots, k)$. Let s_{j,i_j} be the number of right-censored observations in the interval $(x_{j,i_j}, x_{j,i_j+1}), i_j = 0, \ldots, u_j - 1$, with $x_{j,i_j} < c_{j,1}^{i_j} < \ldots < c_{j,s_{j,i_j}}^{i_j} < x_{j,i_j+1}$. Similarly, let s_{j,u_j} be the number of right-censored observations in the interval (x_{j,u_j}, T_0), with $x_{j,u_j} < c_{j,1}^{u_j} < \ldots < c_{j,s_{j,u_j}}^{u_j} < T_0$ and $\sum_{i_j=0}^{u_j} s_{j,i_j} = v_j$, so $n_j - (u_j + v_j)$ units from group j are right-censored at T_0. The following definition and lemma, presented by Coolen and Yan (2003), are needed to derive the lower and upper probabilities.

Definition 1. A partial specification of a probability distribution for a real-valued random quantity X can be provided via probability masses assigned to intervals, without any further restriction on the spread of the probability mass within each interval. A probability mass assigned, in such a way, to an interval (a, b) is denoted by $M_X(a, b)$, and referred to as M-function value for X on (a, b).

These M-functions are basic probability assignments in the sense of Shafer (1976), for which the following lemma, which will be used in the proof of Theorem 1, holds.

Lemma 1. For $s \geq 2$, let $J_l = (j_l, r)$, with $j_1 < j_2 < \ldots < j_s < r$, so we have nested intervals $J_1 \supset J_2 \supset \ldots \supset J_s$ with the same right end-point r (which may be infinity). We consider two independent real-valued random quantities, say X and Y. Let the probability distribution for X be partially specified via M-function values, with all probability mass $P(X \in J_1)$ described by the s M-function values $M_X(J_l)$, $l = 1, \ldots, s$, so $\sum_{l=1}^{s} M_X(J_l) = P(X \in J_1)$ because of the nested structure of the intervals $J_1, \ldots, J_s$. Then, without additional assumptions, $\sum_{l=1}^{s} P(Y < j_l) M_X(J_l) \leq P(Y < X, X \in J_1) \leq P(Y < r) P(X \in J_1)$, provides the maximum lower and minimum upper bounds.

To compare the k groups, we consider a hypothetical further unit from each group which would also have been involved in this experiment, with X_{j,n_j+1} the random failure time for the further unit from group j, assumed to be exchangeable with the failure times of the n_j units of the same group included in the experiment. The assumption rc-$A_{(n_j)}$ implies the lower and upper probabilities presented in Theorem 1, which are optimal bounds under the assumptions made. We restrict attention to the events $X_{l,n_l+1} = \max\limits_{1 \leq j \leq k} X_{j,n_j+1}$, for $l = 1, \ldots, k$. The lower and upper probabilities, in Theorem 1, are specified with the use of the following definition.

Definition 2. For NPI with lifetime data containing right-censored observations, and with early termination of the experiment at time T_0, the assumption rc-$A_{(n_j)}$ implies that the following M-function values apply for a nonnegative random quantity X_{j,n_j+1}, on the basis of data consisting of u_j failure times and $(n_j - u_j)$ right-censored observations:

$$M_{i_j}^{j} = M_{X_{j,n_j+1}}(x_{j,i_j}, x_{j,i_j+1}) = \frac{1}{n_j + 1} \prod_{\{r: c_r < x_{j,i_j}\}} \frac{\tilde{n}_{j,c_r} + 1}{\tilde{n}_{j,c_r}}$$

$$M_{i_j,t_j}^{j} = M_{X_{j,n_j+1}}(c_{j,t_j}^{i_j}, x_{j,i_j+1}) = \frac{1}{(n_j + 1)} (\tilde{n}_{j,c_{j,t_j}^{i_j}})^{-1} \prod_{\{r: c_r < c_{j,t_j}^{i_j}\}} \frac{\tilde{n}_{j,c_r} + 1}{\tilde{n}_{j,c_r}}$$

$$M_{T_0}^{j} = M_{X_{j,n_j+1}}(T_0, \infty) = \frac{n_j - (u_j + v_j)}{n_j + 1} \prod_{\{r: c_r < T_0\}} \frac{\tilde{n}_{j,c_r} + 1}{\tilde{n}_{j,c_r}}$$

where $i_j = 0, \ldots, u_j$, $t_j = 1, \ldots, s_{j,i_j}$, and $\tilde{n}_{j,c_r}$ and $\tilde{n}_{j,c_{j,t_j}^{i_j}}$ are the number of units from group j in the risk set just prior to time c_r and $c_{j,t_j}^{i_j}$, respectively. Also

$$P_{i_j}^j = P(X_{j,n_j+1} \in (x_{j,i_j}, x_{j,i_j+1})) = \frac{1}{n_j+1} \prod_{\{r:c_r<x_{j,i_j+1}\}} \frac{\tilde{n}_{j,c_r}+1}{\tilde{n}_{j,c_r}}$$

$$P_{T_0}^j = P(X_{j,n_j+1} \in (T_0, \infty)) = M_{X_{j,n_j+1}}(T_0, \infty) = M_{T_0}^j$$

Theorem 1. *The lower and upper probabilities for the event that the next observed lifetime from group l will be the maximum of all next observed lifetimes for the k groups in the experiment, with one future lifetime per group considered, are*

$$\underline{P}^{(l)} = \underline{P}\left(X_{l,n_l+1} = \max_{1\le j\le k} X_{j,n_j+1}\right) = \sum_{i_l=0}^{u_l} \left\{ \prod_{\substack{j=1 \\ j\neq l}}^{k} \left[\sum_{i_j=0}^{u_j} 1(x_{j,i_j+1} < x_{l,i_l}) P_{i_j}^j \right] M_{i_l}^l \right.$$
$$\left. + \sum_{t_l=1}^{s_{l,i_l}} \prod_{\substack{j=1 \\ j\neq l}}^{k} \left[\sum_{i_j=0}^{u_j} 1(x_{j,i_j+1} < c_{l,t_l}^{i_l}) P_{i_j}^j \right] M_{i_l,t_l}^l \right\} + M_{T_0}^l \prod_{\substack{j=1 \\ j\neq l}}^{k} \sum_{i_j=0}^{u_j} 1(x_{j,i_j+1} < T_0) P_{i_j}^j \qquad (1)$$

$$\overline{P}^{(l)} = \overline{P}\left(X_{l,n_l+1} = \max_{1\le j\le k} X_{j,n_j+1}\right) = \sum_{i_l=0}^{u_l} P_{i_l}^l \prod_{\substack{j=1 \\ j\neq l}}^{k} \left\{ \sum_{i_j=0}^{u_j} 1\left(x_{j,i_j} < x_{l,i_l+1}\right) M_{i_j}^j \right.$$
$$\left. + \sum_{i_j=0}^{u_j} \sum_{t_j=1}^{s_{j,i_j}} 1(c_{j,t_j}^{i_j} < x_{l,i_l+1}) M_{i_j,t_j}^j + 1\left(T_0 < x_{l,i_l+1}\right) M_{T_0}^j \right\} + P_{T_0}^l \qquad (2)$$

Proof. We can write the probability for the event that the lifetime of the next future observation

from group l is the maximum of all next future lifetimes for the k groups as

$$P^{(l)} = P\left(X_{l,n_l+1} = \max_{1\le j\le k} X_{j,n_j+1}\right) = P\left(\bigcap_{\substack{j=1\\ j\neq l}}^{k}\left\{X_{j,n_j+1} < X_{l,n_l+1}\right\}\right)$$
$$= \sum_{i_l=0}^{u_l} P\left(\bigcap_{\substack{j=1\\ j\neq l}}^{k}\left\{X_{j,n_j+1} < X_{l,n_l+1}, X_{l,n_l+1}\in(x_{l,i_l}, x_{l,i_l+1})\right\}\right)$$
$$+ P\left(\bigcap_{\substack{j=1\\ j\neq l}}^{k}\left\{X_{j,n_j+1} < X_{l,n_l+1}, X_{l,n_l+1}\in(T_0,\infty)\right\}\right)$$

The lower probability is derived as follows

$$P^{(l)} \ge \sum_{i_l=0}^{u_l}\left\{P\left(\bigcap_{\substack{j=1\\ j\neq l}}^{k}\left\{X_{j,n_j+1} < x_{l,i_l}\right\}\right) M^l_{i_l}\right.$$
$$\left.+\sum_{t_l=1}^{s_{l,i_l}} P\left(\bigcap_{\substack{j=1\\ j\neq l}}^{k}\left\{X_{j,n_j+1} < c^{i_l}_{l,t_l}\right\}\right) M^l_{i_l,t_l}\right\} + P\left(\bigcap_{\substack{j=1\\ j\neq l}}^{k}\left\{X_{j,n_j+1} < T_0\right\}\right) M^l_{T_0}$$
$$= \sum_{i_l=0}^{u_l}\left\{\prod_{\substack{j=1\\ j\neq l}}^{k} P\left(X_{j,n_j+1} < x_{l,i_l}\right) M^l_{i_l} + \sum_{t_l=1}^{s_{l,i_l}}\prod_{\substack{j=1\\ j\neq l}}^{k} P\left(X_{j,n_j+1} < c^{i_l}_{l,t_l}\right) M^l_{i_l,t_l}\right\}$$
$$+ \prod_{\substack{j=1\\ j\neq l}}^{k} P\left(X_{j,n_j+1} < T_0\right) M^l_{T_0}$$
$$\ge \sum_{i_l=0}^{u_l}\left\{\prod_{\substack{j=1\\ j\neq l}}^{k}\left[\sum_{i_j=0}^{u_j} 1\big(x_{j,i_j+1} < x_{l,i_l}\big) P^j_{i_j}\right] M^l_{i_l}\right.$$
$$\left.+\sum_{t_l=1}^{s_{l,i_l}}\prod_{\substack{j=1\\ j\neq l}}^{k}\left[\sum_{i_j=0}^{u_j} 1\Big(x_{j,i_j+1}<c^{i_l}_{l,t_l}\Big) P^j_{i_j}\right] M^l_{i_l,t_l}\right\} + \prod_{\substack{j=1\\ j\neq l}}^{k}\left[\sum_{i_j=0}^{u_j} 1\big(x_{j,i_j+1}<T_0\big) P^j_{i_j}\right] M^l_{T_0}$$

The first inequality follows by putting all masses according to the M-functions for X_{l,n_l+1}

corresponding to the intervals (x_{l,i_l}, x_{l,i_l+1}), for $i_l = 0, 1, \ldots, u_l$, and to (T_0, ∞), in the left end points of these intervals, and by using Lemma 1 for the nested intervals. The second inequality follows by putting all M-function masses for X_{j,n_j+1}, for $j = 1, \ldots, k, j \neq l$, corresponding to the intervals (x_{j,i_j}, x_{j,i_j+1}), with $(i_j = 0, 1, \ldots, u_j)$, and to (T_0, ∞), in the right end points of these intervals. The upper probability is obtained in a similar way, but now all M-function masses for the random quantities involved are put at the opposite end points of the respective intervals, which leads to

$$
\begin{aligned}
P^{(l)} &\leq \sum_{i_l=0}^{u_l} P\left(\bigcap_{\substack{j=1 \\ j\neq l}}^{k} \left\{X_{j,n_j+1} < x_{l,i_l+1}\right\}\right) P_{i_l}^l + P\left(\bigcap_{\substack{j=1 \\ j\neq l}}^{k} \left\{X_{j,n_j+1} < \infty\right\}\right) P_{T_0}^l \\
&= \sum_{i_l=0}^{u_l} P_{i_l}^l \prod_{\substack{j=1 \\ j\neq l}}^{k} P\left(X_{j,n_j+1} < x_{l,i_l+1}\right) + P_{T_0}^l \\
&\leq \sum_{i_l=0}^{u_l} P_{i_l}^l \prod_{\substack{j=1 \\ j\neq l}}^{k} \left\{ \sum_{i_j=0}^{u_j} 1\left(x_{j,i_j} < x_{l,i_l+1}\right) M_{i_j}^j \right. \\
&\quad \left. + \sum_{i_j=0}^{u_j} \sum_{t_j=1}^{s_{j,i_j}} 1\left(c_{j,t_j}^{i_j} < x_{l,i_l+1}\right) M_{i_j,t_j}^j + 1\left(T_0 < x_{l,i_l+1}\right) M_{T_0}^j \right\} + P_{T_0}^l
\end{aligned}
$$

□

It should be noted that the bounds above are sharp, in the sense that there are indeed possible assignments of the probabilities, in accordance to the M-function values, for which these bounds are actually attained, and hence they are valid lower and upper probabilities with strong consistency properties (Augustin and Coolen, 2004). It is easily seen that the value of T_0 only influences these lower and upper probabilities through the u_j. If $u_l = 0$ then $\overline{P}^{(l)} = 1$, while if $u_j = 0$ for at least one $j \neq l$ then $\underline{P}^{(l)} = 0$. If the experiment is terminated before a single unit has failed, then $\underline{P}^{(l)} = 0$ and $\overline{P}^{(l)} = 1$ for all groups. These extreme cases illustrate an attractive feature of lower and upper probabilities in quantifying the strength of statistical information, in an intuitive manner that is not possible with precise probabilities. If T_0 increases, $\underline{P}^{(l)}$ never decreases and $\overline{P}^{(l)}$ never increases, and they can only change if further events are observed.

This can be of great benefit on deciding when, during an experiment, the information provided by the experiment is sufficient to terminate it, for example if a lower probability for an event of interest has exceeded a chosen threshold value which is deemed to provide sufficient evidence in favour of this event.

If the experiment is not ended before event times for all units have been observed (whether the units have failed or were right-censored), then the terms including T_0 in formulae (1) and (2) disappear, and we get an extension of the results by Coolen and Yan (2003), who only considered NPI for comparison of two groups of lifetime data. In this case, the lower and upper probabilities are

$$\underline{P}^{(l)} = \sum_{i_l=0}^{u_l} \left\{ \prod_{\substack{j=1 \\ j \neq l}}^{k} \left[\sum_{i_j=0}^{u_j} 1(x_{j,i_j+1} < x_{l,i_l}) P_{i_j}^{j} \right] M_{i_l}^{l} + \sum_{t_l=1}^{s_{l,i_l}} \prod_{\substack{j=1 \\ j \neq l}}^{k} \left[\sum_{i_j=0}^{u_j} 1(x_{j,i_j+1} < c_{l,t_l}^{i_l}) P_{i_j}^{j} \right] M_{i_l,t_l}^{l} \right\}$$

$$\overline{P}^{(l)} = \sum_{i_l=0}^{u_l} P_{i_l}^{l} \prod_{\substack{j=1 \\ j \neq l}}^{k} \left\{ \sum_{i_j=0}^{u_j} 1\big(x_{j,i_j} < x_{l,i_l+1}\big)\, M_{i_j}^{j} + \sum_{i_j=0}^{u_j} \sum_{t_j=1}^{s_{j,i_j}} 1(c_{j,t_j}^{i_j} < x_{l,i_l+1}) M_{i_j,t_j}^{j} \right\}$$

A special case also occurs if there are no right-censored observations before T_0, in this case our method generalizes the NPI results by Coolen-Schrijner et al. (2009), who considered NPI for comparison of two groups with early termination of the experiment, but without earlier right-censoring occurring.

At any value of T_0, we can state that the data provide a strong indication that group l is the best if $\underline{P}^{(l)} > \overline{P}^{(j)}$ for all $j \neq l$. Of course, this may not occur, and we may be happy to have data providing a weak indication that group l is the best. It might seem attractive to state that, if $\underline{P}^{(l)} > \underline{P}^{(j)}$ and $\overline{P}^{(l)} > \overline{P}^{(j)}$ for all $j \neq l$, there would be a weak indication that group l is the best. Indeed, if one has to select one group and there is a group for which such a weak indication of being best holds, then that is the natural candidate. However, such a weak indication can be very weak indeed, in particular as it can already occur for relatively small T_0, with $\underline{P}^{(l)}$ positive but very small, this is discussed in an example in Section 4. If such a weak indication holds for one group, and in addition one judges the lower probability of this group being best to be sufficiently high, then it seems a reasonable basis for the choice of this group as being the

best. In all these considerations, it is an advantage that the difference between corresponding lower and upper probabilities ($\overline{P}^{(l)} - \underline{P}^{(l)}$), also called the imprecision, reflects the amount of information available, and it decreases if more relevant information becomes available. If one judges this difference to be too large, or if one judges each lower probability of a group being best too small to base a choice on the information available, then clearly one must either get more information, e.g. by continuing the experiment or repeating the experiment with more units, or one could explore the use of other statistical approaches with more modeling assumptions.

4 Examples

In this section, three examples with data from the literature are presented to illustrate our new method. In the second example we briefly discuss a classical precedence testing alternative for the same data, and the third example considers a situation with crossing empirical survival curves.

Example I

We use a data set from Desu and Raghavarao (2004, p.263), representing the recorded times (months) until promotion at a large company, for 19 employees in $k = 3$ departments, which we refer to as 'groups' in line with terminology used throughout this paper. The data are as follows: For group 1: 15, 20^+, 36, 45, 58, 60 ($n_1 = 6$), for group 2: 12, 25^+, 28, 30^+, 30^+, 36, 40, 45, 48 ($n_2 = 9$), and for group 3: 30^+, 40, 48, 50 ($n_3 = 4$), where '+' indicates that the employee left the company at that length of service before getting promotion, hence this can be considered to be a right-censored observation (one could argue about whether or not this right-censoring process is independent of the promotion process, but as we only use this data set to illustrate our new method, and have no further circumstantial information, we do not address this in more detail). We consider at which department the data suggest that one needs to work the longest to get a promotion. In this example, as we are looking at maximum time till promotion, the 'best group' in terminology from Section 3, of course actually represents the department where one has to work the longest to achieve a promotion. This data set contains tied observations,

in NPI these are dealt with by assuming that they differ a very small amount, in such a way that the lower (or upper) probability of interest is smallest (largest) over all possible ways to break the ties, which is an attractive manner for dealing with ties.

We use these data to illustrate the NPI method proposed in this paper, for which we also wish to illustrate the effect of possible early termination of a lifetime experiment. To enable this, we now assume that the recorded times until promotion are all measured from the same moment in time, and we consider the effect on our inferences if, instead of having the complete data as given above, the differences in time to promotion were studied after T_0 months. In this case, all observations that are larger than T_0 are replaced by right-censored observations at T_0. For several values of T_0, the lower probabilities $\underline{P}^{(l)}$ and upper probabilities $\overline{P}^{(l)}$, for $l = 1, 2, 3$, are presented in Table 1. For all values of T_0, until it is greater than the largest observation in the data set (60), these lower and upper probabilities are also displayed in Figure 1. At no value for T_0 the data indicate strongly that one of the groups leads to longest time to promotion, according to the formulation of such indications as explained at the end of Section 3.

Table 1: NPI lower and upper probabilities, Example I

T_0	u_1	u_2	u_3	$\underline{P}^{(1)}$	$\overline{P}^{(1)}$	$\underline{P}^{(2)}$	$\overline{P}^{(2)}$	$\underline{P}^{(3)}$	$\overline{P}^{(3)}$
11	0	0	0	0	1	0	1	0	1
14	0	1	0	0	1	0	0.9029	0	1
17	1	1	0	0	0.8629	0	0.9029	0.0114	1
27	1	1	0	0	0.8629	0	0.9029	0.0114	1
33	1	2	0	0	0.8629	0	0.7974	0.0243	1
38	2	3	0	0	0.7140	0	0.6591	0.0887	1
42	2	4	1	0.0678	0.7140	0.0248	0.5398	0.1135	0.8332
47	3	5	1	0.0813	0.6148	0.0315	0.4341	0.1969	0.8332
49	3	6	2	0.1670	0.6148	0.0315	0.3542	0.2161	0.7475
52	3	6	3	0.2392	0.6148	0.0315	0.3542	0.2161	0.6617
59	4	6	3	0.2392	0.6148	0.0315	0.3542	0.2161	0.6617
61	5	6	3	0.2392	0.6148	0.0315	0.3542	0.2161	0.6617

As mentioned before, T_0 only influences the lower and upper probabilities considered here via the u_j, so the actually observed failure times, in the sense that, for increasing T_0, these lower and upper probabilities for each group are constant except when T_0 increases past an observed u_j. For example, for $15 < T_0 < 28$, the lower and upper probabilities for all three groups remain constant since no observed failure times are in this interval, even though there are two right-censored observations in this interval. These right-censored observations affect the

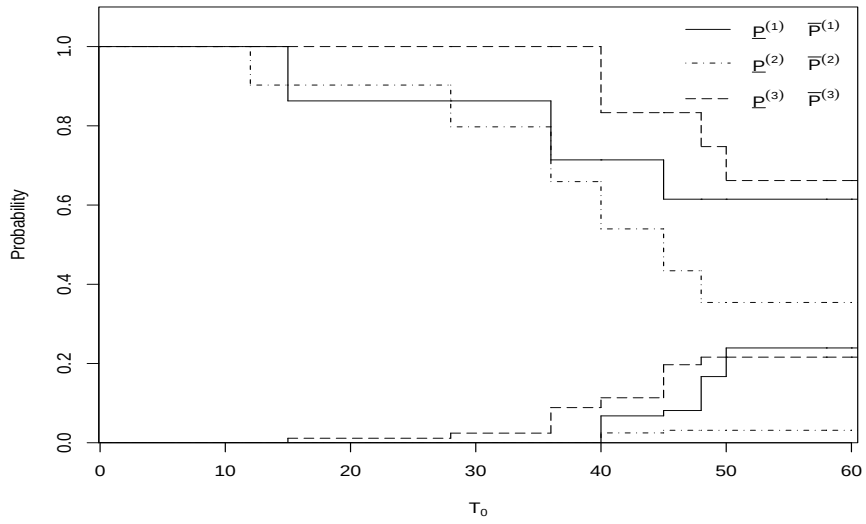

Figure 1: NPI lower and upper probabilities, Example I

lower and upper probabilities with larger values of T_0, at later failure times, as the jump sizes in these functions will increase. Effectively, our method considers the extreme configurations of the probability masses per group, according to the NPI M-functions, and if the largest observation before T_0 is a right-censoring time, this does not affect the most extreme orderings of the observations in different groups which correspond to the lower and upper probabilities, and hence there is no direct effect of such right-censoring times on the presented lower and upper probabilities.

At $T_0 = 28$, when the experiment would include the failure time 28 for a unit of group 2, the upper probability for group 2 to be best decreases and the lower probability for group 3 increases. However, the lower probabilities for group 1 and for group 2 to be best still remain 0 here, as there has not yet been an observed failure for group 3 at that moment in time, so the data do not exclude the possibility here that units in group 3 never fail. If the experiment is ended before the first failure of a particular group occurs, as is the case for group 3 at T_0 less than 40 in this example, than the extreme case corresponding to these lower probabilities for groups 1 and group 2, according to the NPI M-functions, allows the probability mass related to failure for units of group 3 to go to infinity, which explains why the lower probabilities for

group 1 and group 2 remain equal to 0 until T_0 increases past 40, the smallest time at which a unit of group 3 fails, and also why the upper probability for group 3 is equal to 1 for all T_0 up to 40.

If the experiment is stopped at $T_0 \in (15, 50)$, both the lower and upper probabilities for group 3 are greater than the lower and upper probabilities, respectively, for group 1 and group 2, as discussed before one could argue that this provides a weak indication that group 3 leads to the longest times until promotion. However, the very large imprecision in these lower and upper probabilities indicates that the evidence for such a claim is very weak, so care must be taken when formulating any conclusion along these lines. For larger values of T_0, such that event times for most units have been observed in the experiment, group 3 has most imprecision remaining, which reflects that there are only few observations for group 3.

Example II

In this example we use a data set also considered by Lee and Desu (1972), which gives leukemia remission times (in days) for patients undergoing three different treatments, so $k = 3$, and the numbers of patients per treatment are $n_1 = 25$, $n_2 = 19$ and $n_3 = 22$. The data are given in Table 2, where '+' again indicates that an observation is right-censored. In this example, 'better' means that a treatment leads to larger remission times.

Table 2: The remission times (in days) of leukemia patients

Treatment 1				Treatment 2			Treatment 3			
4	5	9	10	8	10	10	8	10	11	12^+
10	12	13	20^+	12	14	20	23	25	25	28
23	28	28	28	48	70	75	28	31	31	40
29	31	32	37	99	103	161^+	48	89	124	143
41	41	57	62	162	169	195	159^+	190^+	196^+	197^+
74	100	139	258^+	199^+	217^+	220	205^+	219^+		
269^+				245^+						

This data set was also used by Chakraborti and van der Laan (1997) to illustrate precedence testing, with Treatment 1 considered as a control treatment, and with focus on the median of the remission times for the control treatment, i.e. $\hat{F}_1^{-1}(0.5) = 29.39$, $\hat{F}_2(29.39) = 0.3334$ and $\hat{F}_3(29.39) = 0.4207$. They tested the null hypothesis that all treatments have the same effect, against the alternative that at least one of Treatments 2 or 3 is better than Treatment 1. They

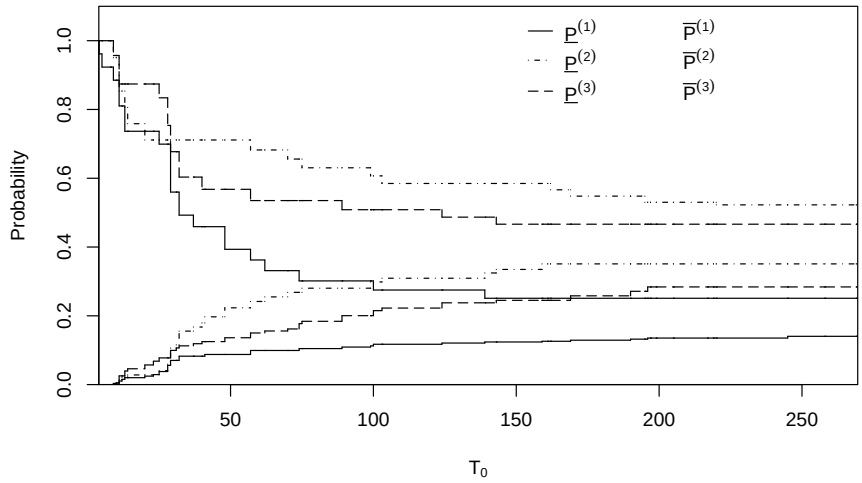

Figure 2: NPI lower and upper probabilities, Example II

concluded that, at 5% significance level, there is no evidence that any of the Treatments 2 or 3 is better than Treatment 1.

Table 3: NPI lower and upper probabilities, Example II

T_0	u_1	u_2	u_3	$\underline{P}^{(1)}$	$\overline{P}^{(1)}$	$\underline{P}^{(2)}$	$\overline{P}^{(2)}$	$\underline{P}^{(3)}$	$\overline{P}^{(3)}$
7	2	0	0	0	0.9232	0	1	0	1
9.5	3	1	1	0.0019	0.8851	0.0045	0.9505	0.0053	0.9570
10.7	5	3	2	0.0103	0.8102	0.0139	0.8535	0.0253	0.9156
15	7	5	3	0.0246	0.7366	0.0281	0.7586	0.0571	0.8741
24	8	6	4	0.0380	0.6992	0.0408	0.7113	0.0776	0.8339
30	12	6	8	0.0703	0.5598	0.1155	0.7113	0.1066	0.6772
33	14	6	10	0.0826	0.4927	0.1674	0.7113	0.1189	0.6034
42	17	6	11	0.0876	0.3934	0.2232	0.7113	0.1362	0.5678
71	19	8	12	0.1047	0.3313	0.2682	0.6558	0.1773	0.5351
101	21	10	13	0.1173	0.2750	0.3093	0.6069	0.2225	0.5085
150	22	11	15	0.1260	0.2510	0.3510	0.5848	0.2450	0.4664
165	22	12	15	0.1291	0.2510	0.3510	0.5664	0.2581	0.4664
170	22	13	15	0.1322	0.2510	0.3510	0.5480	0.2713	0.4664
198	22	14	15	0.1353	0.2510	0.3510	0.5302	0.2840	0.4664
221	22	15	15	0.1404	0.2510	0.3510	0.5226	0.2840	0.4664
270	22	15	15	0.1404	0.2510	0.3510	0.5226	0.2840	0.4664

This data set contains tied observations and we deal with them in the same manner as in Example I. Table 3 presents the NPI lower and upper probabilities for the events that Treatment l ($l = 1, 2, 3$) is the best, i.e. Treatment l leads to larger remission times than the other two treatments, for a number of times T_0 at which the experiment could have been stopped, where

as before all units for which no event had yet been observed at T_0 would be considered to be right-censored at T_0. These lower and upper probabilities are also displayed in Figure 2. If the experiment had been terminated at any time before 162 (i.e. $T_0 < 162$) then there would have been a weak indication that Treatments 2 and 3 are better than Treatment 1, since $\underline{P}^{(1)} < \underline{P}^{(j)}$ and $\overline{P}^{(1)} < \overline{P}^{(j)}$ for $j = 2, 3$. However, if we consider, for example, the case where the experiment would have been stopped at $T_0 = 162$, then the data would provide a strong indication that Treatments 2 and 3 are both better than Treatment 1, since $\overline{P}^{(1)} < \underline{P}^{(j)}$ for $j = 2, 3$. Of course, as these NPI lower (upper) probabilities never decrease (increase), the same indication holds if the experiment would have continued beyond time 162, no matter if or when it would have been terminated. This is an interestingly different conclusion than that reached by Chakraborti and van der Laan (1997), and is a good indication of the importance of using several statistical methods simultaneously, as further discussed in Section 5. It should be noted that, if in this example the experiment is not terminated before an event for each unit has been recorded (so $T_0 > 269$), then the NPI lower and upper probabilities corresponding to Treatment 3 have the largest imprecision, which is caused by the fact that for this treatment more observations are right-censored, particularly the larger observations, than for the other treatments.

Example III

This example considers comparison of $k = 2$ groups of lifetime data in a scenario where the empirical survival curves for the two groups cross. There is a substantial literature on statistical methods for such cases, typically because of the possible variations in model assumptions and alternative hypotheses when testing a null-hypothesis of 'equal distributions' for the lifetimes of both groups, see for example Le (2004) and Logan et al. (2008) for recent overviews and contributions. Fleming et al. (1980) presented data on survival of patients with bile duct cancer from a study to determine whether those treated with a combination of radiation treatment (RoRx) and 5-Fluorouracil (5-FU) would survive significantly longer than a control population. Survival times, in days, for the RoRx+5-FU Treatment group (22 patients) were 30, 67, 79^+, 82^+, 95, 148, 170, 171, 176, 193, 200, 221, 243, 261, 262, 263, 399, 414, 446, 446^+, 464, 777, while for the Control group (25 patients) the survival times were 57, 58, 74, 79, 89, 98, 101,

104, 110, 118, 125, 132, 154, 159, 188, 203, 257, 257.1, 431, 461, 497, 723, 747, 1313, 2636.

Figure 3 presents the Kaplan-Meier survival functions together with the NPI lower and upper survival functions (Coolen and Yan, 2004; Maturi et al., 2010a) for both the treatment and control groups, and with time transferred to months. Note that the figure does not provide the full curves, as including the right tails would make the main part of the plot harder to see. The Kaplan-Meier function is usually considered as the empirical survival curve, and it is bounded by the corresponding NPI lower and upper survival functions. Clearly, the curves for the different groups cross, with survival for members of the treatment group higher than for the control group early on, but later in the experiment the control group seems to have better survival prospects.

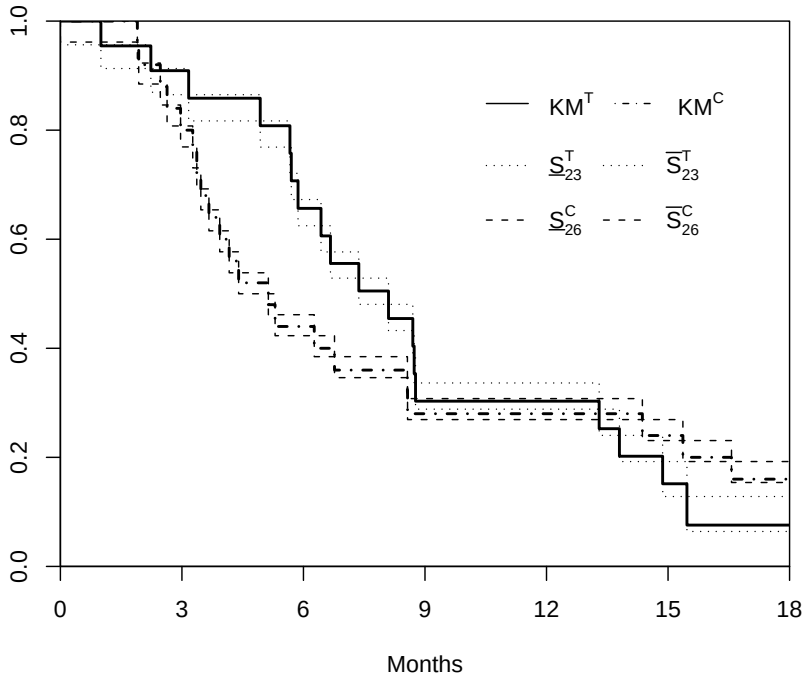

Figure 3: Kaplan-Meier (KM) and NPI lower and upper survival functions for Control and Treatment groups, Example III.

Fleming et al. (1980) applied three different hypothesis tests for these data, testing for equality of the two underlying population survival functions against the one-sided alternative

of the treatment providing longer survival. One of these tests rejected the null-hypothesis at 5% level of significance while the other two tests failed to do so. The reason is that different test statistics for such methods pick up on different aspects of the data, with some reflecting the crossing of the empirical survival curves while others do not pick up this feature.

The results of the NPI method, as presented in this paper, applied to these data are presented in Table 4 for a variety of values T_0, which are times at which it is assumed that the experiment would have been ended. The event considered is that the lifetime of the next unit from the treatment group is greater than the lifetime of the next unit from the control group, $X_{T,23} > X_{C,26}$. The NPI lower and upper probabilities for $T_0 = 88$ are actually the values that hold without early termination of the experiment, as all observations (either deaths or right-censoring times) are smaller than this value. So, taking the full data into account, we have $\underline{P}(X_{T,23} > X_{C,26}) = 0.5468$ and $\overline{P}(X_{T,23} > X_{C,26}) = 0.6345$. These values suggest that there is some evidence in favour of the suggestion that the treatment has a beneficial effect on survival. The NPI method does not involve testing of a hypothesis, but the lower and upper probabilities clearly quantify the strength of the evidence in favour of this specific event. The corresponding lower and upper probabilities in case of early termination of the experiment, so for different T_0 as given in Table 4, shows the effect of the crossing empirical survival functions quite clearly. As mentioned before, for increasing T_0 the NPI lower (upper) probability never decreases (increases), and early on the lower probability for the event $X_{T,23} > X_{C,26}$ increases substantially faster than the corresponding upper probability decreases, implying that early on there is substantially more evidence in favour of this event than against this event. From about $T_0 = 9$ however the upper probability decreases faster than the lower probability increases, which is caused by the same feature reflected by the crossing of the empirical survival curves. In addition to showing the effect of crossing empirical survival curves on the presented NPI method, this example also makes clear that it is often difficult to base a decision on 'best group' solely on the groups' empirical survival curves, and that different classical hypothesis testing methods can lead to varying conclusions by taking different aspects of the data into account. The NPI method provides an alternative way to quantify differences between groups of lifetime data that may often be attractive, in particular in situations where one would naturally be

interested in comparing the outcome of different treatments for a single next patient in each group, or indeed where different treatments can be applied to a single patient and one wishes to make an optimal choice for this patient in terms of longest survival time.

Table 4: NPI lower and upper probabilities for $X_{T,23} > X_{C,26}$, Example III.

T_0 months	$\underline{P}(X_{T,23} > X_{C,26})$	$\overline{P}(X_{T,23} > X_{C,26})$
1	0	0.9582
3	0.1704	0.9197
5	0.3903	0.8587
6	0.4495	0.7977
8	0.4938	0.7441
9	0.5271	0.6887
15	0.5345	0.6517
18	0.5419	0.6394
25	0.5468	0.6394
88	0.5468	0.6345

5 Concluding remarks

This paper has introduced NPI for comparing lifetimes of units from $k \geq 2$ independent groups, including the possibility that, if the data result from a lifetime experiment which started simultaneously for all units, this experiment may be ended before all event times have been observed or one simply wishes to consider the intermediate results at some point during an ongoing experiment, for example to see if evidence from data is already strong enough to end the experiment. For each unit, the event time recorded, if it happens before the experiment is ended, is either the time of an observed failure or a right-censoring time. Where classical frequentist methods in statistics tend to base such comparisons on hypotheses tests, the NPI approach directly compares random failure times of further units from these groups, which are assumed to be related to the observations per group through Hill's assumption $A_{(n)}$, or Coolen and Yan's assumption rc-$A_{(n)}$ if the data contain right-censored observations, without any further assumptions. Attention has been restricted to a single future observation per group, which can, conceptually, be generalized to multiple future observations per group, but the interdependence of such multiple future observations for a single group must be taken into account, and when the data for the group include right-censored observations NPI methods for doing this have not

yet been developed.

As for any new statistical method, it is important to consider how it can be applied. Of course, the assumption $A_{(n)}$ per group is crucial, if for example the data or knowledge about underlying processes are such that one does not consider the exchangeability assumption, implicit to $A_{(n)}$, to be appropriate, then this method should not be applied. All classical nonparametric methods in this application area tend to agree with this exchangeability assumption, but require additional assumptions (e.g. similarities in the probability distribution functions corresponding to different groups). One may well think that, in most applications, there is knowledge about the process and groups considered that can or should be taken into account, which NPI does not take on board. However, in all cases NPI can be considered to be a 'base-line method', it provides inferences without further assumptions or information, and this enables for example useful study of the outcomes of other statistical methods. If other methods lead to conclusions which differ substantially from those resulting from the NPI approach, then this will be due to the assumptions underlying the other method, which may often not have been made under complete awareness but more for mathematical convenience.

It is important to be aware of the fact that problems which appear to be identical are often formulated in substantially different manners in different statistical methods, with each method affected by specific features of the data. As such, we would strongly recommend the use of several statistical methods for a problem of interest, followed by careful study of the resulting inferences. Of course, one does then not report only outcomes of those methods which one may consider to be favourable, but has to report the total study with reasons for inclusion of specific methods, with particular focus on variability in conclusions corresponding to different statistical methods. If all such conclusions point in the same direction, then one can have great confidence in the inferences, but if this is not the case then the value of such an extensive study may well be even greater, as detailed understanding of the different outcomes is likely to provide more insight into specific aspects of the data and the actual inferential problem, as well as into the different methods used. It is somewhat peculiar that many statisticians, explicitly trained to study variability, seem to advocate the use of a single inferential approach, model and method for a given problem, while a detailed study using several methods would often

provide far greater understanding. In practice, careful reporting of such an extensive study using multiple approaches may often not be considered to be feasible, and time may even be too short to actually apply multiple methods. Nevertheless, we believe that the NPI approach presented in this paper provides an attractive different way to approach the important problem of multiple comparisons of different groups of lifetime data, and as it can be easily implemented its application, on its own when deemed suitable but more generally in combination with a variety of more established inferential methods, can lead to valuable insights and powerful inferences, particularly due to the strong consistency properties of NPI as briefly discussed in Section 2.

We consider it an advantage of NPI that, as clearly shown in the examples in this paper for smaller values of T_0, it may be that the lower and upper probabilities are so wide that they do not point towards clear decisions. This makes clear that, in order to derive stronger guidance, more information is needed, which in this application area would imply to either continue the experiment or to repeat it with more units involved. Of course, if there are no possibilities to gain further information, the wide bounds do not lead to indecision (an often heard criticism against the use of lower and upper probabilities), but they just make clear that the data and method used do not strongly indicate a preference for any of the groups. In this case, the data and NPI method may still provide some weak indications to support a specific choice, whereas alternative statistical methods, if they lead to a null hypothesis of 'equal probability distributions' not being rejected, would provide very little guidance on what group to choose if one must do so.

We only considered comparison of different groups by focusing on a single group being best, defined in terms of maximum value of the random lifetime for a future observation. Generalization to consider subsets of groups, either such that they contain the best one or that all selected groups are better than all not-selected groups, is achievable along the lines of Coolen and van der Laan (2001). Sometimes one may wish to compare different groups by focusing on different aspects, for example aiming for maximum lower probability of surviving a specific length of time. It is also of interest to consider the use of score-functions which enable both the lower and upper probabilities for events of interest, or the lower and upper survival functions,

to be taken into account. NPI provides opportunities to solve such problems, as long as they are formulated in a predictive manner, which may well be attractive in many applications, and may be considered to be more intuitive than formulation of multiple comparison problems via hypothesis tests.

Acknowledgement

We are grateful to three anonymous referees for their positive comments about the results in this paper and their constructive suggestions with regard to presentation.

References

Augustin, T. and Coolen, F. P. A. (2004). Nonparametric predictive inference and interval probability. *Journal of Statistical Planning and Inference*, 124(2):251–272.

Balakrishnan, N. and Ng, H. K. T. (2006). *Precedence-type tests and applications*. Wiley-Interscience, Hoboken, NJ.

Chakraborti, S. and Desu, M. M. (1990). Quantile tests for comparing several treatments with a control under unequal right-censoring. *Biometrical Journal*, 32(6):697–706.

Chakraborti, S. and van der Laan, P. (1997). An overview of precedence-type tests for censored data. *Biometrical Journal*, 39(1):99–116.

Coolen, F. P. A. (2006). On nonparametric predictive inference and objective bayesianism. *Journal of Logic, Language and Information*, 15(1-2):21–47.

Coolen, F. P. A. and van der Laan, P. (2001). Imprecise predictive selection based on low structure assumptions. *Journal of Statistical Planning and Inference*, 98(1-2):259–277.

Coolen, F. P. A. and Yan, K. J. (2003). Nonparametric predictive comparison of two groups of lifetime data. In Bernard, J. M., Seidenfeld, T. and Zaffalon, M., editors, *ISIPTA'03: Proceedings of the Third International Symposium on Imprecise Probabilities and their Applications*, pages 148–161, Waterloo (Canada). Proceedings in Informatics 18, Carleton Scientific.

Coolen, F. P. A. and Yan, K. J. (2004). Nonparametric predictive inference with right-censored data. *Journal of Statistical Planning and Inference*, 126(1):25–54.

Coolen-Schrijner, P., Maturi, T. A. and Coolen, F. P. A. (2009). Nonparametric predictive precedence testing for two groups. *Journal of Statistical Theory and Practice*, 3(1):273–287.

de Cristofaro, R. (2004). On the foundations of likelihood principle. *Journal of Statistical Planning and Inference*, 126(2):401–411.

De Finetti, B. (1974). *Theory of probability*. Wiley, London.

Desu, M. M. and Raghavarao, D. (2004). *Nonparametric statistical methods for complete and censored data*. Chapman & Hall/CRC, Boca Raton.

Fleming, T. R., O'Fallon, J. R., O'Brien, P. C. and Harrington, D. P. (1980). Modified Kolmogorov-Smirnov test procedures with application to arbitrarily right-censored data. *Biometrics*, 36(4):607–625.

Hill, B. M. (1968). Posterior distribution of percentiles: Bayes' theorem for sampling from a population. *Journal of the American Statistical Association*, 63(322):677–691.

Hill, B. M. (1988). De Finetti's theorem, induction, and a_n, or Bayesian nonparametric predictive inference (with discussion). In: Bernardo, J. M., DeGroot, M. H., Lindley, D. V. and Smith, A. F. M., editors, *Bayesian Statistics 3*, pages 211–241. Oxford University Press.

Hill, B. M. (1993). Parametric models for $a_{(n)}$: Splitting processes and mixtures. *Journal of the Royal Statistical Society. Series B (Methodological)*, 55(2):423–433.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481.

Lawless, J. F. and Fredette, M. (2005). Frequentist prediction intervals and predictive distributions. *Biometrika*, 92(3):529–542.

Le, C. T. (2004). Statistical methods for the comparison of crossing survival curves. In: Balakrishnan, N. and Rao, C. R., editors, *Handbook of Statistics*, Vol. 23, pages 277–289. Elsevier.

Lee, E. T. and Desu, M. M. (1972). A computer program for comparing k samples with right-censored data. *Computer Programs in Biomedicine*, 2(4):315–321.

Logan, B. R., Klein, J. P. and Zhang, M. J. (2008). Comparing treatments in the presence of crossing survival curves: an application to bone marrow transplantation. *Biometrics*, 64(3):733–740.

Maturi, T. A., Coolen-Schrijner, P. and Coolen, F. P. A. (2010a). Nonparametric predictive inference for competing risks. *Journal of Risk and Reliability*, 224(1):11–26.

Maturi, T. A., Coolen-Schrijner, P. and Coolen, F. P. A. (2010b). Nonparametric predictive comparison of lifetime data under progressive censoring. *Journal of Statistical Planning and Inference*, 140(2):515–525.

Shafer, G. (1976). *A Mathematical Theory of Evidence*. Princeton Universityh Press.

Walley, P. (1991). *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall.

Weichselberger, K. (2001). *Elementare Grundbegriffe einer allgemeineren Wahrscheinlichkeitsrechnung I. Intervallwahrscheinlichkeit als umfassendes Konzept*. Physika, Heidelberg.