Received: 5 November 2021 | Revised: 17 June 2022 | Accepted: 10 October 2022

DOI: 10.1002/pst.2271

MAIN PAPER

WILEY

On the use of extreme value tail modeling for generalized pairwise comparisons with censored outcomes

Mickaël De Backer[1] | **Catherine Legrand**[1] | **Julien Péron**[2,3,4] | **Alexandre Lambert**[5] | **Marc Buyse**[6,7]

[1]ISBA/LIDAM, UCLouvain, Louvain-la-Neuve, Belgium

[2]CNRS, UMR 5558, Laboratoire de Biométrie et Biologie Evolutive, Equipe Biostatistique-Santé, Université Lyon 1, Villeurbanne, France

[3]Service de Biostatistique et Bioinformatique, Hospices Civils de Lyon, Lyon, France

[4]Oncology department, Hospices Civils de Lyon, Lyon, France

[5]Global Biometrics and Data Sciences, Bristol-Myers Squibb, Braine-l'Alleud, Belgium

[6]International Drug Development Institute (IDDI), Louvain-la-Neuve, Belgium

[7]Interuniversity Institute for Biostatistics and Statistical Bioinformatics (I-BioStat), Hasselt University, Diepenbeek, Belgium

Correspondence
Mickaël De Backer, ISBA/LIDAM, UCLouvain. Voie du Roman Pays 20, B-1348 Louvain-la-Neuve, Belgium.
Email: mickael.debacker@uclouvain.be

Funding information
Walloon Region, Grant/Award Number: BioWin Consortium Agreement No 7979

Abstract

In randomized clinical trials, methods of pairwise comparisons such as the 'Net Benefit' or the 'win ratio' have recently gained much attention when interests lies in assessing the effect of a treatment as compared to a standard of care. Among other advantages, these methods are usually praised for delivering a treatment measure that can easily handle multiple outcomes of different nature, while keeping a meaningful interpretation for patients and clinicians. For time-to-event outcomes, a recent suggestion emerged in the literature for estimating these treatment measures by providing a natural handling of censored outcomes. However, this estimation procedure may lead to biased estimates when tails of survival functions cannot be reliably estimated using Kaplan–Meier estimators. The problem then extrapolates to the other outcomes incorporated in the pairwise comparison construction. In this work, we suggest to extend the procedure by the consideration of a hybrid survival function estimator that relies on an extreme value tail model through the Generalized Pareto distribution. We provide an estimator of treatment effect measures that notably improves on bias and remains easily apprehended for practical implementation. This is illustrated in an extensive simulation study as well as in an actual trial of a new cancer immunotherapy.

KEYWORDS

clinical trial, generalized pairwise comparisons, generalized Pareto distribution, Kaplan–Meier, multi-criteria analysis

1 | INTRODUCTION

In randomized clinical trials, several outcomes of interest are commonly evaluated on each individual, among which some may measure the time until an event of interest occurs. Primary analysis usually focuses on a single (primary) endpoint. When this endpoint is a time-to-event variable, traditional estimators of the treatment benefit typically rely on a proportional hazards (PH) assumption, where one assumes that the ratio of hazards of an event in both treatment groups is constant over time. Testing for a treatment effect is then performed via a log-rank procedure, known to have maximum power in case of PH, and the treatment effect is usually summarized by estimating the hazard ratio through a Cox regression model. However, as is well-known, such a measure only has a straightforward interpretation for summarizing the treatment effect when the restrictive PH assumption is met. In practice, there are many situations where

this is not the case, and common examples are anticancer immunotherapy trials, such as the CA184-024 melanoma trial considered in this manuscript.

An alternative analysis for such time-to-event data is based on pairwise comparison methods, see for example, ref. [1–3]. These procedures are praised for their ability to handle multiple outcomes, possibly of a different nature, while still resulting in an easily interpretable treatment effect measure. Additionally, while they also maintain a simple relation with the hazard ratio in case of PH, these nonparametric methods do not rely on PH assumptions. This broadens their range of applicability for appropriately measuring the treatment effect in contrast to the hazard ratio, see ref. [4,5] or [6,7] in the context of immunotherapy trials. As a result, methods of pairwise comparison have gained in popularity for analyzing clinical trial data.

In this context, we concentrate on the treatment effect measure suggested by ref. [1], denoted by Δ_τ and named here 'Net Benefit'. For ease of reading, we present mainly a univariate setting in this paper. Our suggestions are however directly applicable to multivariate settings as well, as will become clear later. Let $T_1^t, ..., T_{n^t}^t$ denote the time-to-events for the individuals in the treatment group, considered to be i.i.d. copies of $T^t \sim F^t$, while $T_1^c, ..., T_{n^c}^c$ are the outcomes of the control group, considered to be i.i.d. copies of $T^c \sim F^c$ with T^t and T^c being independent. The Net Benefit Δ_τ is defined as

$$\Delta_\tau := \mathbb{P}(T^t > T^c + \tau) - \mathbb{P}(T^c > T^t + \tau),$$

for some $\tau \geq 0$ corresponding to a threshold of clinical relevance, see for example, ref. [1] or [3] and references therein. In this simple situation with one outcome of interest, Δ_τ is thus interpreted as the probability that a randomly drawn patient from the treatment group survives at least τ units of time longer with respect to the event of interest than a randomly drawn patient from the control group, minus the probability of the opposite situation.

In practice, when one disposes of fully observed samples for both groups, a natural estimation of Δ_τ, as proposed by ref. [1] and readily extending the U-statistic of the Wilcoxon-Mann–Whitney test, is obtained by

$$\widehat{\Delta}_\tau = \frac{1}{n^t n^c} \sum_{i=1}^{n^t} \sum_{j=1}^{n^c} \left[1\left(T_i^t > T_j^c + \tau\right) - 1\left(T_j^c > T_i^t + \tau\right) \right]. \tag{1.1}$$

Estimating Δ_τ thus amounts to comparing all pairs of observed time-to-events for both groups, and classify the latter pairs as either 'wins' (i.e. $T_i^t > T_j^c + \tau$), 'losses' (i.e., $T_j^c > T_i^t + \tau$) or 'neutral' pairs (i.e., $| T_j^c - T_i^t | < \tau$), an approach that is equivalent to the one required for the computation of the win ratio proposed by ref. [2].

However, a common nuisance when studying time-to-event outcomes is the possible presence of (right-)censoring in the data, as some individuals may leave the study early or may not have yet experienced the event of interest at the time of the analysis. Hence, instead of relying on some realizations of the variables T^t and T^c, one only observes the minimum of it and censoring variables. That is, the observed samples are $\left(Y_i^\ell, \eta_i^\ell\right), i = 1, ..., n^\ell$, where $Y^\ell = \min\left(T^\ell, C^\ell\right)$ and $\eta^\ell = 1\left(T^\ell \leq C^\ell\right)$, $\ell \in \{c, t\}$. Providing an appropriate estimation procedure for Δ_τ then constitutes a challenge since we no longer dispose of fully observed data.

For the situation with $\tau = 0$, a first suggestion in this context was historically proposed by ref. [8], where the rationale is to keep classifying pairs as wins or loses but only if there is no ambiguity regarding the category to which the pair belongs. All pairs which we are not able to classify due to censoring are considered 'uninformative', regardless of the observed times composing the pair. When considering several outcomes, these uninformative pairs can then be classified based on one of the other outcomes. Unsurprisingly, Gehan's procedure for only one outcome is highly sensitive to the censoring proportion in the data, as more and more pairs are uninformative for higher censoring proportions.[4] pointed out that under proportional hazards, the censoring-induced bias in the estimate of the univariate Net Benefit could be removed by using the number of informative pairs (instead of the total number of pairs) in the denominator of (1.1).

The procedure recently proposed by ref. [3], which we will henceforth refer to as the 'Peron' procedure, aims at avoiding the censoring-induced bias by extending the ideas of ref. [9]. In particular, the suggestion is to replace the 'definite' classification of a pair as win, loss, or neutral, by a corresponding probability for each pair to belong to one of these three categories based on all available data. Using this procedure, the estimator of the Net Benefit (1.1) is replaced by

$$\widehat{\Delta}_\tau = \frac{1}{n^t n^c} \sum_{i=1}^{n^t} \sum_{j=1}^{n^c} \left[\widehat{\mathbb{P}}\left(T_i^t > T_j^c + \tau | \Omega_n \right) - \widehat{\mathbb{P}}\left(T_j^c > T_i^t + \tau | \Omega_n \right) \right], \tag{1.2}$$

where Ω_n denotes all relevant information in the data for the estimation of the probabilities. Hence, instead of fully contributing to one of the classification groups, the contributions of pairs of observations may here be split over groups with weights corresponding to conditional probabilities of belonging to each group. Of note, for informative pairs (i.e., not 'uninformative' in Gehan's sense), the methodology uses the same classification as Gehan's rule.

In practice, Equation (1.2) suggests one has to estimate quantities of the type $\mathbb{P}\left(T_i^t > T_j^c + \tau | \Omega_n \right)$, whose calculation will depend on the observed values of Y_i^t, Y_j^c, η_i^t and η_i^c contained in the pair (see Table 3 in ref. [3]). For instance, if one disposes of a pair for which both observed outcomes are censored, that is, $\eta_i^t = \eta_j^c = 0$, then its probability to be a win when the observed outcomes are such that $| Y_i^t - Y_j^c | < \tau$ is given by

$$\mathbb{P}\left(T_i^t > T_j^c + \tau | Y_i^t, Y_j^c, \eta_i^t = 0, \eta_j^c = 0 \right) = \frac{-1}{S^t\left(Y_i^t\right) S^c\left(Y_j^c\right)} \int_{Y_j^c}^{\infty} S^t(s+\tau) \mathrm{d}S^c(s),$$

where $S^\ell(s) = \mathbb{P}\left(T^\ell > s\right)$, for $\ell \in \{c, t\}$.

In terms of estimation, one is required to replace the true survival functions S^ℓ, evaluated over the whole open interval $\left[Y_j^c, \infty\right)$, by appropriate estimators $\widehat{S}^\ell(\cdot)$, $\ell \in \{c, t\}$. Péron et al. used the Kaplan–Meier estimator for this task, although the authors acknowledged that when the largest observation is censored, the Kaplan–Meier estimator is only defined up to that point. A direct consequence is that, due to censoring, one may underestimate the probabilities to be in either of the three categories if the tails of S^t or S^c are not suitably estimated. Hence, these estimated probabilities do not sum up to 1, and the remaining probability that is not assigned to either of the categories is then considered 'uninformative'. In the context of Net Benefit or win ratio estimation, the Peron procedure will then generate a growing proportion of uninformative pairs among all pairs when the proportion of censoring increases, simply because the estimators for the survival functions are no longer appropriate in the tails. Ultimately, this induces a bias for $\widehat{\Delta}_\tau$ towards 0 (value of no treatment effect) as very few pairs with long times to event remain informative, a problem similar to the Gehan procedure although much attenuated in comparison to the latter.

In this paper, we aim at improving on the Peron procedure for pairwise comparison of censored outcomes, by relying on a suitable estimation of $S^t(\cdot)$ and $S^c(\cdot)$ over their whole support. That is, we suggest to replace the use of the Kaplan–Meier estimator by a hybrid semiparametric estimator that is composed of a Kaplan–Meier estimator in most of the observed support of the survival function, and completed with a suitable extrapolation for the tail of the distribution. As will be detailed later, instead of relying on common parametric families of distributions for this extrapolation, our proposal takes its roots in extreme value theory for this task. By adopting this hybrid survival function for the estimation of the Net Benefit or win ratio, we will be able to reduce (if not suppress) the bias towards 0 of the treatment effect estimator. Additionally, by suppressing the need to rely on 'uninformative' weights, our procedure avoids unnecessary complications when considering multiple outcomes; the arguments of ref. [3] can then readily be used for this multivariate setting, be it for the construction of the Net Benefit, the win ratio, or any other relevant treatment effect measure based on pairwise comparison (see e.g., ref. [10]).

The rest of this paper is organized as follows. Section 2 summarizes the results of extreme value theory that will be exploited in this paper for the construction of the hybrid survival function estimator. Section 3 details the procedure using the hybrid survival function, henceforth referred to as 'EVT', for the estimation of Δ_τ. An extensive simulation study is conducted in Section 4 to investigate as many aspects as possible of the EVT procedure in comparison with the Peron procedure. Lastly, an application to an actual randomized trial of a new cancer immunotherapy is developed in Section 5, where we illustrate how the EVT procedure improves over the Peron procedure whilst keeping all of its advantages.

2 | HYBRID SURVIVAL FUNCTION ESTIMATION

In this section, we consider the problem of estimating, in a flexible and general way, a survival function over its whole support in the presence of censoring. Figure 1 illustrates the usual issue of appropriately estimating the tail behavior of

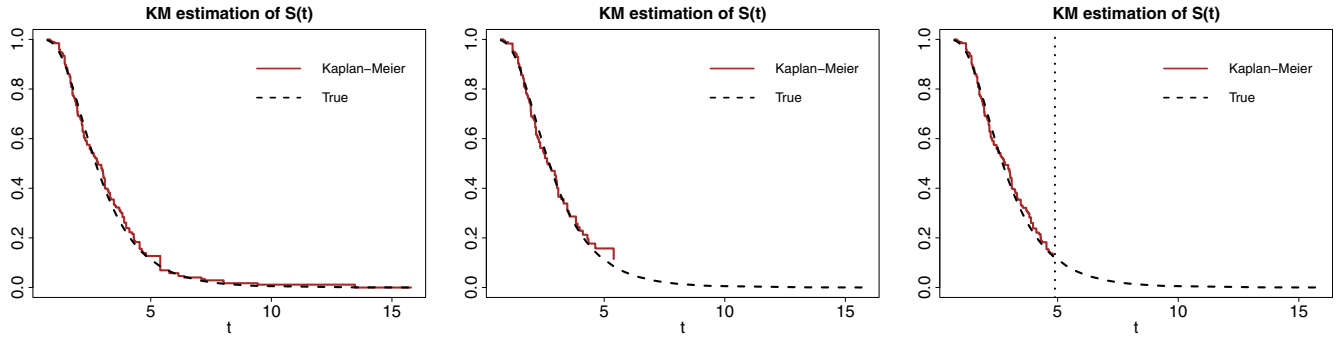

FIGURE 1 Illustration of Kaplan-Meier (KM) estimator of survival functions for samples with either 10% of censoring (left), 50% censoring (middle), or 10% of censoring during the study and 10% of administrative censoring (right).

the survival function when the last observation in the sample is censored. Our context is thus implicitly one where completing survival curves towards 0 is both relevant and reasonably feasible. That is, we assume on the one side that the limited amount of information in the tails of the observable curves is due to practical considerations regarding time-to-event variables, rather than indications about the behavior of the underlying populations (for instance, the presence of a proportion of subjects immune to the event of interest). On the other side, we assume that the observable largest times are representative enough of the underlying population survival curves. Typical situations for our context include the presence of an 'important' censoring proportion in the data, for instance when the event of interest affects a small proportion of patients during time of follow-up, or in interim analyses of a long-term trial. As is conventionally done in survival analysis, we will consider two potential sources of censoring: 'follow-up' censoring which arises from the limited follow-up available on some patients at the time of the analysis (including some patients who dropped out of the trial and for whom no longer follow-up can be obtained, and 'administrative' censoring, which arises from the fact that the analysis takes place whilst follow-up is still on-going).

Several authors have already studied the ambition of completing Kaplan–Meier curves with a parametric tail, mainly in relation to mean (residual) lifetime estimation (MRL). Indeed, MRL can only be estimated in a nonparametric fashion if the last observation of the study is an event, which would lead the Kaplan–Meier estimator of the survival function to drop to 0. Early publications include,[11–13] where extrapolation of the tail of the Kaplan–Meier estimator is carried out based on a parametric fit that is estimated from all the data. As highlighted for instance by ref. [14,15], such a parametric modeling typically provides a poor fit to the tail of the curve. This is to be understood by the fact that most events lie at the beginning of the time to event curve, driving the parametric fit to be much more appropriate for this part of the curve than for the tail. The latter being in fact of primary interest, a natural alternative is to base the parameter estimation solely on tail data, as proposed in ref. [14–16]. All these approaches however suffer from the inherent lack of data in the tail, and still leave the analyst with the delicate choice of a particular parametric model. Goodness-of-fit procedures have been proposed for the latter task, for instance in ref. [13,15], but in practice these have little capacity to detect a poor fit in the presence of censoring due to the limited available information in the data.

In order to circumvent these shortcomings,[17] recently suggested to opt for a hybrid estimation of survival functions for MRL estimation, with extrapolation in the tail based on extreme value theory. In our context, several justifications underpin a similar choice instead of a purely parametric or nonparametric approach. First, the parametric modeling that will result from extreme value theory only applies to the tails of the survival functions, which is the region where the nonparametric Kaplan–Meier starts to fail. Hence, we avoid the need to choose a parametric model for the survival function in a region of its support where a Kaplan–Meier estimator already provides a suitable and flexible tool. Second, the domain of extreme value theory is explicitly designed to model probabilities of events that are more extreme than any previously observed. In contrast to more traditional parametric assumptions for survival functions, this makes extreme value theory well suited for our objective given that we are in a context of limited information in the tail due to censoring. Of note, the use of extreme value theory also allows us to model each survival curve separately, thereby avoiding the need to adopt any other restrictive assumption (e.g., PH) about the joint behavior of the latter. In short thus, we only model the tails of the survival functions parametrically and separately, and we do not require any parametric choice of family distributions.

Turning now to the theory of extreme values, the latter is based on two main approaches for the consideration of 'extreme' events: block maxima and peaks-over-thresholds (see e.g., ref. [18]). We will focus on the second approach, where the idea is to consider as extreme any observation that is above some fixed high time threshold $u \in \mathbb{R}^+$.

More concretely, considering only one group of individuals in order to avoid unnecessary notations, our interest lies here in the behavior of the excesses of our time to event above some fixed threshold, conditionally on the knowledge that the threshold is indeed exceeded. That is, we are interested in the limiting distribution of $T-u \mid T>u$, where $T \sim F_T$ is the time to event of interest. The cornerstone result on which we build the hybrid survival function estimator is the so-called Pickands-Balkema-de Haan theorem:[19,20] under core conditions for F_T mentioned below, we have for $t>0$ and $u \to \infty$,

$$\mathbb{P}(T-u \geq t | T>u) \to G(t) := \left[1+\frac{\xi t}{\sigma_u}\right]_+^{-1/\xi}, \tag{2.1}$$

where $[t]_+ = \max(0,t)$, $\sigma_u = \sigma + \xi(u-\mu)$, and where $\mu, \sigma > 0$ and ξ are location, scale and shape parameters, respectively. The distribution G is called the Generalized Pareto Distribution (GPD). Leaving aside for now the conditions on F_T under which the theorem holds, one observes that regardless of the shape of the center of this distribution, the theorem states that the shape of its tail (i.e., the excesses above some threshold) always takes a very special form, provided we are far enough in the tail ($u \to \infty$). In other words, the GPD offers a natural parametric approximation for the modeling of conditional distributions of excesses, as long as the threshold u is chosen to be sufficiently large for the asymptotic result to be reasonably employed.

The necessary and sufficient condition on F_T under which the above theorem holds is usually called the extended regular variation property.[21] As the name suggests, this condition essentially states that the distribution F_T is sufficiently regular in order for tail behaviors to be captured in a single broad family of parametric distributions. To illustrate the latter condition, some examples of distributions for F_T that are encompassed by the theorem are given below, categorized by values of ξ as this parameter completely determines the upper tail behavior and the support of G (see e.g., ref. [18]):

$\xi > 0$ is the *heavy*-tailed *Fréchet* case. Examples of distributions F_T included here are Pareto or Fréchet distributions.

$\xi = 0$ is the *light*-tailed *Gumbel* case. Examples include the exponential, Weibull and Log-normal distributions.

$\xi < 0$ is the *bounded*-tailed *Weibull* case. Examples include beta or uniform distributions.

The theorem only concerns the limiting behavior of the tail, hence we suggest to flexibly estimate the survival curve in two parts: first using a nonparametric estimator until the threshold u, whose choice is to be discussed further, and then using the above-described result to parametrically extrapolate the tail behavior beyond u. That is, since for $t>0$,

$$\mathbb{P}(T \geq t+u) = \mathbb{P}(T \geq t+u | T \geq u)\mathbb{P}(T \geq u),$$

a natural semiparametric estimator for $S_T(t) = 1 - F_T(t)$ is given by

$$\widehat{S}_T(t) = \begin{cases} \widehat{S}_{KM}(t), & t \leq u \\ \widehat{S}_{KM}(u)\left[1+\frac{\widehat{\xi}(t-u)}{\widehat{\sigma}_u}\right]_+^{-1/\widehat{\xi}}, & t > u \end{cases} \tag{2.2}$$

where $\widehat{S}_{KM}$ is the classical nonparametric Kaplan–Meier estimator of S_T, and where $\widehat{\xi}$ and $\widehat{\sigma}_u$ are appropriate estimators of ξ and σ_u, respectively. An illustration of this estimator is provided in Figure 2.

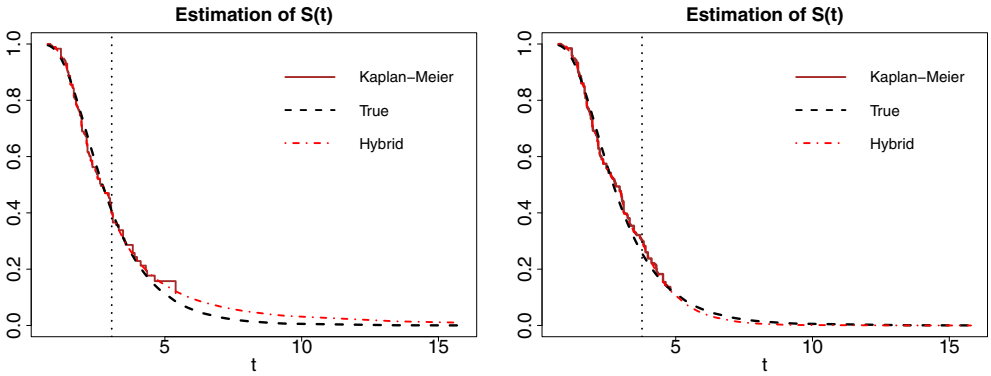

FIGURE 2 Illustration of semi parametric estimator of survival functions for samples with either 50% censoring (left), or 10% of censoring during the study and 10% of administrative censoring (right).

3 | IMPROVED NET BENEFIT ESTIMATION PROCEDURE

To use the suggested hybrid estimator in the context of estimating Δ_τ, Equation (2.2) requires both an appropriate choice for the thresholds $u^\ell, \ell \in \{t, c\}$, and estimation of the parameters of the GPD. We consider again the one sample situation for ease of presentation. That is, we observe $(Y_i, \eta_i), i = 1, ..., n,$ where $Y_i = \min(T_i, C_i)$ and $\eta_i = 1(T_i \leq C_i)$ for C a censoring variable independent of T. Estimation of ξ and σ_u is carried out in two steps:

1. Choose a threshold u and select the m observations Y_i that are above u. The choice of u is a typical bias-variance trade-off: one has to choose u large enough for the asymptotic result of (2.1) to be reasonably employed, but on the other hand u should not be too large in order to dispose of a sufficiently important sample size m of extreme observations for estimation of the parameters. While for complete observations, many advanced methods have been suggested in the literature (see e.g., ref. [22] for a review), for censored data an automatic choice of u is still the subject of ongoing research. The additional difficulty as compared to complete observations is that the threshold that would be in some sense optimal may very well lie beyond the last observed event, an issue that may be especially acute in the case of administrative censoring.

Nevertheless, in our context, extensive simulation studies seem to corroborate the suggestion of ref. [17] (see also ref. [23] for more extensive simulations) consisting in simply setting u as the 80% quantile of the observed event times. In other words, in both our simulation study and our real data analysis, we will leave 20% of observed event times above u in order to estimate ξ and σ_u. While such a choice is arbitrary, simulations suggest that it is reasonable in most situations, and that the bias reduction for the Net Benefit estimator is similar for a wide range of choices.

2. Based on the sample $\left(Y_i^u, \eta_i\right), i = 1, ..., m,$ where $Y_i^u := Y_i - u \mid Y_i > u$, obtain estimators of ξ and σ_u. Several methodologies may potentially be employed, see for example,[24,25,26] suggest a one-step Newton–Raphson approximation to the maximum likelihood estimators. Here, we adopt maximum likelihood estimators for ξ and σ_u obtained by minimizing the censored negative log-likelihood, that is,

$$\left(\widehat{\xi}, \widehat{\sigma}_u\right) = \underset{(\xi, \sigma_u)}{\arg\min} \left\{ \log(\sigma_u) \sum_{i=1}^{m} \eta_i + \sum_{i=1}^{m} \left(\frac{1}{\xi} + \eta_i\right) \log\left[1 + \frac{\xi Y_i^u}{\sigma_u}\right]_+ \right\}.$$

In the presence of censoring, this approach can be constrained to ensure that the obtained GPD distribution has finite k-th moment. This is a consequence of the fact that the latter moment only exists if $\xi < 1/k$. Hence, for real data applications with time-to-events variables, it is sufficient to constrain the optimization with $\xi < 1/2$ so that the distribution has a finite variance.

We can now use the hybrid estimator (2.2) in each treatment group in order to improve on the Peron method. For instance, taking the example of the censored pair reported in Section 1, if the observation Y_j^c and the thresholds u^t and u^c are such that $Y_j^c < u^c < u^t - \tau$, then the estimated probability to be a win becomes

$$\begin{aligned}
&\widehat{\mathbb{P}}\left(T_i^t > T_j^c + \tau | Y_i^t, Y_j^c, \eta_i^t = 0, \eta_j^c = 0\right) \\
&= \frac{-1}{\widehat{S}^t\left(Y_i^t\right)\widehat{S}^c\left(Y_j^c\right)} \left\{ \int_{Y_j^c}^{u^c} \widehat{S}_{KM}^t(s+\tau)\, \mathrm{d}\widehat{S}_{KM}^c(s) \right. \\
&\left. + \int_{u^c}^{u^t-\tau} \widehat{S}_{KM}^t(s+\tau)\, \mathrm{d}\widehat{S}_{EVT}^c(s) + \int_{u^t-\tau}^{\infty} \widehat{S}_{EVT}^t(s+\tau)\, \mathrm{d}\widehat{S}_{EVT}^c(s) \right\},
\end{aligned}$$

where $\widehat{S}_{EVT}^{\ell}(s) = \widehat{S}_{KM}^{\ell}\left(u^{\ell}\right)\left[1 + \widehat{\xi}^{\ell}\left(s - u^{\ell}\right)/\widehat{\sigma}_u^{\ell}\right]_+^{-1/\widehat{\xi}^{\ell}}$, $\ell \in \{c, t\}$.

4 | SIMULATION STUDY

In this section, we use Monte Carlo simulations to assess the performance of the EVT procedure for estimation of Δ_τ, in particular with regard to the censoring-induced bias in a univariate setting. To explore this, we first describe the settings considered, with the simulated data being subject to both administrative and lost to follow-up censoring. Censoring will always be considered non-informative. We then proceed to describe the different estimators of Δ_τ considered here. In order to provide a full picture of the performance of the EVT procedure against its competitors, several settings will then be considered where we let the following characteristics vary: censoring proportions, thresholds of clinical relevance τ, population Net Benefit Δ_τ and type of time-to-event distributions. Of note, in all cases the censoring distribution is considered to be the same in both groups, while actual censoring proportions may differ between groups as these depend on the interplay between censoring times and time to event times. All censoring distributions will be calibrated to obtain the desired censoring proportions defined for the treatment group.

4.1 | Simulation settings

We consider two main settings for data generation:

Setting 1 We consider the following:

a. Time-to-events: $T_1^t, ..., T_n^t$ i.i.d. $Exp(\lambda_t)$ and $T_1^c, ..., T_n^c$ i.i.d. $Exp(1)$ for n_t and $n_c \in \{500, 1000\}$. The parameter λ_t is chosen such that $\Delta_\tau \in \{0, 1/3, 1/2\}$, where we may determine that

$$\Delta_\tau = \frac{e^{-\tau\lambda_t} - \lambda_t e^{-\tau}}{1 + \lambda_t}.$$

Note that this is a situation of proportional hazards, where the simple correction of ref. [4] (dividing by the number of informative pairs) works well.

b. Censoring times: $C_1, ..., C_{n^t+n^c}$ i.i.d. $Unif(0, M)$ and additional administrative censoring with end of study at time t_M. The parameters M and t_M are chosen in order to attain censoring proportions during the study $p_c \in \{0\%, ..., 45\%\}$ and administrative censoring $p_c^M \in \{0\%, 10\%, 20\%\}$, focusing only on the treatment group. The procedure to calibrate M and t_M will be described next. For now, note that given $\Delta_\tau \geq 0$ in our settings, we will almost always have a greater censoring proportion in the treatment group than in the control group, and the difference will grow as Δ_τ grows.

Setting 2 We consider the following:

a. Time-to-events: Log-normal distributions with shift in mean, that is $T_1^t, ..., T_n^t$ i.i.d. Log - $Normal(\mu_t, 0.5)$ and $T_1^c, ..., T_n^c$ i.i.d. Log - $Normal(1, 0.5)$ for $n_t, n_c \in \{500, 1000\}$. The parameter μ_t is again chosen such that $\Delta_\tau \in \{0, 1/3, 1/2\}$. This is a situation where the simple correction of[4] breaks down.
b. Censoring times: as in setting 1.

Now, the procedure to define M and t_M in settings 1 and 2, with $M > t_M$ to be coherent, is similar to the paper of.[17] That is, since p_c is the proportion of censoring during the study and C is uniformly distributed, we first have

$$p_c = \mathbb{P}(C < T, C < t_M) = \frac{1}{M}\int_0^{t_M} S_T(c)\mathrm{d}c.$$

On the other hand, for an observation to be administratively censored, either it is fully observed but at a time greater than t_M, or it is classically censored but again at a time greater than t_M. Hence, some straightforward calculations lead to

$$p_c^M = \mathbb{P}(t_M < T < C) + \mathbb{P}(t_M < C < T) = (M - t_M)\frac{S_T(t_M)}{M}.$$

Depending on the particular type of distribution for T^t and the choices of p_c and p_c^M, we thus have a system of two equations with two unknowns that can be numerically solved for M and t_M.

As a remark, note that the procedure used to mimic administratively censored data is here simplified by the neglect of accrual times, and hence somehow considers that all observations that are censored at t_M entered the study at the same time. This is of course of no impact on the validity of the methodology nor on the major tendencies that will be observed in this section, as the importance of the procedure lies in correcting for incomplete survival function estimators, and the latter would also be present if all patients had different accrual times.

Lastly, the following estimators of Δ_τ are considered:

Peron: estimator of ref. [3] implemented in the R package BuyseTest.

Gehan: estimator of ref. [8] implemented in the R package BuyseTest.

EVT: proposed estimator with the weighting scheme of ref. [3] along with the above-exposed hybrid survival function estimators.

EVT ben: Benchmark for EVT, as it is the exact same estimator as EVT, but with parameters ξ and σ_u of the GPD distribution estimated with full knowledge of observations, that is, as if there was no censoring. This will serve in order to illustrate the impact of appropriately estimating the GPD parameters.

4.2 | Varying censoring proportions

We start our simulation study with an investigation of the impact of increasing censoring proportions on the behavior of the considered estimators. To this end, we focus on setting 1 with fixed $\tau = 0$ and $\Delta_\tau = 1/3$. Figure 3 reports the simulation results for different sample sizes and different censoring proportions, due to lost to follow-up and administrative censoring.

As expected, the Gehan estimator is extremely sensitive to growing censoring proportions, as already discussed in the literature. The Peron, EVT and EVT benchmark estimators have a very similar behavior when the censoring proportion is moderate. This is because appropriate tail modeling is not yet crucial, given that Kaplan–Meier estimation is still mostly available in the latter region as well. However, when censoring proportions increase, the anticipated bias towards 0 of the Peron estimator begins to manifest itself, whereas the EVT and EVT benchmark estimators resist better: tail modeling becomes more crucial. With heavy censoring, for example, if $p_c = 45\%$ and $p_c^M = 20\%$, EVT worsens as appropriately modeling the tail becomes challenging based on the available information. It should however be noted that the deterioration primarily concerns the variance. These results suggest that the somewhat arbitrary choice of a threshold u^ℓ corresponding to the 80% quantile of the observed events seems acceptable when applied to the estimation of Δ_τ estimation, although the choice is of course not to be taken as indisputable in practice.

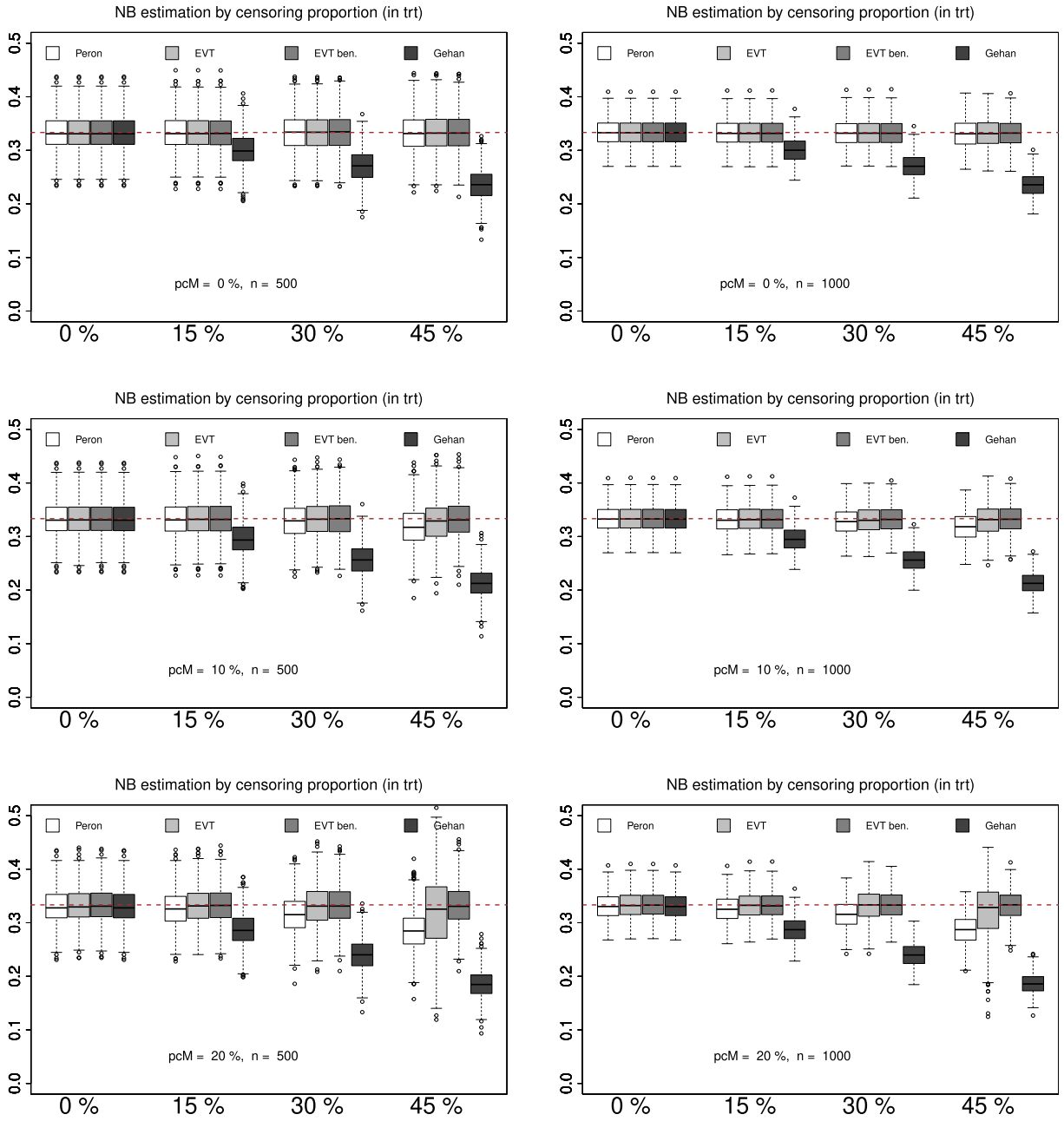

FIGURE 3 Δ_τ (Net Benefit, 'NB') estimation for 500 datasets for setting 1 based on $\tau = 0$, $\Delta_\tau = 1/3$, for n^t and $n^c \in \{500,1000\}$, $p_c \in \{0, ..., 45\%\}$ and $p_C^M \in \{0, 10\%, 20\%\}$. The true Δ_τ is in dotted brown.

Two last comments may be warranted here. First, when considering increasing sample sizes (i.e., left column versus right column), it is interesting to observe that for most situations where the Peron estimator is biased, it does not seem to fade away when sample sizes increase. This is a direct consequence of the fact that the Kaplan–Meier estimator is not consistent in the tail for such simulation settings. And as a last comment, it is worth emphasizing that the variances of the Peron and Gehan estimators may not be compared with the variance of the proposed estimator. Indeed, both Gehan and Peron will classify less and less observations as either wins, losses or neutral when censoring increases, and their variance may therefore slightly decrease with increasing p_c and p_c^M, since classifying fewer observations induces less variability. In contrast, EVT still attempts to classify every pair, and exhibits therefore an increasing variance for greater censoring proportions.

4.3 | Varying thresholds τ

In this section, we consider the impact of an increasing threshold of clinical relevance τ. In particular, we consider again setting 1 with a Net Benefit $\Delta_0 = 1/3$ for $\tau = 0$, with $p_c \in \{15\%, 30\%\}$ and $p_c^M \in \{10\%, 20\%\}$ (i.e., similar to Figure 3 for $\tau = 0$). Here we let τ increase, and consider again two different samples sizes for each group, as n^t and $n^c \in \{500, 1000\}$. Hence, everything is similar to the previous section except for the definition of the population parameter of interest.

The results of this simulation study are presented in Figure 4. It is important to realize that, in the Peron procedure, considering higher thresholds of clinical significance amounts to exploring further the tails of S^ℓ, $\ell \in \{c, t\}$. It then comes as no surprise that the performance of the Peron estimator deteriorates as τ grows, since we are here in settings where the Kaplan–Meier curves do not drop to 0 in addition to being highly variable in the tail. It should however be noted that this deterioration for higher thresholds already happens for fairly moderate censoring proportions here, in contrast to the results of the previous section where $\tau = 0$. In contrast, when one is able to appropriately model the tails of S^t and S^c with respect to the amount of censoring (i.e., compare results of EVT and EVT benchmark), there is no striking deterioration of the quality of estimation of Δ_τ, as suggested by the results of EVT for increasing values of τ.

4.4 | Varying net benefits Δ_τ

In this section, we consider a slightly different setting as we here focus on the impact of the actual value Δ_τ on the estimation procedures. The rationale for this section is related to the previous section, and is based on two observations; first, higher (absolute) values of the population parameter Δ_τ are related to the survival functions of both groups being more distant from one another. And second, the weights involved in the Peron procedure require to explore the tails of S^c calculated with the observed times of the treatment group, and vice-versa. Hence, one may conjecture that the

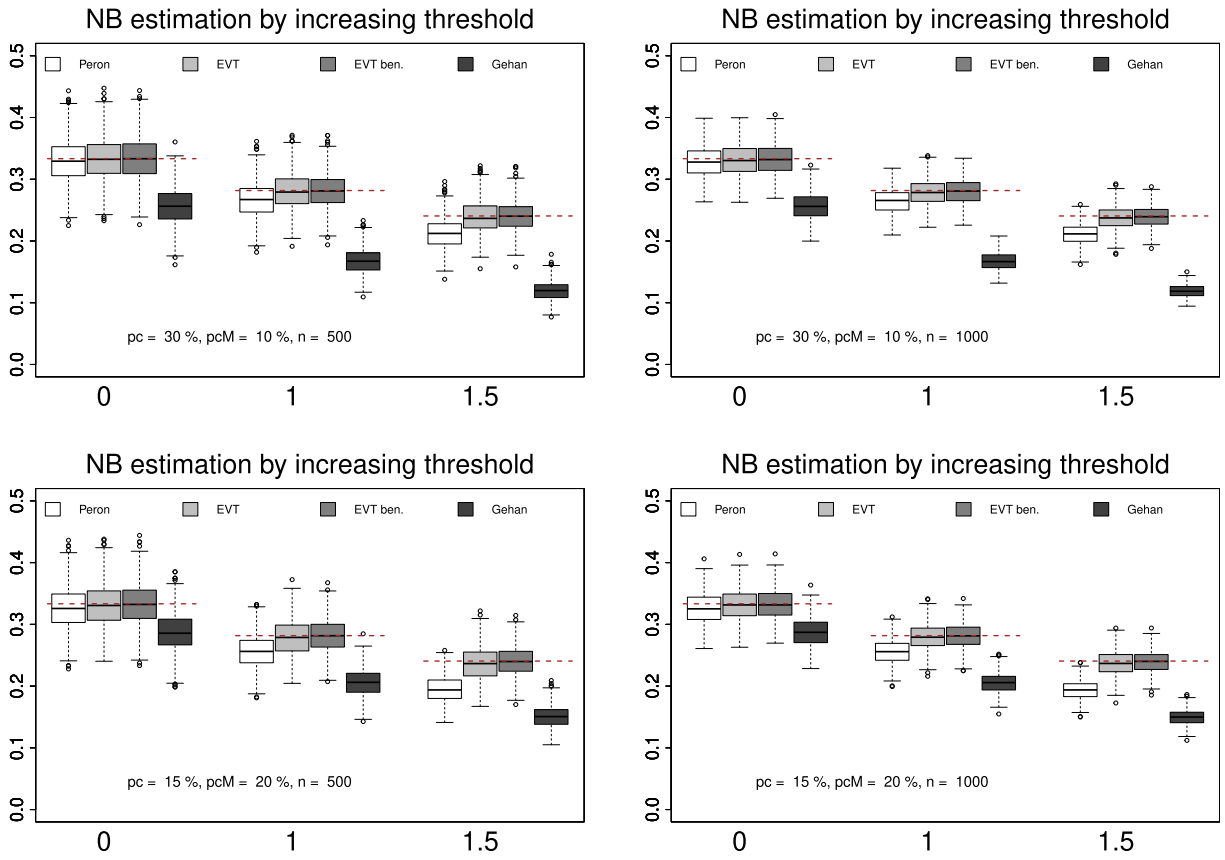

FIGURE 4 Δ_τ estimation for 500 datasets for setting 1 based on $\Delta_0 = 1/3$ with increasing $\tau \in \{0, 1, 1.5\}$, for n^t and $n^c \in \{500, 1000\}$, $p_c \in \{15\%, 30\%\}$ and $p_C^M \in \{10\%, 20\%\}$. The true Δ_τ is in dotted brown.

greater the value of the population parameter Δ_τ, the further we will need to explore the tails of S^c, and thus the worse the impact on the Peron estimator if censoring is such that Kaplan–Meier estimators do not suitably capture the tail behavior. Considering fixed censoring and fixed value of τ, we thus expect that the quality of the Peron estimator deteriorates when Δ_τ increases. We analyze if this phenomenon has a similar impact on the EVT estimator.

We keep the same simulation setting as above, but instead of having a unique censoring distribution for both groups, and hence different censoring proportions for both groups, we now simulate different censoring distributions in order to have similar censoring proportions. If we do no adopt this operation, we will alter the results because increasing Δ_τ implies having higher time-to-event values for the treatment group. This, in turn, implies higher values of M and t_M if we want to keep p_c and p_c^M fixed, which at last implies *less* censoring for the control group. Hence, if we do not pay attention to this, by increasing Δ_τ, we reduce the overall censoring proportion (i.e., in both groups) which of course affects the performance of the estimators, while our objective is here to study the impact of Δ_τ for fixed censoring. So, we simulate two separate distributions that are such that p_c and p_c^M are fixed and the same for both groups, regardless of the value of Δ_τ.

Figure 5 shows the simulation results for values of $\Delta_\tau \in \{0, 1/3, 1/2\}$, with smaller sample sizes than before. The performance of the Peron estimator deteriorates for increasing values of Δ_τ, given that this induces exploring more the tails of the Kaplan–Meier estimators. In contrast, if tail modeling is appropriate using the suggested hybrid estimators for the survival functions, then exploring further the tails has little impact on the quality of estimation of Δ_τ. This is confirmed by the good behavior of the EVT estimator for all values of Δ_τ.

4.5 | Varying time-to-event distributions

To complete our simulation study, we investigate the impact of an alternative distribution for the time-to-event variable. In particular, we consider setting 2 where the distributions of survival times are no longer exponential, but rather

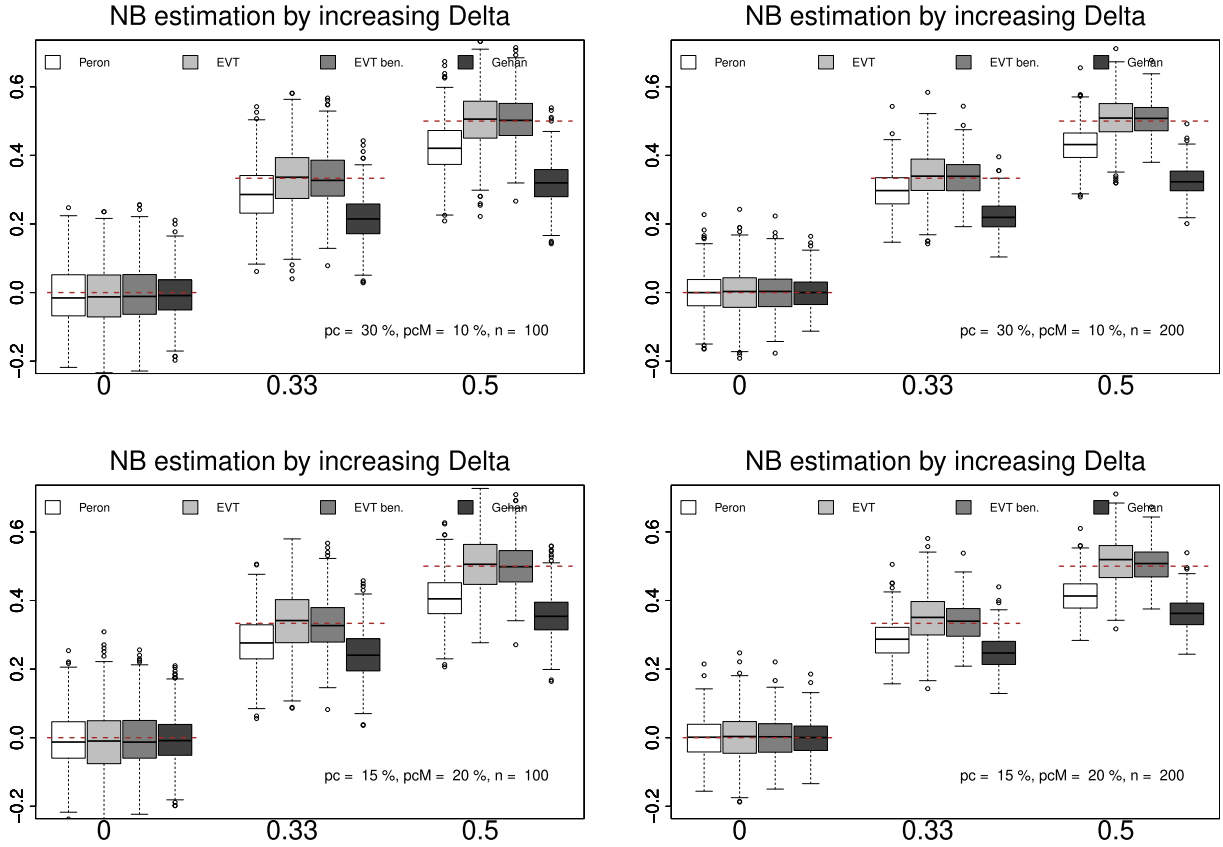

FIGURE 5 Δ_τ estimation for 500 datasets for setting 1 based on $\tau = 0$ with increasing $\Delta_\tau \in \{0, 1/3, 1/2\}$, for n^t and $n^c \in \{100, 200\}$, $p_c \in \{15\%, 30\%\}$ and $p_C^M \in \{10\%, 20\%\}$. The true Δ_τ is in dotted brown.

log-normal with common variance between groups. We then conduct the same simulations as above to study the effects on all estimators of varying censoring proportions, varying values of τ and varying values of Δ_τ. For sake of brevity, we only display here the results for varying thresholds τ. This is depicted in Figure 6, which is in all aspects similar to Figure 4, except for the survival time distributions.

As can be observed, very similar conclusions to the ones holding for Figure 4 can be drawn here: the EVT estimator resists very well to the need of exploring further the tails of the survival functions as compared to the Peron estimator when the threshold τ increases.

5 | ACTUAL RANDOMIZED TRIAL

We consider the data from the CA184-024 trial (ClinicalTrials.gov number, NCT00324155), which was an international anticancer immunotherapy trial having recruited 502 patients with previously untreated metastatic melanoma. These patients were assigned in a 1:1 ratio to a new treatment (dacarbazine plus ipilimumab) or a control treatment (dacarbazine plus placebo). We focus here on overall survival (OS), which was the primary outcome of the study.[27]

An interesting feature of this study was the delayed effect of the experimental treatment versus the control. This is displayed in the left side of Figure 7, where survival curves only start to diverge after several months. This feature of delayed effect becomes increasingly encountered in modern immunotherapies, as argued in ref. [7] where the same trial is analyzed using the Peron procedure. A direct consequence is that the proportional hazards assumption is violated. Hence, traditional methods such as log-rank tests have to be adapted and interpretation of the associated hazard ratios become even more cumbersome for communication with patients and clinicians.[6] In contrast, treatment effects measured through pairwise comparisons still lead to accessible interpretations whether or not hazards are proportional. In addition, the simulation study in ref. [7] suggests that statistical testing based on simple permutation techniques for Δ_τ tends to have higher power than standard log-rank testing in contexts of delayed effects, which makes the use of Δ_τ particularly appealing for analyzing the present data.

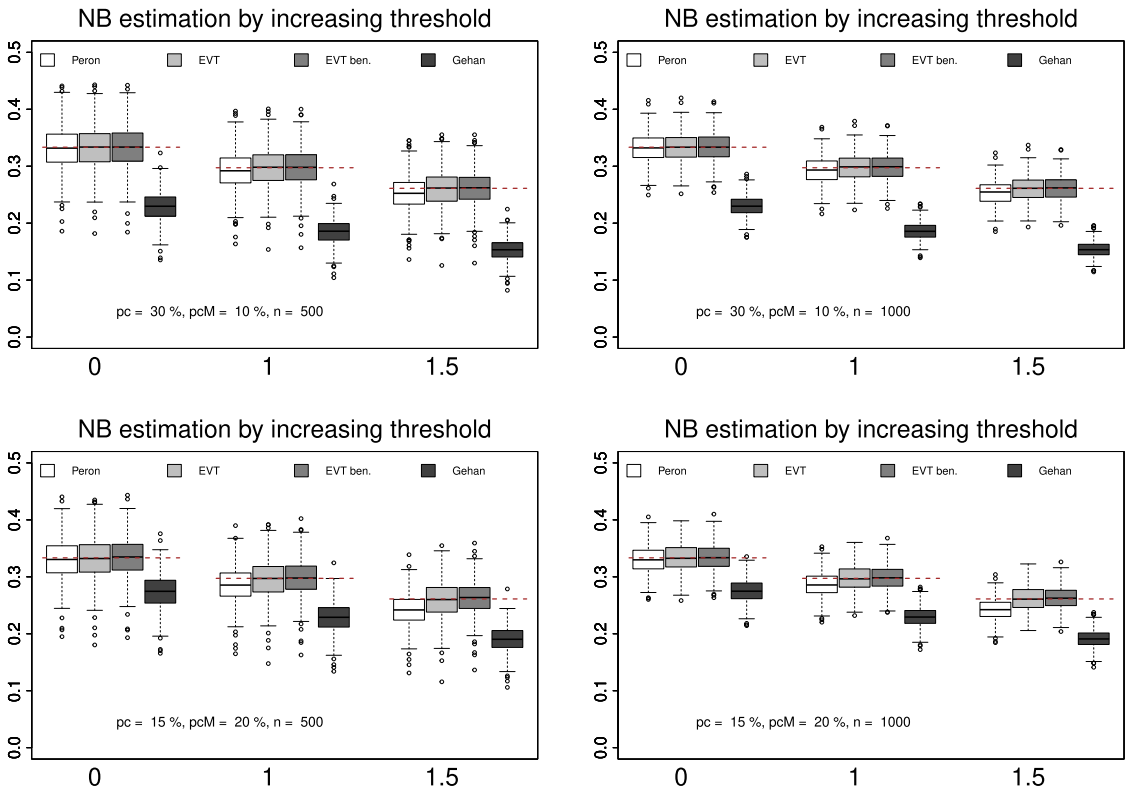

FIGURE 6 Δ_τ estimation for 500 datasets for setting 2 based on $\Delta_0 = 1/3$ with increasing $\tau \in \{0, 1, 1.5\}$, for n^t and $n^c \in \{500, 1000\}$, $p_c \in \{15\%, 30\%\}$ and $p_C^M \in \{10\%, 20\%\}$. The true Δ_τ is in dotted brown.

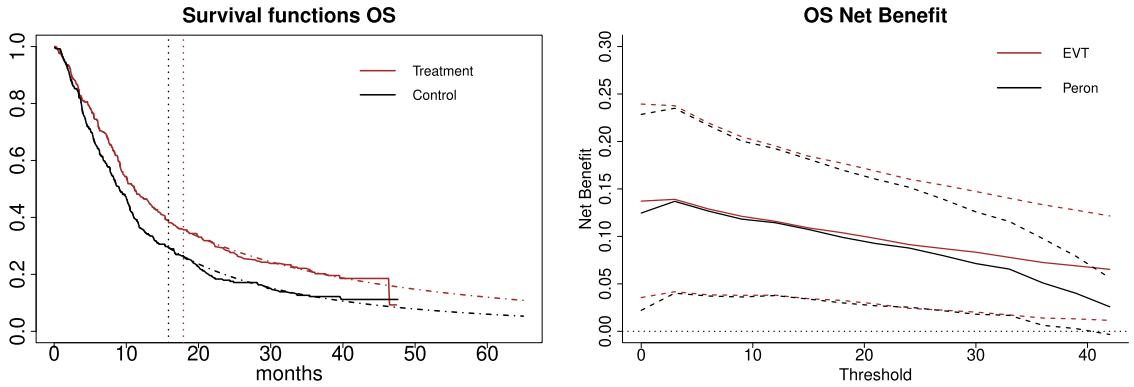

FIGURE 7 Left: survival functions of OS for the CA184-024 trial data. For both groups, the Kaplan-Meier estimator is displayed in full line, while tail extrapolation using the tools of the paper is in dotted line. Vertical dotted lines indicate the threshold above which parametric tail extrapolation is operated. Right: estimated Net Benefits Δ_τ for OS in the CA184-024 trial data. Pointwise 95% two-sided confidence intervals are constructed using ten thousand permuted samples.

The right side of Figure 7 displays the graphs of $\widehat{\Delta}_\tau$ computed for OS and different values of τ ranging from 0 to 42 months. This corresponds to a maximal threshold of clinical relevance that leaves five patients at risk in the control group. In this graph, all pointwise 95% confidence intervals are constructed using ten thousand permutations of the original dataset, and are hence constructed under the null hypothesis $H_0 : \Delta_\tau = 0$. It should be noted that using permutations implicitly amounts to assuming similar censoring distributions in both groups. Of course, one could also consider constructing confidence intervals by bootstrapping samples instead of simply permuting group allocations, leading to arguably more appropriate intervals under the alternative. See for example, Chapter 9 of ref. [28] for an overview of permutation techniques in survival analysis, and ref. [29] on the use of permutation and bootstrap techniques for inference on the Mann–Whitney effect (a linear transformation of the Net Benefit) and the win ratio.

Returning to Figure 7, an interesting feature of the right side is the divergence that starts to appear between EVT and its competitors for higher values of τ, while being in good agreement for smaller values of τ. Since the censoring in the sample is fairly moderate (21.6% for treatment, 13.5% for control) and is such that the end of the Kaplan–Meier curves are arguably not too far away from 0, it is natural to recognize that the EVT and Peron estimators present similar results for small values of τ as appropriate tail modeling is expected to be of moderate impact here. On the other hand, when considering higher values of τ, the weighting scheme of the Peron estimator is such that one needs to explore the tails of the survival curves further and further. At some point, for high values of τ, most of the pairs will thus heavily rely for their classification on appropriate tail modeling and the Peron estimator is expected to be biased towards 0, as more and more uninformative weights are introduced in the procedure. In contrast, no uninformative weight is introduced at all in the EVT estimator, which explains the divergence of curves for $\widehat{\Delta}_\tau$ observed in the graph. An interesting consequence is that, since the resampling technique uses the exact same procedure for every permuted sample as for the original one, the two-sided pointwise confidence intervals of the Peron and Gehan estimators are both driven towards 0 for higher values of τ. As a result, the latter two procedures fail at detecting a statistical significant treatment effect at a 5% level for the highest values of τ. This again suggests that the EVT estimator is more able to detect long-term and delayed effects, as suggested by the elevated and sustained values of $\widehat{\Delta}_\tau$ for the EVT estimator as a function of τ. Next, in order to complete our analysis, we also consider an artificial situation where we increase the proportion of administrative censoring for OS by introducing a limit of 39 months above which every observation will now be considered censored as well. This is intended to roughly mimic the situation of an analysis of early data that would have been conducted, for example, for an interim analysis of the trial. The resulting Kaplan–Meier curves for both treatment arms are depicted on the left-side of Figure 8.

We conduct a similar study for the evolution of Δ_τ as a function of τ. In particular, we consider here varying values of τ ranging from 0 to 33 months, and compare again the Peron and EVT estimators of Δ_τ based on this artificial dataset. This is depicted on the right-side of Figure 8, where we also include the curves for $\widehat{\Delta}_\tau$ based on the original dataset. The most interesting feature of the graph is that, contrary to the Peron estimator, the EVT estimators based on either the original or the artificial dataset are extremely similar for every value of τ. This is a natural consequence of providing a tail modeling of survival curves that is quite similar between the original and artificial datasets, provided the proportion of administrative censoring is reasonable. In contrast, the Peron procedure evidently suffers from the

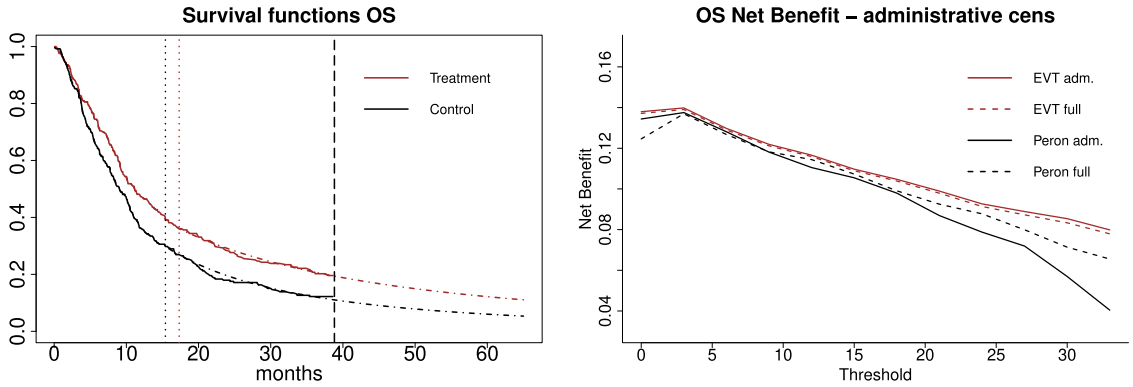

FIGURE 8 OS data for the CA184-024 trial with artificial censoring at 39 months. Left: for both groups, the Kaplan-Meier estimator is displayed in full line, while tail extrapolation using the tools of the paper is in dotted line. Right: comparison of estimated Δ_τ for EVT and Péron based on either the original or the artificial dataset that includes more administrative censoring.

lack of information in the Kaplan–Meier curves with the artificial dataset, driving the resulting estimator even more towards 0 than with the original dataset.

6 | DISCUSSION

Methods based on pairwise comparison form an attractive toolbox for the analysis of clinical trial data, as they allow for the handling of several outcomes of interest, possibly including thresholds of clinical relevance, and still manage to keep an interpretation that is arguably better suited for communication with patients than traditional statistical methodologies. These advantages come at the cost of several challenges, among which the handling of censored outcomes is most typical in practice. Our work aims at tackling this challenge, building on the recent suggestion of ref. [3] and improving on the latter by providing a suitable tail estimation for the survival functions involved in the methodology. This improvement is required when the censoring mechanism in the data is such that the tails of survival functions are no longer reliably estimated using a Kaplan–Meier estimator. In such situations, the Peron estimator is affected by a bias for the treatment effect estimator towards the value indicating no treatment effect. Additionally, the Peron methodology will relies on the concept of 'uninformative' weights, which is more of an artifact of the estimator rather than a population concept in the context of pairwise comparisons. Besides the univariate setting, this uninformative quantity is further bothersome when including additional outcomes in the pairwise comparison construction.

In order to avoid these two issues, we suggested to use cornerstone results from extreme value theory that allow to parametrically model the tails of the survival functions of each treatment group separately, and where the parametric choice is specifically designed for tail modeling. A crucial aspect of this strategy is that, by modeling tails for each treatment arm, no assumption is made about the joint behavior of the latter. In that sense, our suggestion preserves one of the strengths of pairwise comparisons methodologies when dealing with censoring. Indeed, the latter are often praised to be free of assumptions concerning the joint behavior of survival functions, which contrasts particularly with the hazard ratio relying crucially on the PH assumption. Hence, we improve on the pitfalls of the Gehan and Peron estimators, and most importantly this improvement is the product of a set of assumptions that are arguably the least restrictive one could reasonably invoke.

As a result, when tail modeling is appropriately performed, the EVT estimator improves over the Peron estimator in all situations where tail exploration is influential. This was confirmed in our simulation study as well as in the analysis of the CA184-024 trial. An important highlight of our simulations is the stability of the EVT estimator with respect to increasing values of the threshold τ when tail modeling is appropriate. As sensitivity analysis of Δ_τ with respect to τ are often highlighted as one of the main strengths of the definition of Δ_τ compared to alternative measures of treatment effect, this represents a considerable improvement of the EVT estimator over the Peron estimator. In a similar fashion, in the CA184-024 trial it was also illustrated that interim results obtained during the study may be expected to vary less over time with the inclusion of supplementary information, provided these interim results do not happen too early on during the study. As a consequence, one may speculate that the EVT estimator may also help in detecting earlier on during the study statistically significant long-term treatment effects than the Peron or Gehan estimators. Of course, the

generalization of these conclusions is conditional to the appropriate use of the tail extrapolation methods employed in this paper.

A natural extension of the results of this paper is the consideration of multiple prioritized outcomes. In this case, pairs that are not completely classified on a first outcome that serves as highest priority, are classified on the next prioritized outcome with a weight that is given by the (estimated) probability to be neutral on the first outcome. The comparative results between our procedure and the Peron and Gehan estimator will naturally extend to this case: bias issues will spread throughout the prioritization rule for the latter estimators, while our procedure will be able to avoid such issues in all settings where survival tail modeling is appropriate.

Nevertheless, as for any procedure, the range of applicability of our methodology is in practice constrained by the appropriateness of our underlying assumptions. In particular, we suggest here to improve on tail modeling of survival functions through the use of extreme value theory, as this alleviates the choice of a specific family of distributions for the tails of survival functions. This however also implies that we consider the latter objective to be realistic with respect to the censoring mechanism in the data. If, on the contrary, the dataset is contaminated by too much censoring, in particular in the tail, then the procedure may no longer be appropriate in the sense that one will attempt to employ tools of extreme value theory based on the complete observations that are, essentially, not extreme observations. A direct consequence is that one will rely on an asymptotic result for parametric modeling, that is, $u \to \infty$ in the Pickands-Balkema-de Haan theorem, while disposing of too little information in the data to let this threshold be high enough. Eventually, this may drive the pairwise comparison estimator to be biased, through a mechanism that is very much similar to the one occurring when choosing the 'wrong' parametric distribution in parametric statistics. In practice thus, a first basic step before embracing our methodology is the visual inspection of both Kaplan–Meier curves in order to evaluate how appropriate a parametric tail modeling would be for the 'extreme event' region. Of note, the restriction induced by too heavy censoring is not unique to our methodology, as it is also applicable to the procedures of Peron and Gehan. In such situations, one should instead turn to alternative modeling assumptions that are however out of the scope of this paper.

A related restriction for our methodology concerns the underlying assumption of regularity of the survival functions. By aiming to extrapolate on the tails of the latter, we indeed assume the information in the sample suffices for the latter task. This essentially amounts to assuming that the survival functions are sufficiently regular in the tails with respect to the region of their support containing the observations we consider to be extreme. If, on the contrary, the tails present some unexpected behavior with respect to the available information, which could for instance occur with particular mixture of distributions, one should not expect to appropriately capture this unforeseen behavior with the approach described in this paper.

To further extend the underlying assumptions of our context, we point out that, by attempting to complete survival functions towards 0, we implicitly assume all individuals are essentially subject to the event of interest. In practice, one should therefore be careful to distinguish situations where observing a Kaplan–Meier estimator that is not dropping towards 0 is either only an implication of classical censoring, or is instead an indication that not all subjects are in fact susceptible of experiencing the event of interest (e.g., in cure rate models). The tools suggested in this paper are to be considered only valid for the former situation, that is, when no cured observations are present at all. Extensions of pairwise comparisons to the latter situation are under current research but beyond the scope of this paper. Once the objective of tail modeling seems attainable and justified in practice, two choices remain to be made for employing our procedure; first, we need to define the thresholds u^t and u^c above which survival times will be considered extreme, and second we need to define an estimation procedure for the parameters of the GPD. Extensive simulations seem to illustrate the fact that these choices can be quite generic in our setting, as results were stable with respect to different investigated options. This comes as a consequence of the fact that appropriate tail modeling of survival functions are, in our context, building blocks in a broader pairwise comparison estimation procedure. Hence, fine-tuning of these choices may not be as crucial as in other contexts (e.g., for mean residual lifetime estimation). In practice thus, relying on maximum likelihood estimation for the GPD parameters and defining u^t and u^c as 80% quantiles of fully observed times is a simple option to start using our procedure. Of course, potential improvements in the literature on extreme value theory with censored observations may further spread to our applied context as well.

FUNDING INFORMATION

This research was partially funded by the regions of Wallonia (BioWin Consortium Agreement No 7979).

CONFLICT OF INTEREST

None.

DATA AVAILABILITY STATEMENT

Data subject to third party restrictions: the data of the actual clinical trial are property of Bristol-Myers Squibb, and are hence only potentially available from Alexandre Lambert with the permission of Bristol-Myers Squibb.

ORCID

Mickaël De Backer https://orcid.org/0000-0003-1669-6391

REFERENCES

1. Buyse M. Generalized pairwise comparisons of prioritized outcomes in the two-sample problem. *Stat Med*. 2010;29(30):3245-3257.
2. Pocock SJ, Ariti CA, Collier TJ, Wang D. The win ratio: a new approach to the analysis of composite endpoints in clinical trials based on clinical priorities. *Eur Heart J*. 2012;33(2):176-182.
3. Péron J, Buyse M, Ozenne B, Roche L, Roy P. An extension of generalized pairwise comparisons for prioritized outcomes in the presence of censoring. *Stat Methods Med Res*. 2018;27(4):1230-1239.
4. Buyse M. Reformulating the hazard ratio to enhance communication with clinical investigators. *Clin Trials*. 2008;5(6):641-642.
5. Oakes D. On the win-ratio statistic in clinical trials with multiple types of event. *Biometrika*. 2016;103(3):742-745.
6. Buyse M, Saad ED, Burzykowski T, Péron J. Assessing treatment benefit in immunooncology. *Stat Biosci*. 2020;12:83-103.
7. Péron J, Lambert A, Munier S, et al. Assessing long-term survival benefits of immune checkpoint inhibitors using the net survival benefit. *J Natl Cancer Inst*. 2019;111:1186-1191.
8. Gehan EA. A generalized Wilcoxon test for comparing arbitrarily singly-censored samples. *Biometrika*. 1965;52(1–2):203-224.
9. Efron B. The two-sample problem with censored data. *Proc Fifth Berkeley Symp Math Stat*. 1967;IV:831-853.
10. Ramchandani R, Schoenfeld DA, Finkelstein DM. Global rank tests for multiple, possibly censored, outcomes. *Biometrics*. 2016;72(3):926-935.
11. Moeschberger ML, Klein JP. A comparison of several methods of estimating the survival function when there is extreme right censoring. *Biometrics*. 1985;41(1):253-259.
12. Klein JP, Lee S-C, Moeschberger ML. A partially parametric estimator of survival in the presence of randomly censored data. *Biometrics*. 1990;46:795-811.
13. Gelber RD, Goldhirsch A, Cole BF. Parametric extrapolation of survival estimates with applications to quality of life evaluation of treatments. *Control Clin Trials*. 1993;14(6):485-499.
14. Su Z, Fang L. A novel method to calculate mean survival time for time-to-event data. *Commun Stat Simulation Comput*. 2012;41(5):611-620.
15. Gong Q, Fang L. Asymptotic properties of mean survival estimate based on the Kaplan- Meier curve with an extrapolated tail. *Pharm Stat*. 2012;11(2):135-140.
16. Grama I, Tricot JM, Petiot JF. Estimation of the extreme survival probabilities from censored data. *Buletinul Academiei de Stiinte a Republicii Moldova - Mathematica*. 2014;74(1):33-62.
17. Alvarez-Iglesias A, Newell J, Scarrott C, Hinde J. Summarising censored survival data using the mean residual life function. *Stat Med*. 2015;34(11):1965-1976.
18. Beirlant J, Goegebeur Y, Segers J, Teugels JL. *Statistics of Extremes: Theory and Applications*. John Wiley & Sons; 2006.
19. Balkema AA, de Haan L. Residual life time at great age. *Ann Probability*. 1974;2:792-804.
20. Pickands J. Statistical inference using extreme order statistics. *Ann Stat*. 1975;3(1):119-131.
21. de Haan L. Slow variation and characterization of domains of attraction. *Statistical Extremes and Applications*. Springer; 1984:31-48.
22. Scarrott Cl, MacDonald A. A review of extreme value threshold es-timation and uncertainty quantification. *REVSTAT-Stat J*. 2012;10(1):33-60.
23. Alvarez-Iglesias A. Extensions and Applications of Survival Trees in Medical Data. PhD thesis, School of Mathematics, Statistics and Applied Mathematics, National University of Ireland, 2012.
24. Beirlant J, Guillou A, Dierckx G, Fils-Villetard A. Estimation of the extreme value index and extreme quantiles under random censoring. *Extremes*. 2007;10(3):151-174.
25. Einmahl JHJ, Fils-Villetard A, Guillou A. Statistics of extremes under random censoring. *Ther Ber*. 2008;14(1):207-227.
26. Beirlant J, Guillou A, Toulemonde G. Peaks-over-threshold modeling under random censoring. *Commun Stat Theory Methods*. 2010;39(7):1158-1179.
27. Robert C, Thomas L, Bondarenko I, et al. Ipilimumab plus dacarbazine for previously untreated metastatic melanoma. *New England J Med*. 2011;364(26):2517-2526.
28. Pesarin F, Salmaso L. *Permutation Tests for Complex Data: Theory, Applications and Software*. Wiley; 2010.
29. Dobler D, Pauly M. Bootstrap-and permutation-based inference for the Mann-Whitney effect for right-censored and tied data. *Test*. 2018;27(3):639-658.

How to cite this article: De Backer M, Legrand C, Péron J, Lambert A, Buyse M. On the use of extreme value tail modeling for generalized pairwise comparisons with censored outcomes. *Pharmaceutical Statistics*. 2023;22(2):284-299. doi:10.1002/pst.2271