Exceedance Probabilities Using Nonparametric Predictive Inference: A Supercentenarian Dataset Case

Ali M.Y. Mahnashi[1], Frank P. A. Coolen[2], Tahani Coolen-Maturi[2*]

[1]Department of Mathematics, College of Science, Jazan University, Jazan, Saudi Arabia.

[2*]Department of Mathematical Sciences, Durham University, Durham, DH1 3LE, United Kingdom.

*Corresponding author(s). E-mail(s): tahani.maturi@durham.ac.uk;
Contributing authors: amahnashi@jazanu.edu.sa;
frank.coolen@durham.ac.uk;

Abstract

Some statistical methods for extreme value analysis assume that the maximum observed value represents the endpoint of the support. However, in cases involving right-censored observations, it is often unclear whether the true value of a censored observation exceeds the largest observed value. This paper is inspired by the Supercentenarian dataset, which contains the ages at death of individuals who lived beyond 110 years, with right-censored data for those still alive at the time of data collection. This study employs Nonparametric Predictive Inference (NPI), a method that provides probability statements for a range of events of interest. NPI is a frequentist method that relies on minimal assumptions, focusing explicitly on future observations. It uses imprecise probabilities based on Hill's assumption $\boldsymbol{A}_{(n)}$ to quantify uncertainty. In this paper, we derive the probability that the true lifetime of at least one right-censored observation—or one of the future observations—exceeds the largest observed value. Furthermore, we extend this analysis to the exceedance of multiple largest observations, provided that they exceed the largest censored observation. We also investigate the time between any two of these largest observations, deriving the lower and upper probabilities for the exceedance of this time. We then demonstrate the proposed method using the Supercentenarian dataset, where the analysis is performed separately for men and women. We show how this approach can help to assess the likelihood of future extreme observations and provide insights into the validity of

assuming the largest observed value as the endpoint of support. This work highlights the strengths of NPI in handling right-censored data and its application to real-world datasets.

Keywords: Nonparametric Predictive Inference, Supercentenarian data, right-censored data, Exceedance

1 Introduction

Statistical methods for analysing extreme values typically assume that the largest value in a data set represents the upper bound of its support. However, this assumption may be problematic when the data set includes right-censored observations. In such cases, the true value of a censored observation could exceed the largest observed value. This paper is motivated by the literature on extreme value theory, where several studies assume that the maximum value of the random quantities under consideration corresponds to the largest observed value in the data set [1, 2]. Specifically, the paper draws inspiration from the Supercentenarian data set [2], which contains the ages at death of individuals who lived beyond 110 years. This data set includes right-censored observations for those who were still alive at the time of data collection. The focus of this paper is to investigate the likelihood that one or more of the right-censored observations correspond to a value greater than the largest observed value.

To address this, we propose using Nonparametric Predictive Inference (NPI), a predictive method that provides probability statements for various events of interest. In particular, we compute the probability that the actual lifetime of one or more right-censored observations exceeds the largest observed value, either for the censored data or for future observations. NPI is a frequentist method that makes minimal assumptions and focuses on quantifying uncertainty about future outcomes. It uses imprecise probabilities based on Hill's assumption $A_{(n)}$ to account for this uncertainty.

We extend this analysis to the exceedance of the second, third, fourth, and up to the j-th largest observations, as long as they exceed the largest censored observation. Additionally, we consider the time between any two of these largest observations and calculate the lower and upper probabilities for the exceedance of this time interval.

The paper is structured as follows: Section 2 introduces Hill's assumption $A_{(n)}$ and its generalisation for handling right-censored data. In Section 3, we analyse the exceedance of the largest observed value from the NPI perspective and extend this analysis to include future observations. Section 4 explores the exceedance of the j-th largest observations, considering the time between two of the largest values and calculating the lower and upper probabilities for the exceedance of a given time interval. The proposed methods are demonstrated using the Supercentenarian dataset in Section 5. Finally, Section 6 provides concluding remarks.

2 Nonparametric Predictive Inference (NPI)

Over the past few decades, Nonparametric Predictive Inference (NPI) has been developed for various data types and applied to a range of problems in statistics, as well as in fields like risk, reliability, operations research, and finance [3]. NPI is a statistical method that relies on minimal assumptions, particularly Hill's assumption $A_{(n)}$ [4], and uses imprecise probabilities to quantify uncertainty [5, 6].

Let $X_1, X_2, \ldots, X_n, X_{n+1}$ be real-valued, absolutely continuous, and exchangeable random quantities. The ordered observed values are denoted by $x_1 < x_2 < \cdots < x_n$, with $x_0 = -\infty$ and $x_{n+1} = \infty$ (or $x_0 = 0$ for non-negative random quantities) [7]. It is assumed that there are no ties among the data; if ties exist, they are handled by assuming small differences between tied observations, a common approach in statistics [8]. The observations divide the real line into $n+1$ intervals $I_j = (x_j, x_{j+1})$ for $j = 0, 1, \ldots, n$. Hill's assumption $A_{(n)}$ [9] states that the probability of the next observation X_{n+1} falling into any of these intervals is equally likely, i.e.,

$$P_{X_{n+1}}(x_j, x_{j+1}) = \frac{1}{n+1}, \quad \text{for} \quad j = 0, 1, \ldots, n. \tag{1}$$

NPI, based on Hill's assumption $A_{(n)}$, provides direct probabilities for future random quantities based on observed values. In NPI, uncertainty is quantified using lower and upper probabilities, which are the sharpest bounds for events of interest when $A_{(n)}$ is assumed to hold [5].

However, $A_{(n)}$ is not suitable for handling right-censored data. Coolen and Yan [10] introduced rc-$A_{(n)}$, a generalization of $A_{(n)}$, for right-censored observations. This generalization assumes that at the time of censoring, the residual lifetime of a censored observation is exchangeable with the residual lifetimes of other observations that have not failed or been censored. For handling censored data, two additional assumptions, the $\tilde{A}_{(n)}$ assumption and the shifted-$\tilde{A}_{(n)}$ assumption, are introduced. These assumptions differ in how probability mass is assigned to intervals or subintervals formed by failure and censoring times. In this work, we focus on the shifted-$\tilde{A}_{(n)}$ assumption, which allows for the application of $A_{(n)}$ but with the starting point shifted from 0 to the right-censoring time $c^i_{i^*}$.

Let $X_1, X_2, \ldots, X_n, X_{n+1}$ be non-negative, exchangeable, and continuous random quantities representing lifetimes. Suppose there are n total observations, including u failure time observations, $x_1 < x_2 < \cdots < x_u$, and $\nu = n - u$ right-censoring times, $c_1 < c_2 < \cdots < c_\nu$. For simplicity, let $x_0 = 0$ and $x_{u+1} = \infty$. Additionally, suppose there are s_i right-censored observations in the interval $I^i = (x_i, x_{i+1})$, denoted by $c^i_1 < c^i_2 < \cdots < c^i_{s_i}$, with $\sum_{i=1}^{u} s_i = \nu$, such that $c^i_{i^*} \in (x_i, x_{i+1})$ for $i = 0, 1, \ldots, u$ and $i^* = 1, 2, \ldots, s_i$. Let $X_{c^i_{i^*}}$ represent the random quantity corresponding to the right-censoring time $c^i_{i^*}$. The shifted-$\tilde{A}_{(n)}$ assumption [10, 11] specifies the probability distribution for $X_{c^i_{i^*}}$, conditioned on $X_{c^i_{i^*}} > c^i_{i^*}$, via the following M-function values:

$$M_{X_{c^i_{i^*}}}(x_k, x_{k+1}) = \frac{1}{\tilde{n}_{c^i_{i^*}} + 1}, \quad \text{for} \quad k = i+1, \ldots, u, \tag{2}$$

$$M_{X_{c^i_{i^*}}}(c^i_{i^*}, x_{k+1}) = \frac{1}{\tilde{n}_{c^i_{i^*}} + 1}, \tag{3}$$

$$M_{X_{c^i_{i^*}}}(c^i_l, \infty) = \frac{1}{\tilde{n}_{c^i_{i^*}} + 1}, \quad \text{for} \quad l = i^* + 1, \ldots, \nu, \tag{4}$$

where $\tilde{n}_{c^i_{i^*}}$ is the number of observations in the risk set at time $c^i_{i^*}$, for $c^i_{i^*} \in (x_i, x_{i+1})$, and $i^* = 1, 2, \ldots, s_i$. This assumption is consistent with the idea that exchangeability of random quantities in the risk set prior to censoring implies exchangeability for those that exceed a given censoring time, aligning with the assumption of non-informative censoring [10, 11].

In practice, when dealing with tied observations in NPI, it is common to assume that the tied observations differ by small amounts [8, 12]. If there is a tie between an event time and a right-censoring time, the standard approach is to assume that the right-censoring time occurs just after the event time [13]. In this paper, we assume there are no ties in the data, but the same approach is used if ties are present (as discussed in [8, 10, 12, 13]).

3 Exceedance of the largest observed value

Building on the $A_{(n)}$ assumption and the concept of non-informative right censoring described in Section 2, this section introduces a new method to determine the probability that the largest observed value in a dataset with right-censored observations will be exceeded. Specifically, the method addresses the question: *What is the probability that one or more lifetimes among the censored observations exceed the largest observed value?*

Let $X_1, X_2, \ldots, X_n$ be non-negative, exchangeable, and continuous random quantities. We have u observed event times, denoted by $x_1 < x_2 < \cdots < x_u$, and $v = n - u$ right-censored observations, with censoring times denoted by $c_1 < c_2 < \cdots < c_v$. For simplicity, we define $x_0 = 0$ and $x_{u+1} = \infty$. The random variable corresponding to a censored observation at time c_r is denoted by X_{c_r}, where $r = 1, 2, \ldots, v$. The largest observed event time in the dataset is represented by $\mathcal{R} = x_u$.

In addition to assuming exchangeability of $X_1, X_2, \ldots, X_n$, we adopt the assumption that, at any right-censoring time c_r, the remaining time until the event for a censored observation is exchangeable with the remaining times for all other observations in the risk set at c_r [10, 11]. Under non-informative right censoring [10, 11], we use the shifted-$\tilde{A}_{(n)}$ assumption. This assumption generalizes $A_{(n)}$ by shifting the reference point from 0 to the observed censoring time c_r, which partially specifies the distribution of X_{c_r} via the following M-function values:

$$M_{X_{c_r}}(x_i, x_{i+1}) = \frac{1}{\tilde{n}_{c_r} + 1}, \quad i = 0, \ldots, u \tag{5}$$

where $c_r \in (x_i, x_{i+1})$, and $\tilde{n}_{c_r}$ is the number of observations in the risk set just before c_r $(r = 1, 2, \ldots, v)$.

Using this framework, we can calculate the probability that at least one censored lifetime exceeds the largest observed event time $\mathcal{R}$, considering only the current dataset of n observations. For convenience, let $G_{\mathcal{R}}(0)$ denote this event of interest. This notation is used because we are focusing solely on the current dataset, with no future observations (i.e., $m = 0$) included. The probability for this event is expressed as:

$$P(G_{\mathcal{R}}(0)) = 1 - \prod_{r=1}^{v} \frac{\tilde{n}_{c_r}}{\tilde{n}_{c_r} + 1} \tag{6}$$

where $\tilde{n}_{c_r}$ represents the number of observations in the risk set (those still functioning or uncensored) just before c_r. The proof of this result and an illustrative example are provided in Appendix A.

Extending this analysis, we now consider predictions for future observations. Let $X_{n+1}, X_{n+2}, \ldots, X_{n+m}$ represent future non-negative, exchangeable, continuous random quantities. Define $\tilde{n}_{x_0} = n$, the initial size of the risk set at $x_0 = 0$, and recall that $\mathcal{R}$ denotes the largest observed event time. The event of interest is now extended to include whether at least one lifetime, either among the censored observations or among the $m \geq 1$ future individuals, exceeds $\mathcal{R}$. The probability for this extended event is given by:

$$P(G_{\mathcal{R}}(m)) = 1 - \left[\prod_{i=1}^{m} \frac{n+i-1}{n+i} \prod_{r=1}^{v} \frac{\tilde{n}_{c_r}}{\tilde{n}_{c_r} + 1} \right] = 1 - \left[\frac{n}{n+m} \prod_{r=1}^{v} \frac{\tilde{n}_{c_r}}{\tilde{n}_{c_r} + 1} \right] \tag{7}$$

where $\tilde{n}_{c_r}$ represents the number of observations in the risk set just before c_r.

This expression shows that $P(G_{\mathcal{R}}(m))$ increases as m grows. Notably, as $m \to \infty$, the second term tends to zero, leading to $P(G_{\mathcal{R}}(m)) \to 1$. This implies that as more individuals are included, the probability of exceeding $\mathcal{R}$ becomes increasingly likely. The proof of this result and an illustrative example are provided in Appendix B.

4 Exceedance of multiple largest observations and time intervals

In the previous section, we examined the exceedance of the largest observed value, $\mathcal{R}$, in the context of right-censored data. In this section, we extend this analysis to the exceedance of the second, third, fourth, ..., up to the jth largest observations, as long as they exceed the largest censored observation, X_{c_v}. We then consider the time t between any two of these largest observations and calculate the lower and upper probabilities for the exceedance of time t.

We maintain the notation introduced in the previous section, with a few additions. To simplify notation, we redefine $\mathcal{R}$ (previously $\mathcal{R} = x_u$) as $\mathcal{R}_1 = x_u$, representing the first largest event time in the dataset. Similarly, let $\mathcal{R}_2 = x_{u-1}$ denote the second largest event time, $\mathcal{R}_3 = x_{u-2}$ the third largest, and so on, up to the largest observed event time greater than the censored observation c_v. Thus, we have the ordering $\mathcal{R}_1 >$

$\mathcal{R}_2 > \mathcal{R}_3 > \ldots > \mathcal{R}_j$, corresponding to $x_u > x_{u-1} > x_{u-2} > \ldots > x_{u-i}$, where $x_{u-i} > c_v$ for $i = 0, 1, \ldots, u$ and $j = 1, 2, \ldots, u$. Recall that $\tilde{n}_{c_r}$, for $r = 1, 2, \ldots, v$, represents the number of observations in the risk set just before time c_r. We assume no ties occur among the observations, and the method is based on the shifted $\tilde{A}_{(n)}$ in Equation (5), under the exchangeability and non-informative right censoring assumptions [10, 11].

As in the previous section, we can directly compute the probability for the event that at least one of the right-censored individuals has a lifetime greater than any of the largest observed values, provided that it exceeds the largest censored observation at c_v. That is, for $\mathcal{R}_j = x_{u-j+1}$, where $x_{u-j+1} > c_v$ for $j = 1, \ldots, u$, the probability is given by

$$P(G_{\mathcal{R}_j}(0)) = 1 - \prod_{r=1}^{v} \frac{\tilde{n}_{c_r} - j + 1}{\tilde{n}_{c_r} + 1}. \tag{8}$$

Next, we extend the analysis by considering the addition of future individuals to the study, as we did in Section 3. We focus on the event that at least one of the right-censored individuals, or one of the $m \geq 1$ future individuals, has a lifetime greater than the jth largest observed value, $\mathcal{R}_j = x_{u-j+1}$, where $x_{u-j+1} > c_v$. Let $G_{\mathcal{R}_j}(m)$ denote this event.

The probability that a lifetime exceeds any of the largest observed values, when calculated backwards from the largest value to the jth largest (as long as it exceeds the largest censored observation), is given by

$$P(G_{\mathcal{R}_j}(m)) = 1 - \left[\prod_{i=1}^{m} \frac{n + i - j}{n + i} \prod_{r=1}^{v} \frac{\tilde{n}_{c_r} - j + 1}{\tilde{n}_{c_r} + 1} \right], \tag{9}$$

or equivalently,

$$P(G_{\mathcal{R}_j}(m)) = 1 - \left[\frac{n(n-1)\ldots(n-j+1)}{(n+m)(n+m-1)\ldots(n+m-j+1)} \prod_{r=1}^{v} \frac{\tilde{n}_{c_r} - j + 1}{\tilde{n}_{c_r} + 1} \right]. \tag{10}$$

The proofs of these results are provided in Appendix C.

We now consider the event that at least one of the right-censored individuals, or one of the $m \geq 1$ future individuals, has a lifetime greater than a specific time t, where t lies between two consecutive largest observed values, say x_i and x_{i+1}, for $i = 0, 1, \ldots, u$, as long as $x_i > c_v$. For simplicity, let $G_{t\in(x_i,x_{i+1})}(m)$ denote this event.

The lower probability for the event $G_{t\in(x_i,x_{i+1})}(m)$ is the probability of the event $G_{x_{i+1}}(m)$, i.e.,

$$\underline{P}(G_{t\in(x_i,x_{i+1})}(m)) = P(G_{x_{i+1}}(m)),$$

while the upper probability is the probability of the event $G_{x_i}(m)$, i.e.,

$$\overline{P}(G_{t\in(x_i,x_{i+1})}(m)) = P(G_{x_i}(m)).$$

For example, if $t \in (x_{u-1}, x_u)$, where $\mathcal{R}_1 = x_u$ and $\mathcal{R}_2 = x_{u-1}$ represent the first and second largest event times in the dataset, with $x_{u-1} > c_v$, then the lower and

upper survival of t, respectively, are

$$\underline{P}(G_{t\in(\mathcal{R}_2,\mathcal{R}_1)}(m)) = P(G_{\mathcal{R}_1}(m)),$$
$$\overline{P}(G_{t\in(\mathcal{R}_2,\mathcal{R}_1)}(m)) = P(G_{\mathcal{R}_2}(m)),$$

which is given by Equation (9). which can be computed using Equation (9)

In the next section, we apply the methods introduced in this paper to the full supercentenarian dataset, separately for women and men.

5 Application to the Supercentenarian data

This paper analyses a dataset used by Alves et al. [2], which includes the ages at death of 1,740 individuals who lived past 110 years old, along with the ages of those who were still alive when the data were collected. The dataset was compiled by the Gerontology Research Group (GRG) and collected on April 22, 2018, from Tables B and C of their records.[1] For analysis, ages are presented in days, though here we will also use years. It is assumed that there are no ties in age between individuals, and for simplicity, we treat each year as 365 days, ignoring leap years.

Notably, the dataset highlights the extreme lifespans of supercentenarians, with Jeanne Calment from France holding the record for the oldest verified age at 122.5 years, and Jiroemon Kimura from Japan holding the record for men at 116.2 years. The dataset includes 1,580 lifetimes of supercentenarian women and 160 of supercentenarian men, with women generally living longer. Of the 1,580 supercentenarian women, 72 were still alive on April 22, 2018, and are thus considered right-censored. In contrast, only two supercentenarian men out of 160 were alive at the time of data collection.

This study aims to estimate the probability that at least one of the right-censored supercentenarian women will live beyond Jeanne Calment's age, and similarly, the probability that at least one of the right-censored supercentenarian men will exceed Jiroemon Kimura's age. The methods outlined in Sections 3 and 4 will be applied separately to the supercentenarian men and women. The first two examples will illustrate the methods from Section 3, while the latter two will demonstrate the methods from Section 4.

Example 1. *(Supercentenarian Women Data)* In this example, we analyse the supercentenarian data for women, which includes $n = 1580$ supercentenarian women, of whom 72 were still alive at the time of the study and their lifetimes are thus right-censored. Jeanne Calment's age of 122.5 years was the largest age recorded in the dataset, so we set $\mathcal{R} = 122.5$. The objective is to determine the probability, $P(G_{\mathcal{R}}(0))$, that at least one of the 72 right-censored supercentenarian women will have a lifetime

[1] The dataset is available at *http://www.grg.org/Adams/Tables.htm*, and further details can be found in [2].

exceeding $\mathcal{R} = 122.5$ years. This probability is given by the following formula:

$$P(G_{122.5}(0)) = 1 - \prod_{r=1}^{72} \frac{\tilde{n}_{c_r}}{\tilde{n}_{c_r} + 1} = 1 - 0.6567 = 0.3433$$

Thus, based on our model with non-informative right censoring, there is a 34.33% chance that at least one of the 72 right-censored supercentenarian women will live longer than Jeanne Calment's age.

Next, consider adding $m = 1$ future supercentenarian woman, denoted X_{n+1}, to the study. Conditional on the assumption that all 72 right-censored supercentenarian women have failed before reaching $\mathcal{R} = 122.5$, the probability, $P(G_{122.5}(1))$, that at least one of the 72 right-censored women or the new supercentenarian woman X_{n+1} will exceed $\mathcal{R}$ is given by:

$$P(G_{122.5}(1)) = 1 - \left[\frac{1580}{1580 + 1} \prod_{r=1}^{72} \frac{\tilde{n}_{c_r}}{\tilde{n}_{c_r} + 1} \right] = 1 - 0.6563 = 0.3437$$

Now, suppose $m = 2$ future supercentenarian women, X_{n+1} and X_{n+2}, are added to the study. Conditional on the assumption that all 72 right-censored women and the first future supercentenarian woman X_{n+1} have failed before reaching $\mathcal{R} = 122.5$, the probability, $P(G_{122.5}(2))$, that at least one of the 72 right-censored women or any of the two new supercentenarian women will exceed $\mathcal{R}$ is:

$$P(G_{122.5}(2)) = 1 - \left[\frac{1580}{1580 + 2} \prod_{r=1}^{72} \frac{\tilde{n}_{c_r}}{\tilde{n}_{c_r} + 1} \right] = 1 - 0.6559 = 0.3441$$

For larger values of $m \geq 2$, the probability $P(G_{122.5}(m))$ that at least one of the 72 right-censored supercentenarian women or any of the m future supercentenarian women will live longer than $\mathcal{R} = 122.5$ can be calculated as:

$$P(G_{122.5}(m)) = 1 - \left[\frac{1580}{1580 + m} \prod_{r=1}^{72} \frac{\tilde{n}_{c_r}}{\tilde{n}_{c_r} + 1} \right] = 1 - \left[\frac{1580}{1580 + m} \times 0.6567 \right]$$

As $m \to \infty$, the probability $P(G_{122.5}(m))$ approaches 1, indicating that with enough future supercentenarians, the event becomes almost certain, as illustrated in Figure 1.

One interesting aspect to consider is determining the smallest m such that the value of the probability $P(G_{\mathcal{R}}(m))$ exceeds a specified probability P, where $P \in [0, 1]$. For example, from Figure 1, we find that:

$$P(G_{122.5}(m)) = 1 - \left[\frac{1580}{1580 + m} \times 0.6567 \right] > P$$

where $P(G_{122.5}(m))$ exceeds $P = 0.95$ when $m \geq 19200$ future supercentenarian women are considered.

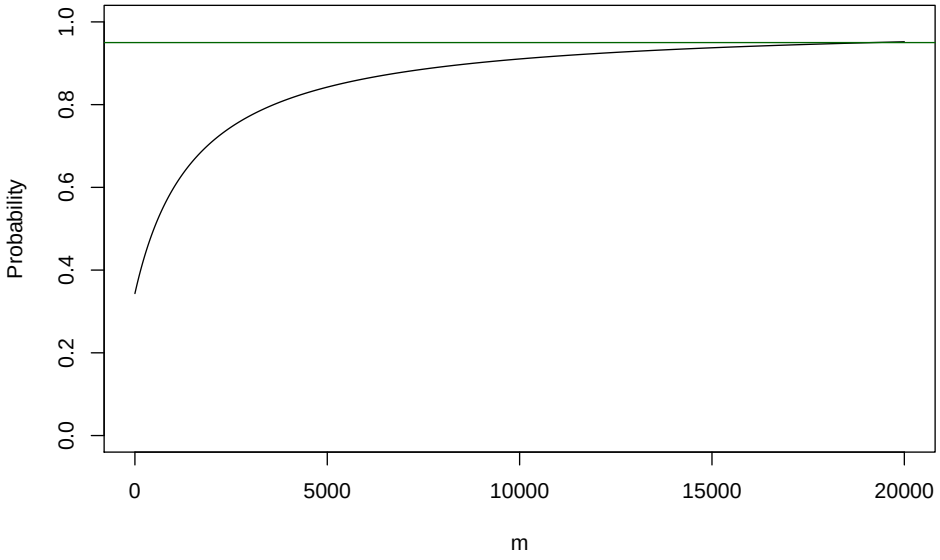

Fig. 1 The probability $P(G_{122.5}(m))$ for the supercentenarian women dataset, as in Example 1. This figure illustrates the likelihood that at least one of the 72 right-censored women or any of the m future supercentenarian women will exceed the age of 122.5 years. In particular, for $m = 2$, the probability is 0.3441, and as m increases, the probability approaches 1, indicating that with a larger number of future supercentenarians, the event becomes almost certain.

Example 2. *(Supercentenarian men data)* In this example, we examine the supercentenarian data for men. The dataset consists of 160 supercentenarian men, two of whom were still alive at the time of the study, and therefore, their lifetimes are right-censored. Since Jiroemon Kimura's age of 116.2 years was the largest recorded age in the dataset, we set $\mathcal{R} = 116.2$. The focus is on determining the probability of the event $G_{116.2}(0)$, which is the probability that at least one of the two right-censored supercentenarian men has a lifetime exceeding the largest observed age, $\mathcal{R} = 116.2$. This probability is given by Equation (6) as follows:

$$P(G_{116.2}(0)) = 1 - \prod_{r=1}^{2} \frac{\tilde{n}_{c_r}}{\tilde{n}_{c_r} + 1} = 1 - 0.9444 = 0.0556$$

Thus, there is a 5.56% chance that at least one of the two right-censored supercentenarian men would live beyond Jiroemon Kimura's age of 116.2 years.

Next, consider $m = 1$ future supercentenarian man, X_{n+1}, added to the study, in addition to the $n = 160$ supercentenarian men. The lifetime of X_{n+1} is considered, conditional on the assumption that both right-censored supercentenarian men have

already failed before the age $\mathcal{R} = 116.2$. The probability of the event $G_{116.2}(1)$, which is the probability that at least one of the two right-censored men or the lifetime of X_{n+1} exceeds $\mathcal{R} = 116.2$, is calculated using Equation (7) as:

$$P(G_{116.2}(1)) = 1 - \left[\frac{160}{160+1}\prod_{r=1}^{2}\frac{\tilde{n}_{c_r}}{\tilde{n}_{c_r}+1}\right] = 1 - 0.9385 = 0.0615$$

Now, consider $m = 2$ future supercentenarian men, X_{n+2}, added to the study, in addition to the $n = 160$ supercentenarian men and the first future supercentenarian, X_{n+1}. The lifetime of X_{n+2} is considered, conditional on the assumption that both right-censored supercentenarian men and X_{n+1} have already failed before the age $\mathcal{R} = 116.2$. The probability of the event $G_{116.2}(2)$, that at least one of the two right-censored men or one of the lifetimes of X_{n+1} and X_{n+2} exceeds $\mathcal{R} = 116.2$, is given by:

$$P(G_{116.2}(2)) = 1 - \left[\frac{160}{160+2}\prod_{r=1}^{2}\frac{\tilde{n}_{c_r}}{\tilde{n}_{c_r}+1}\right] = 1 - 0.9327 = 0.0673$$

Considering $m \geq 2$ future supercentenarian men added to the study, the probability of the event $G_{116.2}(m)$, that at least one of the two right-censored supercentenarian men or one of the lifetimes of the $m \geq 2$ future supercentenarian men exceeds $\mathcal{R} = 116.2$, is calculated using Equation (7). The results are displayed in Figure 2 as:

$$P(G_{116.2}(m)) = 1 - \left[\frac{160}{160+m}\prod_{r=1}^{2}\frac{\tilde{n}_{c_r}}{\tilde{n}_{c_r}+1}\right] = 1 - \left[\frac{160}{160+m} \times 0.9444\right]$$

An interesting point to consider is identifying the smallest value of m for which the probability $P(G_{\mathcal{R}}(m))$ exceeds a given threshold P, where $P \in [0, 1]$. For instance, as shown in Figure 2, we observe that:

$$P(G_{116.2}(m)) = 1 - \left[\frac{160}{160+m} \times 0.9444\right] > P$$

where $P(G_{116.2}(m))$ exceeds $P = 0.95$ when $m \geq 2900$ future supercentenarian men are added to the study.

Example 3. *(Supercentenarian women data)* In this example, we use data from $n = 1580$ supercentenarian women (as in Example 1). Among them, 72 women were still alive at the time of the study, and their lifetimes are right-censored. Additionally, there are eight supercentenarian women whose ages exceed the largest censored value, with the oldest recorded age being 117.1. In this case, we consider the second-largest age, $\mathcal{R}_2 = 119.3$, and the third-largest age, $\mathcal{R}_3 = 117.8$, in contrast to the first largest age of $\mathcal{R}_1 = 122.5$, as considered in Example 1.

We are interested in determining the probability of the event $G_{\mathcal{R}_2}(0)$, which refers to the probability that at least one of the 72 right-censored supercentenarian women has a lifetime exceeding the second-largest observed age, $\mathcal{R}_2 = 119.3$. This probability

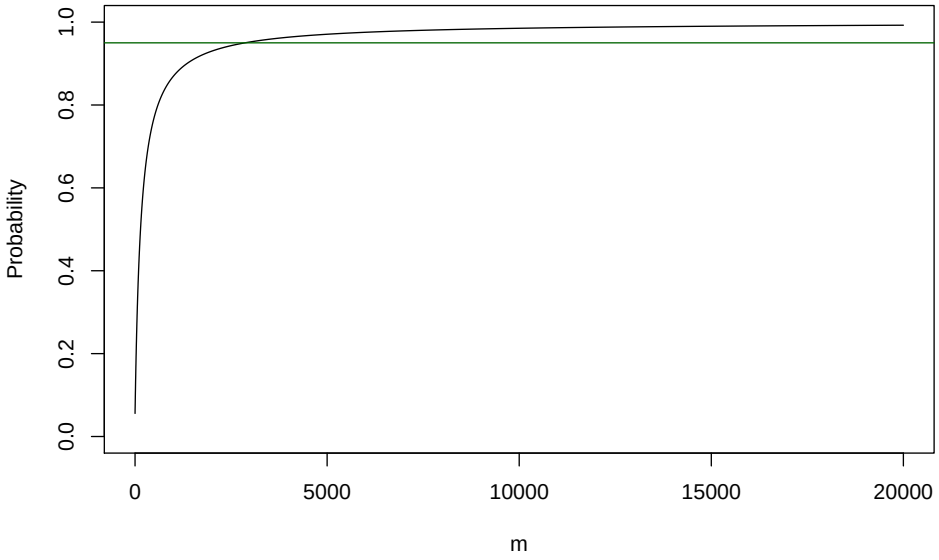

Fig. 2 The probability $P(G_{116.2}(m))$ for the supercentenarian men dataset, as in Example 2. This figure shows the likelihood that at least one of the two right-censored supercentenarian men or any of the m future supercentenarian men will exceed the age of $\mathcal{R} = 116.2$ years. For $m = 0$, the probability is 0.0556, indicating a 5.56% chance that one of the right-censored men would exceed this age. As more future supercentenarians are added to the study, the probability increases, with the probability approaching 1 as m becomes larger. Notably, the probability exceeds 0.95 when at least 2900 future supercentenarian men are included in the study.

is computed as follows:

$$P(G_{119.3}(0)) = 1 - \prod_{r=1}^{72} \frac{\tilde{n}_{c_r} - 1}{\tilde{n}_{c_r} + 1} = 1 - 0.4228 = 0.5772.$$

Next, we compute the probability for the event $G_{\mathcal{R}_3}(0)$, which is the probability that at least one of the 72 right-censored supercentenarian women has a lifetime exceeding the third-largest observed age, $\mathcal{R}_3 = 117.8$. This probability is:

$$P(G_{117.8}(0)) = 1 - \prod_{r=1}^{72} \frac{\tilde{n}_{c_r} - 2}{\tilde{n}_{c_r} + 1} = 1 - 0.2655 = 0.7345.$$

Thus, based on our model assumptions, which involve the $A_{(n)}$ assumption and non-informative right censoring, we find that the probability of at least one of the 72 supercentenarian women surviving beyond $\mathcal{R}_2 = 119.3$ is 0.5772. This probability

increases to 0.7345 for surviving beyond $\mathcal{R}_3 = 117.8$. Additionally, it is more likely that someone will survive any of the eight supercentenarian women, as the analysis proceeds from the first largest age to the eighth largest age, all of which exceed the largest censored age of 117.1.

Now, let us consider the scenario where $m = 2$ future supercentenarian women, X_{n+1} and X_{n+2}, are added to the study, alongside the existing $n = 1580$ supercentenarian women. The lifetime of X_{n+2} is considered, conditional on the assumption that all 72 right-censored supercentenarian women and X_{n+1} have failed before reaching $\mathcal{R}_2 = 119.3$. The probability of the event $G_{\mathcal{R}_2}(2)$, which is the probability that at least one of the 72 right-censored women or one of the future women (X_{n+1} or X_{n+2}) survives beyond $\mathcal{R}_2 = 119.3$, is computed as:

$$P(G_{119.3}(2)) = 1 - \left[\frac{1580 \cdot 1579}{(1580+2)(1580+1)} \prod_{r=1}^{72} \frac{\tilde{n}_{c_r} - 1}{\tilde{n}_{c_r} + 1}\right] = 1 - 0.4217 = 0.5783.$$

Similarly, for the survival beyond the third largest observed age $\mathcal{R}_3 = 117.8$, we compute the probability for the event $G_{\mathcal{R}_3}(2)$, which is the probability that at least one of the 72 right-censored supercentenarian women or X_{n+1} and X_{n+2} survives beyond $\mathcal{R}_3 = 117.8$:

$$P(G_{117.8}(2)) = 1 - \left[\frac{1580 \cdot 1579 \cdot 1578}{(1580+2)(1580+1)(1580)} \prod_{r=1}^{72} \frac{\tilde{n}_{c_r} - 2}{\tilde{n}_{c_r} + 1}\right] = 1 - 0.2645 = 0.7355.$$

Thus, for survival times between $\mathcal{R}_2 = 119.3$ and $\mathcal{R}_3 = 117.8$, the upper survival probability is 0.7355 (corresponding to $\mathcal{R}_3 = 117.8$) and the lower survival probability is 0.5783 (corresponding to $\mathcal{R}_2 = 119.3$).

Considering $m \geq 2$ future supercentenarian women added to the study, the probabilities for the events $G_{\mathcal{R}_1}(m)$, $G_{\mathcal{R}_2}(m)$, and $G_{\mathcal{R}_3}(m)$—that at least one of the 72 right-censored supercentenarian women or one of the $m \geq 2$ future women survives beyond $\mathcal{R}_1 = 122.5$, $\mathcal{R}_2 = 119.3$, or $\mathcal{R}_3 = 117.8$—are shown in Figure 3.

From Figure 3, if we consider a specific probability value, say $P = 0.95$, we can determine the smallest m for which the probabilities $P(G_{122.5}(m))$, $P(G_{119.3}(m))$, and $P(G_{117.8}(m))$ exceed $P = 0.95$. It is evident from the figure that as the largest recorded age decreases (moving backward through the ordered ages), the smallest m that results in a probability greater than $P = 0.95$ also decreases. Specifically, for the event $G_{122.5}(m)$, the smallest m such that $P(G_{122.5}(m)) > 0.95$ is $m \geq 19200$ future supercentenarian women. For the event $G_{119.3}(m)$, the smallest m for which $P(G_{119.3}(m)) > 0.95$ is $m \geq 3050$ future supercentenarian women. Finally, for the event $G_{117.8}(m)$, the smallest m that makes $P(G_{117.8}(m)) > 0.95$ is $m \geq 1180$ future supercentenarian women.

Example 4. *(Supercentenarian Men Data)* In this example, we again use the data on $n = 160$ supercentenarian men, as in Example 2. Two of these men are still alive at the time of the study, so their lifetimes are right-censored. Additionally, there are 33 supercentenarian men whose ages exceed the largest censored supercentenarian age,

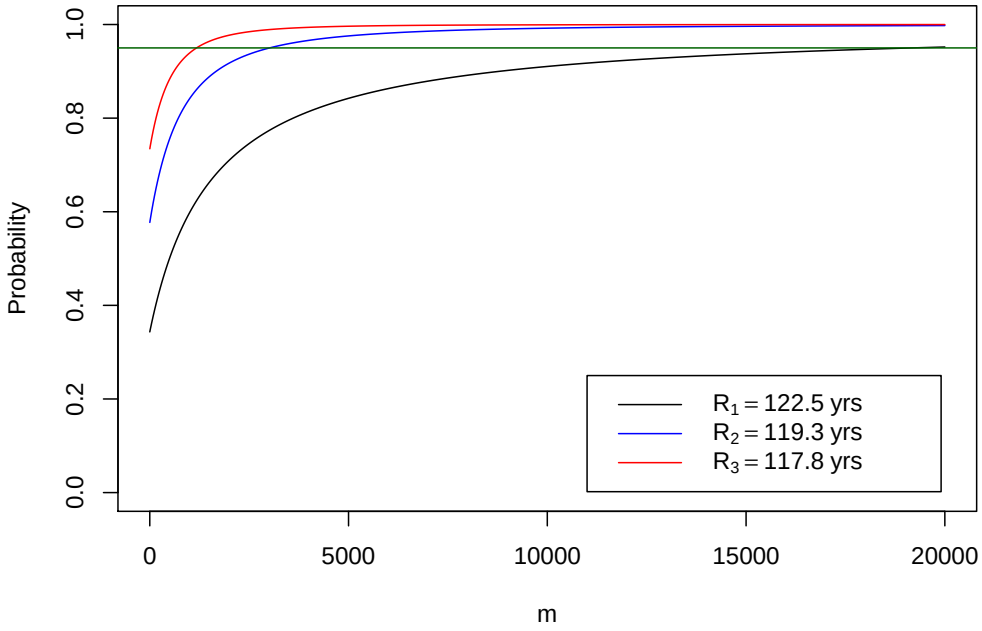

Fig. 3 The probabilities $P(G_{\mathcal{R}_1}(m))$, $P(G_{\mathcal{R}_2}(m))$, and $P(G_{\mathcal{R}_3}(m))$ for the supercentenarian women dataset, as in Example 3. These probabilities represent the likelihood that at least one of the 72 right-censored supercentenarian women, or any of the m future supercentenarian women, will survive beyond the specified age thresholds: $\mathcal{R}_1 = 122.5$, $\mathcal{R}_2 = 119.3$, and $\mathcal{R}_3 = 117.8$. For each case, the probability increases with the addition of future supercentenarians. As illustrated, the smallest number of future supercentenarian women required for the probability to exceed 0.95 is $m \geq 19200$ for $\mathcal{R}_1$, $m \geq 3050$ for $\mathcal{R}_2$, and $m \geq 1180$ for $\mathcal{R}_3$. This demonstrates that as the reference age decreases, the number of future supercentenarians needed to achieve a high probability decreases as well.

which is 111.9. In Example 2, we considered the first largest recorded age, $\mathcal{R}_1 = 116.2$. In this case, we focus on the second and third largest ages recorded, $\mathcal{R}_2 = 115.7$ and $\mathcal{R}_3 = 115.5$, respectively.

The interest is in calculating the probability for the event $G_{\mathcal{R}_2}(0)$, which represents the probability that at least one of the two right-censored supercentenarian men has a lifetime greater than the second largest observed value, $\mathcal{R}_2 = 115.7$. This probability is obtained as follows:

$$P(G_{115.7}(0)) = 1 - \prod_{r=1}^{2} \frac{\tilde{n}_{c_r} - 1}{\tilde{n}_{c_r} + 1} = 1 - 0.8903 = 0.1097.$$

Similarly, the probability for the event $G_{\mathcal{R}_3}(0)$, where at least one of the two right-censored supercentenarian men has a lifetime greater than the third largest recorded age, $\mathcal{R}_3 = 115.5$, is

$$P(G_{115.5}(0)) = 1 - \prod_{r=1}^{2} \frac{\tilde{n}_{c_r} - 2}{\tilde{n}_{c_r} + 1} = 1 - 0.8378 = 0.1622.$$

Under the assumptions of the model, which is based on the $A_{(n)}$ assumption and non-informative right censoring, the probability that at least one of the right-censored supercentenarian men will live longer than the second largest observed age, $\mathcal{R}_2 = 115.7$, is 0.1097. This probability increases to 0.1622 if considering survival beyond the third largest age, $\mathcal{R}_3 = 115.5$. Additionally, it is more likely that one of the 33 supercentenarian men, whose ages exceed 111.9, will survive any of the ages from $\mathcal{R}_1$ to $\mathcal{R}_3$.

Next, consider the addition of $m = 2$ future supercentenarian men, denoted X_{n+1} and X_{n+2}, to the study. The lifetime of X_{n+2} is considered, conditional on the fact that the two right-censored supercentenarian men and X_{n+1} have all failed before reaching the second largest recorded age, $\mathcal{R}_2 = 115.7$. The probability for the event $G_{\mathcal{R}_2}(2)$, that at least one of the right-censored supercentenarian men or the future supercentenarian men X_{n+1} and X_{n+2} will live longer than $\mathcal{R}_2$, is

$$P(G_{115.7}(2)) = 1 - \left[\frac{160 \cdot 159}{(160+2)(160+1)} \prod_{r=1}^{2} \frac{\tilde{n}_{c_r} - 1}{\tilde{n}_{c_r} + 1} \right] = 1 - 0.8684 = 0.1316.$$

Similarly, considering the survival of the third largest age, $\mathcal{R}_3 = 115.5$, and the addition of $m = 2$ future supercentenarian men, the probability for the event $G_{\mathcal{R}_3}(2)$ is

$$P(G_{115.5}(2)) = 1 - \left[\frac{160 \cdot 159 \cdot 158}{(160+2)(160+1)(160)} \prod_{r=1}^{2} \frac{\tilde{n}_{c_r} - 2}{\tilde{n}_{c_r} + 1} \right] = 1 - 0.8070 = 0.1930.$$

Now, consider the event $G_{t\in(\mathcal{R}_3,\mathcal{R}_2)}(2)$, where t lies between $\mathcal{R}_3 = 115.5$ and $\mathcal{R}_2 = 115.7$, in the case of adding $m = 2$ future supercentenarian men. The lower probability for the event $G_{t\in(115.5,115.7)}(2)$ is derived from the probability for the event $G_{115.7}(2)$, as shown in Equation (4), yielding:

$$\underline{P}(G_{t\in(115.5,115.7)}(2)) = P(G_{115.7}(2)) = 0.1316.$$

The corresponding upper probability for the event $G_{t\in(115.5,115.7)}(2)$ is derived from the probability for the event $G_{115.5}(2)$, as shown in Equation (4), yielding:

$$\overline{P}(G_{t\in(115.5,115.7)}(2)) = P(G_{115.5}(2)) = 0.1930.$$

Finally, consider the case where $m \geq 2$ future supercentenarian men are added to the study. The probabilities for the events $G_{\mathcal{R}_1}(m)$, $G_{\mathcal{R}_2}(m)$, and $G_{\mathcal{R}_3}(m)$, which

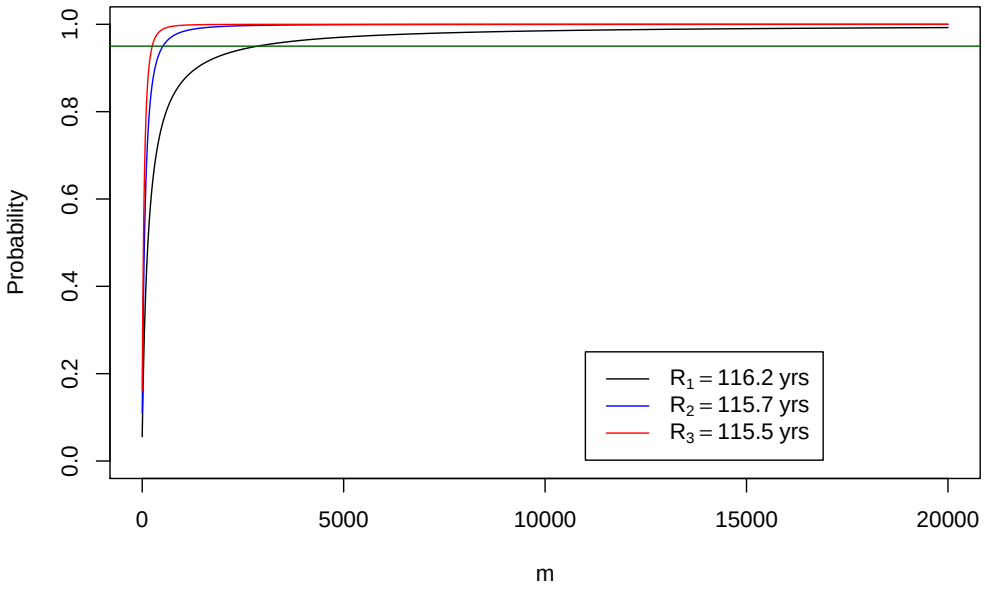

Fig. 4 The probabilities $P(G_{\mathcal{R}_1}(m))$, $P(G_{\mathcal{R}_2}(m))$, and $P(G_{\mathcal{R}_3}(m))$ for the supercentenarian men dataset, as in Example 4. These probabilities represent the likelihood that at least one of the two right-censored supercentenarian men, or any of the m future supercentenarian men, will survive beyond the specified age thresholds: $\mathcal{R}_1 = 116.2$, $\mathcal{R}_2 = 115.7$, and $\mathcal{R}_3 = 115.5$. The probability increases with the addition of future supercentenarians. From the figure, it is evident that as the largest recorded age decreases, the smallest number of future supercentenarians required to exceed a probability of 0.95 decreases. Specifically, for the event $G_{116.2}(m)$, the smallest m such that $P(G_{116.2}(m)) > 0.95$ is $m \geq 2900$; for $G_{115.7}(m)$, it is $m \geq 515$; and for $G_{115.5}(m)$, it is $m \geq 250$.

represent the probabilities that at least one of the two right-censored supercentenarian men or one of the $m \geq 2$ future supercentenarian men will live longer than $\mathcal{R}_1 = 116.2$, $\mathcal{R}_2 = 115.7$, and $\mathcal{R}_3 = 115.5$, respectively, are shown in Figure 4.

From Figure 4, if we consider a specific probability value, say $P = 0.95$, we can determine the smallest m for which the probabilities $P(G_{116.2}(m))$, $P(G_{115.7}(m))$, and $P(G_{115.5}(m))$ exceed $P = 0.95$. It is evident from the figure that as the largest recorded age decreases (moving backward through the ordered ages), the smallest m that results in a probability greater than $P = 0.95$ also decreases. Specifically, we see that $P(G_{116.2}(m))$ exceeds 0.95 when $m \geq 2900$ future supercentenarian men, $P(G_{115.7}(m))$ exceeds 0.95 when $m \geq 515$ future supercentenarian men, and $P(G_{115.5}(m))$ exceeds 0.95 when $m \geq 250$ future supercentenarian men.

6 Concluding remarks

In this paper, we introduced a method to estimate the probability that the true lifetime corresponding to a right-censored observation exceeds the largest observed value in a dataset. This method also extends to consider future observations, calculating the probability that at least one future or right-censored observation has a lifetime exceeding the largest observed value. Furthermore, we extended the analysis to the exceedance of the second, third, and up to the j-th largest observations, provided they exceed the largest censored observation. Additionally, we examined the time between any two of these largest observations, calculating the lower and upper probabilities for the exceedance of the time between them.

The method is built upon the Nonparametric Predictive Inference (NPI) framework, particularly utilising the shifted $A_{(n)}$ assumption. This assumption, combined with the exchangeability assumption and non-informative right censoring, offers a flexible and assumption-minimal approach for deriving predictive probabilities. These assumptions focus on the remaining times to the event of interest for individuals reaching a certain age and allow for the quantification of uncertainty in future observations. The use of NPI is especially suited for extreme value analysis because it does not rely on parametric assumptions, making it well-suited for data with extreme values, such as the Supercentenarian dataset, where the true tail behaviour of the distribution is of particular interest.

We applied these methods to the Supercentenarian dataset, with separate analyses for women and men. The results show that assuming the largest observed value as the endpoint of support is not appropriate in the context of extreme value analysis. For instance, the probabilities of surviving beyond the largest observed age were notably high, demonstrating the importance of considering exceedance probabilities rather than treating the largest observed value as a definitive endpoint.

While the NPI method, with its minimal assumptions, provides valuable insights, it does have limitations. It is not sufficient to make detailed predictions beyond the largest observed value without incorporating additional distributional assumptions or accounting for other complexities in the data. Future research could consider extending this methodology by integrating additional assumptions about the underlying distribution or by using alternative approaches that allow for covariates and random effects to refine survival probability estimates. Such extensions could enhance the predictive power of the analysis, especially when applied to larger and more diverse datasets that consider other factors, such as health status, socio-economic factors, and geography.

This work provides a foundation for understanding the uncertainty associated with extreme survival outcomes and has practical applications in fields such as healthcare, demography, and insurance. The exceedance probabilities we have derived could inform resource allocation for long-term care, inform population models of extreme longevity, and guide insurance risk assessments for longevity-related products. Further exploration of these implications will help shape policies that consider the growing number of individuals living to extreme ages.

It is also of interest to develop the NPI approach to include further information through modelling the dependence of lifetimes on covariates. This will provide an

alternative to inferences such as the Proportional Hazards model and will be enabled by the development of NPI for regression problems, which is currently ongoing.

In conclusion, while NPI offers a robust framework for extreme value analysis in the context of right-censored data, integrating additional assumptions and expanding the data set will be key to refining survival predictions and enhancing the practical relevance of this approach. NPI for right-censored data is closely related to the Kaplan-Meier (KM) estimate for the population survival function for such data [14]. The NPI-based lower and upper survival functions bound the KM estimate, but they have strong consistency properties for prediction, which do not hold for the KM estimate. Furthermore, NPI provides predictive inference which is exactly calibrated [15], a strong consistency property in frequentist statistics.

Acknowledgements. The research described in this article was conducted during Ali Mahnashi's PhD studies at the Department of Mathematical Sciences, Durham University. Ali wishes to express his sincere gratitude to Jazan University in Saudi Arabia for providing financial support that enabled him to complete his PhD studies.

Appendix A Proof of Equation (6) with an illustrative example

Proof of Equation (6): We first consider the individual X_{c_v}, who is the last individual to be censored at censoring time c_v, such that there are no further censorings beyond it. For X_{c_v}, we can apply the shifted-$\tilde{A}_{(n)}$, as given in Equation (5), which allows us to apply $A_{(n)}$ with the starting point shifted from 0 to the highest right-censoring time c_v. The lifetime of this individual X_{c_v} will either exceed $\mathcal{R}$ or not. If the lifetime of X_{c_v} exceeds $\mathcal{R}$, then based on the shifted-$\tilde{A}_{(n)}$, the probability that $X_{c_v} > \mathcal{R}$ is

$$P(X_{c_v} > \mathcal{R}) = \frac{1}{\tilde{n}_{c_v} + 1}$$

If the lifetime of X_{c_v} does not exceed $\mathcal{R}$, then the probability for the event $X_{c_v} < \mathcal{R}$, knowing the value of $\tilde{n}_{c_v}$, is

$$P(X_{c_v} < \mathcal{R}) = 1 - \frac{1}{\tilde{n}_{c_v} + 1} = \frac{\tilde{n}_{c_v}}{\tilde{n}_{c_v} + 1}$$

where $\tilde{n}_{c_v}$ is the number of observations in the risk set just prior to time c_v.

Next, we consider the previous individual with the second censoring time c_{v-1}, namely $X_{c_{v-1}}$, conditional on $X_{c_v} < \mathcal{R}$. It is important to note that for $X_{c_{v-1}}$, it does not matter where exactly the final individual's failure time or lifetime, X_{c_v}, is, as long as it occurs before $\mathcal{R}$. Specifically, we do not need to take censoring into account for X_{c_v} because we are conditioning on what happens before $\mathcal{R}$. Thus, it does not matter what the exact value of X_{c_v} is within the interval $(X_{c_{v-1}}, \mathcal{R})$. Therefore, the

probability that $X_{c_{v-1}}$ exceeds $\mathcal{R}$, given that $X_{c_v} < \mathcal{R}$, based on the shifted-$\tilde{A}_{(n)}$, is

$$P(X_{c_{v-1}} > \mathcal{R} \mid X_{c_v} < \mathcal{R}) = \frac{1}{\tilde{n}_{c_{v-1}} + 1}$$

and the probability for the event of interest, $X_{c_{v-1}} < \mathcal{R}$, given $X_{c_v} < \mathcal{R}$, knowing the value of $\tilde{n}_{c_{v-1}}$, is

$$P(X_{c_{v-1}} < \mathcal{R} \mid X_{c_v} < \mathcal{R}) = 1 - \frac{1}{\tilde{n}_{c_{v-1}} + 1} = \frac{\tilde{n}_{c_{v-1}}}{\tilde{n}_{c_{v-1}} + 1}$$

where $\tilde{n}_{c_{v-1}}$ is the number of observations in the risk set just prior to time c_{v-1}.

The same procedures are repeated for all other individuals whose lifetimes have been right-censored at censoring times c_r, where $r = 1, 2, \ldots, v-3, v-2$. If the lifetime of an individual X_{c_r} does not exceed $\mathcal{R}$, we check the previous individuals at those censoring times c_r. The important thing to note is that for these individuals, it does not matter exactly where their failure times occur as long as they have already failed before $\mathcal{R}$. Generally, for the lifetime of these later individuals, censoring does not need to be taken into account, since it is based on what happens before $\mathcal{R}$. Therefore, for an individual X_{c_r} at time c_r, we only know the number of individuals between X_{c_r} and $\mathcal{R}$, and we also know that all of them failed before $\mathcal{R}$. The probability that $X_{c_r} > \mathcal{R}$, given that $X_{c_{r+1}} < \mathcal{R}, \ldots, X_{c_{v-1}} < \mathcal{R}, X_{c_v} < \mathcal{R}$, based on the shifted-$\tilde{A}_{(n)}$, as in Equation (5), is

$$P(X_{c_r} > \mathcal{R} \mid X_{c_{r+1}} < \mathcal{R}, \ldots, X_{c_{v-1}} < \mathcal{R}, X_{c_v} < \mathcal{R}) = \frac{1}{\tilde{n}_{c_r} + 1}$$

and the probabilities for the event of interest, that no one survives beyond $\mathcal{R}$, knowing the values of $\tilde{n}_{c_r}$ for $r = 1, 2, \ldots, v-3, v-2$, are

$$P(X_{c_r} < \mathcal{R} \mid X_{c_{r+1}} < \mathcal{R}, \ldots, X_{c_{v-1}} < \mathcal{R}, X_{c_v} < \mathcal{R}) = 1 - \frac{1}{\tilde{n}_{c_r} + 1} = \frac{\tilde{n}_{c_r}}{\tilde{n}_{c_r} + 1} \quad \text{(A1)}$$

It is crucial to emphasize that for the event of interest above, we do not need to apply $A_{(n)}$ with censoring, since it is written as a conditional event that all individuals have lifetimes less than $\mathcal{R}$. If an individual's lifetime exceeds $\mathcal{R}$, then we know that the event of all individuals being less than $\mathcal{R}$ is not true.

Consequently, the probability for the event of interest $G_{\mathcal{R}}(0)$, denoted by $P(G_{\mathcal{R}}(0))$, is

$$P(G_{\mathcal{R}}(0)) = 1 - \prod_{r=1}^{v} \frac{\tilde{n}_{c_r}}{\tilde{n}_{c_r} + 1}$$

Thus, the proof is complete.

The following example illustrates the probabilities presented in this section.

Example 5. Suppose we have a dataset consisting of $n = 10$ observations. Of these ten individuals, seven died at ages 111, 113, 115, 116, 119, 120, and 122, while three observations were still alive at the time the data were collected. Their lifetimes were right-censored at ages 112, 114, and 117. Note that the largest recorded observation is 122, so $\mathcal{R} = 122$. Let X_{c_1}, X_{c_2}, and X_{c_3} denote the random quantities corresponding to the right-censoring times at 112, 114, and 117, respectively.

We first consider the individual X_{c_3}, who was censored last at age 117, such that there are no further censorings beyond this point. The lifetime of X_{c_3} will either survive beyond $\mathcal{R}$ or not. If $X_{c_3} > 122$, then based on the shifted-$\tilde{A}_{(3)}$ with 3 observations in the risk set just prior to time c_3, the probability that $X_{c_3} > 122$ is:

$$P(X_{c_3} > 122) = \frac{1}{4}$$

If $X_{c_3} < 122$, then the probability that $X_{c_3} < 122$ is $1 - \frac{1}{4} = \frac{3}{4}$.

Next, we consider X_{c_2}, who was censored at age 114, conditional on $X_{c_3} < 122$. For X_{c_2}, we do not need to account for censoring of X_{c_3}, since $X_{c_3} < 122$ and thus does not influence the probability. We only know that there were 3 deaths between 116 and 122. Thus, the probability that $X_{c_2} > 122$, given $X_{c_3} < 122$, based on the shifted-$\tilde{A}_{(6)}$ with $\tilde{n}_{c_2} = 6$, is:

$$P(X_{c_2} > 122 \mid X_{c_3} < 122) = \frac{1}{7}$$

Therefore, the probability for $X_{c_2} < 122$, given $X_{c_3} < 122$, is $1 - \frac{1}{7} = \frac{6}{7}$.

Next, we consider X_{c_1}, who was censored at age 112, conditional on both $X_{c_2} < 122$ and $X_{c_3} < 122$. For X_{c_1}, we again do not need to account for censoring for X_{c_2} and X_{c_3}, since both died before 122. It does not matter what the exact values of X_{c_2} and X_{c_3} are within the interval $(114, 122)$, and we only know that there were 6 deaths between 113 and 122. Thus, the probability that $X_{c_1} > 122$, given that $X_{c_2} < 122$ and $X_{c_3} < 122$, based on the shifted-$\tilde{A}_{(8)}$ with $\tilde{n}_{c_1} = 8$, is:

$$P(X_{c_1} > 122 \mid X_{c_2} < 122, X_{c_3} < 122) = \frac{1}{9}$$

Therefore, the probability that $X_{c_1} < 122$, given $X_{c_2} < 122$ and $X_{c_3} < 122$, is $1 - \frac{1}{9} = \frac{8}{9}$.

Consequently, the probability that at least one of the three individuals X_{c_1}, X_{c_2}, and X_{c_3}, whose lifetimes are right-censored at ages 112, 114, and 117, respectively, would have a lifetime greater than $\mathcal{R} = 122$, is:

$$P(G_{122}(0)) = 1 - \prod_{r=1}^{3} \frac{\tilde{n}_{c_r}}{\tilde{n}_{c_r} + 1} = 1 - \left[\frac{3}{4} \times \frac{6}{7} \times \frac{8}{9}\right] = 1 - \frac{4}{7} = 0.4286$$

This example illustrates how to derive the probability for the event of interest $G_{122}(0)$. We do not need to account for censoring in the $A_{(n)}$ setting because we are conditioning on the fact that all individuals are less than $\mathcal{R} = 122$.

Appendix B Proof of Equation (7) with an illustrative example

Proof of Equation 7: For $m = 1$, we consider the lifetime of the first future individual, X_{n+1}, conditional on the fact that all individuals whose lifetimes have been right-censored at censoring times c_r (where $r = 1, 2, \ldots, v$) have failed before the value $\mathcal{R}$. It is crucial to note that, for all right-censored individuals, it does not matter exactly where their lifetimes are, as long as they occur before $\mathcal{R}$. Therefore, the only information we need is the number of individuals in the risk set at time x_0, denoted $\tilde{n}_{x_0} = n$. The probability of the event that $X_{n+1} > \mathcal{R}$, given that all $X_{c_r} < \mathcal{R}$, $r = 1, 2, \ldots, v$, based on the shifted-$\tilde{A}_{(n)}$ as in Equation (5), with $\tilde{n}_{x_0} = n$, is

$$P(X_{n+1} > \mathcal{R} \mid X_{c_1} < \mathcal{R}, X_{c_2} < \mathcal{R}, \ldots, X_{c_v} < \mathcal{R}) = \frac{1}{\tilde{n}_{x_0} + 1} = \frac{1}{n+1}.$$

The probability of the complementary event, that $X_{n+1} < \mathcal{R}$, given the same conditions, is

$$P(X_{n+1} < \mathcal{R} \mid X_{c_1} < \mathcal{R}, X_{c_2} < \mathcal{R}, \ldots, X_{c_v} < \mathcal{R}) = 1 - \frac{1}{n+1} = \frac{n}{n+1}.$$

For $m = 2$, we consider the lifetime of the second future individual, X_{n+2}, conditional on the lifetime of the first future individual, X_{n+1}, and all individuals whose lifetimes have been right-censored at times c_r (where $r = 1, 2, \ldots, v$) having failed before $\mathcal{R}$. Again, the probability that $X_{n+2} > \mathcal{R}$, given that $X_{n+1} < \mathcal{R}$ and all $X_{c_r} < \mathcal{R}$, is derived based on the shifted-$\tilde{A}_{(n)}$, now with $\tilde{n}_{x_0} + 1 = n + 1$, as X_{n+1} has been added. Thus, the probability is

$$P(X_{n+2} > \mathcal{R} \mid X_{n+1} < \mathcal{R}, X_{c_1} < \mathcal{R}, \ldots, X_{c_v} < \mathcal{R}) = \frac{1}{(\tilde{n}_{x_0} + 1) + 1} = \frac{1}{n+2}.$$

The probability for the complementary event, $X_{n+2} < \mathcal{R}$, is

$$P(X_{n+2} < \mathcal{R} \mid X_{n+1} < \mathcal{R}, X_{c_1} < \mathcal{R}, \ldots, X_{c_v} < \mathcal{R}) = 1 - \frac{1}{n+2} = \frac{n+1}{n+2}.$$

In general, for an event $X_{n+i} > \mathcal{R}$ where $i = 2, 3, \ldots, m$, conditional on the previous future individuals and all right-censored individuals, the probability based on the shifted-$\tilde{A}_{(n)}$ is

$$P(X_{n+i} > \mathcal{R} \mid X_{n+1} < \mathcal{R}, \ldots, X_{n+i-1} < \mathcal{R}, X_{c_1} < \mathcal{R}, \ldots, X_{c_v} < \mathcal{R}) = \frac{1}{(\tilde{n}_{x_0} + i - 1) + 1} = \frac{1}{n+i}.$$

For the complementary event $X_{n+i} < \mathcal{R}$, the probability is

$$P(X_{n+i} < \mathcal{R} \mid X_{n+1} < \mathcal{R}, \ldots, X_{n+i-1} < \mathcal{R}, X_{c_1} < \mathcal{R}, \ldots, X_{c_v} < \mathcal{R}) = 1 - \frac{1}{n+i} = \frac{n+i-1}{n+i}. \tag{B2}$$

Since the event of interest, $G_{\mathcal{R}}(m)$, accounts for both future observations and the data set containing the n observations, Equation (A1), which is related to the data set with only the n observations, and Equation (B2), which is related to future observations, are required to compute the probability that all right-censored times exceed $\mathcal{R}$. Therefore, the probability for the event $G_{\mathcal{R}}(m)$, denoted $P_{\mathcal{R}}(G(m))$, is

$$P(G_{\mathcal{R}}(m)) = 1 - \left[\prod_{i=1}^{m} \frac{n+i-1}{n+i} \times \prod_{r=1}^{v} \frac{\tilde{n}_{c_r}}{\tilde{n}_{c_r}+1}\right] = 1 - \left[\frac{n}{n+m} \prod_{r=1}^{v} \frac{\tilde{n}_{c_r}}{\tilde{n}_{c_r}+1}\right].$$

Thus, the proof is complete.

The following example illustrates the probabilities presented in this section.

Example 6. We again use the same data on $n = 10$ observations (as in Example 5). We consider that X_{n+1} and X_{n+2} are the lifetimes of the first and second future individuals to be included in the study. We now ask for the probability that at least one of the three individuals, $X_{c_1}, X_{c_2}, X_{c_3}$, with lifetimes right-censored at ages 112, 114, and 117, or one of the future individuals, X_{n+1} and X_{n+2}, has a lifetime greater than the largest observed value $\mathcal{R} = 122$.

We first consider the lifetime of X_{n+1}, conditional on that $X_{c_1}, X_{c_2}, X_{c_3}$, with right-censored lifetimes at ages 112, 114, and 117, have failed before $\mathcal{R} = 122$. The probability that $X_{n+1} > 122$, given that $X_{c_1} < 122, X_{c_2} < 122, X_{c_3} < 122$, is

$$P(X_{n+1} > 122 \mid X_{c_1} < 122, X_{c_2} < 122, X_{c_3} < 122) = \frac{1}{11}.$$

The probability that $X_{n+1} < 122$, given the same conditions, is

$$P(X_{n+1} < 122 \mid X_{c_1} < 122, X_{c_2} < 122, X_{c_3} < 122) = 1 - \frac{1}{11} = \frac{10}{11}.$$

Next, we consider the lifetime of the second future individual X_{n+2}, conditional on that $X_{n+1} < 122$ and the right-censored individuals have also failed before 122. The probability that $X_{n+2} > 122$, given that $X_{n+1} < 122, X_{c_1} < 122, X_{c_2} < 122, X_{c_3} < 122$, is

$$P(X_{n+2} > 122 \mid X_{n+1} < 122, X_{c_1} < 122, X_{c_2} < 122, X_{c_3} < 122) = \frac{1}{12}.$$

The probability that $X_{n+2} < 122$, given the same conditions, is

$$P(X_{n+2} < 122 \mid X_{n+1} < 122, X_{c_1} < 122, X_{c_2} < 122, X_{c_3} < 122) = 1 - \frac{1}{12} = \frac{11}{12}.$$

Consequently, the probability for the event $G_{122}(2)$, denoted by $P(G_{122}(2))$, is derived as

$$P(G_{122}(2)) = 1 - \left[\prod_{i=1}^{2} \frac{n+i-1}{n+i} \prod_{r=1}^{3} \frac{\tilde{n}_{c_r}-1}{\tilde{n}_{c_r}}\right]$$
$$= 1 - \left[\left(\frac{10}{11} \times \frac{11}{12}\right) \times \left(\frac{3}{4} \times \frac{6}{7} \times \frac{8}{9}\right)\right] = 1 - \frac{40}{84} = 0.5238.$$

Appendix C Proof of the results in Section 4

We now take into account the second largest observed value in the dataset, denoted by $\mathcal{R}_2 = x_{u-1}$, as long as there are no censored observations past it, such that $x_{u-1} > c_v$. We consider the event of interest: for at least one of the individuals whose lifetimes have been right-censored, the actual lifetime value is larger than the second largest observed value, $\mathcal{R}_2$. For ease of notation, let $G_{\mathcal{R}_2}(0)$ denote this event. The probability for the event $G_{\mathcal{R}_2}(0)$ is then found in the same manner as for the event $G_{\mathcal{R}}(0)$ (exceeding the first largest observation), as derived in Section 3.

For individuals whose lifetimes have been right-censored at time c_r, where $r = 1, 2, \ldots, v$, censoring does not need to be considered as long as all these individuals failed before $\mathcal{R}_2$, and we only know the number of individuals between X_{c_r} and $\mathcal{R}_2$. Based on the shifted-$\tilde{A}_{(n)}$ as given in Equation (5), we have

$$P_{X_{c_r}}(\mathcal{R}_2, \mathcal{R}) = P_{X_{c_r}}(\mathcal{R}, \infty) = \frac{1}{\tilde{n}_{c_r}+1},$$

where $\tilde{n}_{c_r}$ is the number of observations in the risk set just before time c_r, for $r = 1, 2, \ldots, v$. Therefore, the probability that $X_{c_r} > \mathcal{R}_2$, given that all other individuals failed before $\mathcal{R}_2$, is

$$P(X_{c_r} > \mathcal{R}_2 | X_{c_{r+1}} < \mathcal{R}_2, \ldots, X_{c_{v-1}} < \mathcal{R}_2, X_{c_v} < \mathcal{R}_2) = \frac{2}{\tilde{n}_{c_r}+1}.$$

The probability for the event of interest, that nobody survives the value $\mathcal{R}_2$, is then

$$P(X_{c_r} < \mathcal{R}_2 | X_{c_{r+1}} < \mathcal{R}_2, \ldots, X_{c_{v-1}} < \mathcal{R}_2, X_{c_v} < \mathcal{R}_2) = 1 - \frac{2}{\tilde{n}_{c_r}+1} = \frac{\tilde{n}_{c_r}-1}{\tilde{n}_{c_r}+1}.$$

Thus, the probability for the event $G_{\mathcal{R}_2}(0)$, denoted by $P(G_{\mathcal{R}_2}(0))$, is given by

$$P(G_{\mathcal{R}_2}(0)) = 1 - \prod_{r=1}^{v} \frac{\tilde{n}_{c_r}-1}{\tilde{n}_{c_r}+1}.$$

Following the same reasoning, we obtain the probability for the event $G_{\mathcal{R}_3}(0)$, where at least one individual, whose lifetime has been right-censored, has an actual

lifetime greater than the third largest observed value, $\mathcal{R}_3 = x_{u-2}$, with $x_{u-2} > c_v$. The probability is

$$P(G_{\mathcal{R}_3}(0)) = 1 - \prod_{r=1}^{v} \frac{\tilde{n}_{c_r} - 2}{\tilde{n}_{c_r} + 1}.$$

In a similar fashion, one can compute the probability for any event where at least one individual's lifetime exceeds a specified largest observed value, as long as it is greater than the largest censored observation at c_v. Specifically, for $\mathcal{R}_j = x_{u-j+1}$, where $x_{u-j+1} > c_v$, we have

$$P(G_{\mathcal{R}_j}(0)) = 1 - \prod_{r=1}^{v} \frac{\tilde{n}_{c_r} - j + 1}{\tilde{n}_{c_r} + 1}.$$

We are now considering the addition of future items to the study, as we did in Section 3. We then examine the event of interest where, for at least one of the individuals whose lifetimes have been right-censored, or one of the $m \geq 1$ future individuals added to the study, the actual lifetime would be larger than the second largest observed value, $\mathcal{R}_2 = x_{u-1}$, with $x_{u-1} > c_v$. Let $G_{\mathcal{R}_2}(m)$ denote this event.

The probability for the event $X_{n+i} > \mathcal{R}_2$, for $i = 1, 2, \ldots, m$, conditional on the failure of all previous future individuals and those whose lifetimes have been right-censored at times c_r, where $r = 1, 2, \ldots, v$, before $\mathcal{R}_2 = x_{u-1}$, is derived based on the shifted-$\tilde{A}_{(n)}$ in Equation (5) as

$$P(X_{n+i} > \mathcal{R}_2 | X_{n+1} < \mathcal{R}_2, \ldots, X_{n+i-1} < \mathcal{R}_2, X_{c_1} < \mathcal{R}_2, \ldots, X_{c_v} < \mathcal{R}_2) = \frac{2}{n+i}.$$

The probability for the event $X_{n+i} < \mathcal{R}_2$, given that $X_{n+1} < \mathcal{R}_2, \ldots, X_{n+i-1} < \mathcal{R}_2$, is

$$P(X_{n+i} < \mathcal{R}_2 | X_{n+1} < \mathcal{R}_2, \ldots, X_{n+i-1} < \mathcal{R}_2, X_{c_1} < \mathcal{R}_2, \ldots, X_{c_v} < \mathcal{R}_2) = 1 - \frac{2}{n+i} = \frac{n+i-2}{n+i}.$$

Thus, the probability for the event of interest $G_{\mathcal{R}_2}(m)$, denoted by $P(G_{\mathcal{R}_2}(m))$, is given by

$$P(G_{\mathcal{R}_2}(m)) = 1 - \left[\prod_{i=1}^{m} \frac{n+i-2}{n+i} \prod_{r=1}^{v} \frac{\tilde{n}_{c_r} - 1}{\tilde{n}_{c_r} + 1} \right],$$

which simplifies to

$$P(G_{\mathcal{R}_2}(m)) = 1 - \left[\frac{n(n-1)}{(n+m)(n+m-1)} \prod_{r=1}^{v} \frac{\tilde{n}_{c_r} - 1}{\tilde{n}_{c_r} + 1} \right].$$

Using the same reasoning, the probability for the event of interest $G_{\mathcal{R}_3}(m)$, where at least one individual (from either the right-censored individuals or the m future individuals) has a lifetime exceeding the third largest observed value $\mathcal{R}_3 = x_{u-2}$, is

given by

$$P(G_{\mathcal{R}_3}(m)) = 1 - \left[\prod_{i=1}^{m} \frac{n+i-3}{n+i} \prod_{r=1}^{v} \frac{\tilde{n}_{c_r}-2}{\tilde{n}_{c_r}+1}\right],$$

which simplifies to

$$P(G_{\mathcal{R}_3}(m)) = 1 - \left[\frac{n(n-1)(n-2)}{(n+m)(n+m-1)(n+m-2)} \prod_{r=1}^{v} \frac{\tilde{n}_{c_r}-2}{\tilde{n}_{c_r}+1}\right].$$

Similar to the above explanation, one could straightforwardly obtain the probability for the event that at least one of the individuals whose lifetimes have been right-censored, or one of the $m \geq 1$ future individuals added to the study, has an actual lifetime value larger than any other largest observed value, as long as it is greater than the largest censored observation at c_v. Consequently, the probability that someone survives any largest observed value recorded in a data set, when it exceeds the largest censored observation, increases as it is calculated backwards from the largest recorded value to the j-th largest observed value, provided that it surpasses the largest censored observation. In general, for $\mathcal{R}j = x_{u-j+1}$, where $x_{u-j+1} > c_v$, the probability is given as follows:

$$P(G_{\mathcal{R}_j}(m)) = 1 - \left[\prod_{i=1}^{m} \frac{n+i-j}{n+i} \prod_{r=1}^{v} \frac{\tilde{n}_{c_r}-j+1}{\tilde{n}_{c_r}+1}\right]$$

which simplifies to:

$$P(G_{\mathcal{R}_j}(m)) = 1 - \left[\frac{n(n-1)\ldots(n-j+1)}{(n+m)(n+m-1)\ldots(n+m-j+1)} \prod_{r=1}^{v} \frac{\tilde{n}_{c_r}-j+1}{\tilde{n}_{c_r}+1}\right].$$

Thus, the proof is complete.

References

[1] Alves, I.F., Neves, C.: Estimation of the finite right endpoint in the gumbel domain. Statistica Sinica, 1811–1835 (2014)

[2] Alves, I.F., Neves, C., Rosário, P.: A general estimator for the right endpoint with an application to supercentenarian women's records. Extremes **20**, 199–237 (2017)

[3] Coolen, F.P.A.: Nonparametric predictive inference. In: Lovric, M. (ed.) International Encyclopedia of Statistical Science, pp. 968–970. Springer, Berlin, Heidelberg (2011)

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

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

[6] Coolen, F.P.A.: On nonparametric predictive inference and objective bayesianism. Journal of Logic, Language and Information, 21–47 (2006)

[7] Hill, B.M.: Parametric models for a(n): splitting processes and mixtures. Journal of the Royal Statistical Society. Series B, 423–433 (1993)

[8] Hill, B.M.: De finetti's theorem, induction, and bayesian nonparametric predictive inference (with discussion). In: Bayesian Statistics, pp. 211–241 (1988). Oxford University Press

[9] Hill, B.M.: Bayesian nonparametric prediction and statistical inference. In: Geisser, S., Hodges, J.S., Press, S.J., Zellner, A. (eds.) Bayesian Analysis in Statistics and Econometrics, pp. 43–94. Springer, New York, NY (1992)

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

[11] Yan, K.J.: Nonparametric predictive inference with right-censored data. PhD Thesis, Durham University. URL https://maths.durham.ac.uk/stats/people/fc/thesis-KJY.pdf (2002)

[12] Maturi, T.A.: Nonparametric predictive inference for multiple comparisons. PhD-thesis, University of Durham. URL https://maths.durham.ac.uk/stats/people/fc/thesis-AAN.pdf (2010)

[13] Kaplan, E.L., Meier, P.: Nonparametric estimation from incomplete observations. Journal of the American Statistical Association, 457–481 (1958)

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

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