Statistics and Computing (2024) 34:101
https://doi.org/10.1007/s11222-024-10420-w

ORIGINAL PAPER

Resampling-based confidence intervals and bands for the average treatment effect in observational studies with competing risks

Jasmin Rühl[1] · Sarah Friedrich[1,2]

Received: 26 September 2023 / Accepted: 29 February 2024

Abstract
The g-formula can be used to estimate the treatment effect while accounting for confounding bias in observational studies. With regard to time-to-event endpoints, possibly subject to competing risks, the construction of valid pointwise confidence intervals and time-simultaneous confidence bands for the causal risk difference is complicated, however. A convenient solution is to approximate the asymptotic distribution of the corresponding stochastic process by means of resampling approaches. In this paper, we consider three different resampling methods, namely the classical nonparametric bootstrap, the influence function equipped with a resampling approach as well as a martingale-based bootstrap version, the so-called wild bootstrap. For the latter, three sub-versions based on differing distributions of the underlying random multipliers are examined. We set up a simulation study to compare the accuracy of the different techniques, which reveals that the wild bootstrap should in general be preferred if the sample size is moderate and sufficient data on the event of interest have been accrued. For illustration, the resampling methods are further applied to data on the long-term survival in patients with early-stage Hodgkin's disease.

Keywords Average treatment effect · Bootstrap · Confidence interval · G-formula · Time-to-event data

1 Introduction

Causal inference provides tools to compare treatment strategies in studies that do not permit random allocation of subjects to therapy groups, e.g., for ethical reasons or simply because it is not feasible. Special analysis methods are necessary because in non-randomized trials, risk factors are likely to be distributed unequally across treatment groups and as a consequence, side-by-side comparisons will lead to biased estimation of the direct treatment effect (Yang et al. 2010; Nørgaard et al. 2017). Randomized trials benefit from causal analysis tools, too, for instance when dealing with non-compliance or selection bias. In this manuscript, we focus on the control of confounding bias. The idea of the counterfactual approach to causal inference is to model the mean outcome in a hypothetical world where all participants of the study are exposed to the same intervention—possibly 'counter to the fact', i.e., contrary to the treatment they actually received. Causal conclusions can then be drawn by contrasting the obtained estimates for the treatment levels of interest (Rubin 1974); (Hernán and Robins 2020 Sect. I.1).
In case of time-to-event endpoints, statisticians need to take additional difficulties into account, however, as the analysis of right-censored data requires particular techniques. The hazard ratio, which is the common measure of the treatment effect for time-to-event data, comes along with several issues when the aim is to draw causal inferences: In the first place, it is non-collapsible. Thus, the causal effect estimate in the entire population may differ fundamentally from the average of the causal effect estimates across subgroups, even if the variable defining these subgroups is no confounder (Martinussen and Vansteelandt 2013; FDA 2023). Another drawback is selection bias, which has e.g., been described by Aalen et al. (2015). Selection bias occurs because the hazard function only takes survivors into account, but if treatment does indeed affect survival, the distribution of the risk factors will deviate between survivors in the two treatment groups as time progresses. Apart from that, the hazard ratio—as a single value—fails to convey potentially time-varying effects

✉ Jasmin Rühl
jasmin.ruehl@math.uni-augsburg.de

Sarah Friedrich
sarah.friedrich@math.uni-augsburg.de

[1] Mathematical Statistics and Artificial Intelligence in Medicine, University of Augsburg, Augsburg, Germany

[2] Centre for Advanced Analytics and Predictive Sciences (CAAPS), University of Augsburg, Augsburg, Germany

Published online: 21 March 2024

and also depends on the duration of the study (Hernán 2010). We therefore consider the risk difference as effect measure instead. Our target estimand is based on the cumulative incidence function, which quantifies the risk of experiencing a specific event type out of one or more possible causes until a given time point. This way, a competing risks framework is accommodated on top, which covers the standard survival setting as a special case. Examples of observational studies that compare treatment effects using the cumulative incidence function include Philipps et al. (2020); Butt et al. (2021); Chauhan et al. (2022).

Beside the estimated average treatment effect, researchers are often also interested in further statistical inference. The stochastic process associated with the estimated cumulative incidence function is rather complex, making it difficult to derive exact confidence intervals and bands, though. A commonly applied remedy is the classical nonparametric bootstrap proposed by Efron (1981) (cf. Neumann and Billionnet 2016; Stensrud et al. 2020; Stensrud et al. 2016), even though this resampling method is not optimal in several situations, e.g., when dealing with dependent data (Singh 1981; Friedrich et al. 2017). Ozenne et al. (2020) presented an alternative approach based on the influence function, and as counting processes are inherent to time-to-event analysis, resampling methods relying on martingale theory further suggest themselves.

In this paper, we illustrate that apart from the method proposed by Ozenne et al. (2020), the classical bootstrap as well as the martingale-based wild bootstrap also accurately approximate the distribution of the stochastic process at hand. We compare the performance of these resampling approaches in terms of the resulting confidence intervals and bands by means of simulations as well as an applied data example recording the long-term outcomes of early-stage Hodgkin's disease patients.

The remainder of this manuscript is organized as follows: Sect. 2 establishes the setting and notation as well as the causal estimator for the average treatment effect. In Sect. 3, we introduce the three mentioned resampling approaches. The simulation study and the analysis of the Hodgkin's disease data are presented in Sects. 4 and 5. Finally, the paper concludes with a discussion.

2 Average treatment effect for right-censored data with competing risks

We consider a competing risks setting with K failure types. Let the absolutely continuous random variables T and C denote an individual's event and censoring time, respectively. The observed data include $T \wedge C$, the minimum of T and C, as well as an indicator $D \in \{0, 1, \ldots, K\}$, which represents the type of failure. W.l.o.g., let $D = 1$ imply that a subject experienced the event of interest. If $D = 0$, the event time is censored, i.e., $C < T$. Besides, we observe a binary treatment indicator A and a bounded, p-dimensional vector $\boldsymbol{Z}$ of baseline covariates. Throughout this paper, suppose that the data sample $\{(T_i \wedge C_i, D_i, A_i, \boldsymbol{Z_i})\}_{i\in\{1,\ldots,n\}}$ is independent and identically distributed (i.i.d.), and does not include any tied event times. It is further assumed that T_i and C_i are conditionally independent given $(A_i, \boldsymbol{Z_i})$.

In the presence of competing events, one may be interested in either the direct or the total effect of treatment on the event of interest (Young et al. 2020). The direct effect reflects the impact of the studied therapy in a hypothetical setting where all competing events have been eliminated, whereas the total effect additionally takes the impact of the therapy mediated by competing events into account. Neither of these characterizations is generally preferable over the other: While the direct effect may help to better understand the mechanisms by which the treatment affects the outcome, interventions that eradicate competing events are rare, and thus, the total effect is typically more relevant in practice. We will focus on the estimation of total effects hereafter.

For a fixed time point t within the study time interval $[0, \tau]$, we define the average treatment effect of interest in the entire population as $ATE(t) = \mathbb{E}\left(F_1^1(t) - F_1^0(t)\right)$. The expression $F_1^a(t) = P(T^a \leq t,\ D^a = 1)$ refers to the potential cumulative incidence function for cause 1 under treatment $a \in \{0, 1\}$, applying the counterfactual notation as in Hernán and Robins (2020). Accordingly, $F_1^a(t)$ describes the probability of observing the event of interest until time t, had all study participants received treatment a.

In order to ensure identifiability of ATE, the subsequent assumptions need to be fulfilled (see e.g., Hernán and Robins 2020, Sect. I.3 for a thorough description): Conditional exchangeability holds if there are no unmeasured confounders. For given covariate values, the risk among the treated subjects is thereby equal to the risk among the untreated subjects, had they been treated, and vice versa. A formal definition of conditional exchangeability requires independence between $\mathbb{1}\{T^a \leq \tau,\ D^a = 1\}$ and A, conditional on $\boldsymbol{Z}$, for $a \in \{0, 1\}$. (We use $\mathbb{1}\{\cdot\}$ here and in the following to denote the indicator function.) Furthermore, the positivity assumption applies if the conditional treatment probability $P(A = 1 \mid \boldsymbol{z})$ is bounded away from 0 and 1 for covariate values $\boldsymbol{z}$ on the support of $f_{\boldsymbol{Z}}(\boldsymbol{z})$, so that both therapies $A = 0$ and $A = 1$ are possible. Lastly, the interventions $A = 0$ and $A = 1$ need to be well-defined, with $\mathbb{1}\{T \leq \tau,\ D = 1\} = \mathbb{1}\{T^A \leq \tau, D^A = 1\}$. This condition is referred to as consistency, and it ensures that the observed and potential risks are equal if the actual and counterfactual therapy coincide.

Assuming that exchangeability, positivity and consistency apply and there is no interference between the potential outcomes of distinct individuals, the g-formula yields an esti-

mate of the average treatment effect (Ozenne et al. 2020):

$$\widehat{ATE}(t) = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{F}_1(t \mid A=1, \boldsymbol{Z_i}) - \hat{F}_1(t \mid A=0, \boldsymbol{Z_i})\right).$$

Here, any assumptions made when modelling $\hat{F}_1$ need to be fulfilled in order to obtain a meaningful estimator. Despite the issues pointed out by Aalen et al. (2015), it is reasonable to derive the cumulative incidence function—and hence $\widehat{ATE}$—from hazard rates; the key point is that the causal interpretation of the effect estimate relies on $\hat{F}_1$. Let $\hat{\Lambda}_k(t \mid a, z)$, $k \in \{1, \ldots, K\}$ be the estimator of the cause-specific, conditional cumulative hazard, and define

$$\hat{F}_1(t \mid a, z) = \int_0^t \exp\left(-\sum_{k=1}^{K}\hat{\Lambda}_k(s \mid a, z)\right) \mathrm{d}\hat{\Lambda}_1(s \mid a, z),$$

in line with the characterization proposed by Benichou and Gail (1990). One possibility to obtain $\hat{\Lambda}_k(t \mid a, z)$ is to fit a cause-k specific Cox model with covariates A and $\boldsymbol{Z}$, i.e.,

$$\hat{\Lambda}_k(t \mid a, z) = \hat{\Lambda}_{0k}(t)\exp(\hat{\beta}_{kA}a + \hat{\boldsymbol{\beta}}_{\boldsymbol{kZ}}^T z),$$

with $\hat{\boldsymbol{\beta}}_{\boldsymbol{k}} = (\hat{\beta}_{kA}, \hat{\boldsymbol{\beta}}_{\boldsymbol{kZ}}^T)^T$ representing the estimated vector of regression coefficients. The covariates may in fact vary for the individual causes, since different event types are possibly associated with distinct risk factors—provided that A is included in the model for the cause of interest. The Breslow estimator eventually yields the following approximation of the cumulative baseline hazard (Breslow 1972):

$$\hat{\Lambda}_{0k}(t) = \int_0^t \frac{\mathrm{d}N_k(s)}{\sum_{i=1}^{n} Y_i(s)\exp(\hat{\beta}_{kA}A_i + \hat{\boldsymbol{\beta}}_{\boldsymbol{kZ}}^T \boldsymbol{Z_i})}.$$

We define the counting process $N_k(t)$ as $\sum_{i=1}^{n} N_{ki}(t)$ with $N_{ki}(t) = \mathbb{1}\{T_i \wedge C_i \leq t,\ D_i = k\}$, such that $\mathrm{d}N_k(t)$ represents the increment of $N_k(t)$ over the infinitesimal time interval $[t, t+dt)$. The at-risk indicator $Y_i(t) = \mathbb{1}\{T_i \wedge C_i \geq t\}$ further specifies whether subject i is part of the risk set just prior to time t.

3 Confidence intervals and bands

Pointwise confidence intervals and time-simultaneous confidence bands are routinely reported in clinical trials as they help to assess the (un)certainty of an estimate. In a series of studies with underlying average treatment effect ATE, it is expected that $(1-\alpha)\cdot 100\%$ of the confidence intervals for $ATE(t)$ at level $(1-\alpha)$ include the true average treatment effect at a given time t. Confidence bands extend this concept to time intervals, meaning that $(1-\alpha)\cdot 100\%$ of the confidence bands for ATE at level $(1-\alpha)$ will cover the true average treatment effect over the entire interval of interest. It is not straightforward to define such confidence regions for ATE, however, due to the complexity of the stochastic process $U_n(t) = \sqrt{n}\,(\widehat{ATE}(t) - ATE(t))$. As a workaround, we aim to approximate the limiting distribution of U_n by means of different resampling approaches.

3.1 Efron's bootstrap

The most common way to derive confidence intervals for ATE is the use of the classical nonparametric bootstrap (Efron 1981), which does not require knowledge of the true underlying distribution. By repeatedly drawing with replacement from the data and calculating a statistical functional of interest in each of the drawn samples, one tries to approach the distribution of the functional in the target population. In the given context, we obtain the estimates $\{\widehat{ATE}_b^*(t)\}_{b\in\{1,\ldots,B\}}$ from B bootstrap samples of the original data, each having size n. An asymptotic confidence interval at level $(1-\alpha)$ can, for instance, be determined by setting the empirical $\frac{\alpha}{2}$ and $(1-\frac{\alpha}{2})$ quantiles of the bootstrap estimates as limits. Furthermore, we construct an asymptotic simultaneous confidence band over the time interval $[t_1, t_2]$ as

$$\left[\widehat{ATE}(t) - q_{1-\alpha}^{EB}\sqrt{\hat{v}^{EB}(t)},\ \widehat{ATE}(t) + q_{1-\alpha}^{EB}\sqrt{\hat{v}^{EB}(t)}\right],$$

with $\hat{v}^{EB}(t)$ referring to the empirical variance of the bootstrap estimates and $q_{1-\alpha}^{EB}$ denoting the $(1-\alpha)$ quantile of

$$\left\{\sup_{t\in[t_1,t_2]}\left|\frac{\widehat{ATE}_b^*(t) - \frac{1}{B}\sum_{\tilde{b}=1}^{B}\widehat{ATE}_{\tilde{b}}^*(t)}{\sqrt{\hat{v}^{EB}(t)}}\right|\right\}_{b\in\{1,\ldots,B\}}.$$

Note that the absolute value is considered here and in the following to increase the stability of the empirical quantiles. The classical bootstrap yields asymptotically correct results in many less intricate settings (as long as the considered data are i.i.d.), and its theoretical validity in the given context is proven by Rühl and Friedrich (2023) based on martingale arguments. While the implementation of Efron's bootstrap is rather simple, the computation time can become excessive with large sample sizes and multiple bootstrap iterations, though.

3.2 Influence function

Another method to obtain confidence intervals for ATE has been described by Ozenne et al. (2020). Supposing that the underlying model is correct, the functional delta method yields an approximation of the asymptotic distribution of

U_n at a given time point w.r.t. the influence function of the average treatment effect. More specifically,

$$U_n(t) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} IF(t;\ T_i \wedge C_i, D_i, A_i, \boldsymbol{Z_i}) + o_P(1)$$
$$\xrightarrow{\mathscr{D}} \mathcal{N}\left(0, \int (IF(t;\ s,d,a,z))^2 \, \mathrm{d}P(s,d,a,z)\right),$$

as n tends to infinity. Here, $P(t, d, a, z)$ denotes the joint probability distribution of the data $(T \wedge C, D, A, \boldsymbol{Z})$. The definition of the influence function IF according to Ozenne et al. (2020, 2017) can be found in Sect. 1.1 of the supplementary material. Besides, we use $\mathcal{N}$ throughout this paper to symbolize the normal distribution. It follows that the plug-in estimator $\hat{v}^{IF}(t) = \frac{1}{n}\sum_{i=1}^{n} (\widehat{IF}(t;\ T_i \wedge C_i, D_i, A_i, \boldsymbol{Z_i}))^2$ is consistent for the asymptotic variance of $U_n(t)$ and thus, asymptotic confidence intervals are easy to calculate without the need of resampling. The construction of confidence bands, on the other hand, is more involved. This is because the dependence between the increments of the process U_n must be taken into account when making inferences concerning multiple time points. It can be shown that U_n converges weakly to a zero-mean Gaussian process on the Skorokhod space $\mathcal{D}[0, \tau]$ (Rühl and Friedrich 2023), and thus, we can derive an asymptotic $(1-\alpha)$ confidence band for ATE over the interval $[t_1, t_2]$ in line with the resampling approach described by Scheike and Zhang (2008):

$$\left[\widehat{ATE}(t) - q^{IF}_{1-\alpha}\sqrt{\hat{v}^{IF}(t)},\ \widehat{ATE}(t) + q^{IF}_{1-\alpha}\sqrt{\hat{v}^{IF}(t)}\right].$$

Here, $q^{IF}_{1-\alpha}$ denotes the $(1-\alpha)$ quantile of

$$\left\{\sup_{t\in[t_1,t_2]}\left|\sum_{i=1}^{n} \frac{\widehat{IF}(t;\ T_i \wedge C_i, D_i, A_i, \boldsymbol{Z_i})}{\sqrt{\hat{v}^{IF}(t)}} \cdot G_i^{IF;(b)}\right|\right\}_{b\in\{1,\ldots,B\}},$$

for B independent standard normal vectors $\{(G_1^{IF;(b)}, \ldots, G_n^{IF;(b)})^T\}_{b\in\{1,\ldots,B\}}$.
As compared to the classical bootstrap, the influence function approach significantly reduces the computation time, considering that the resampling step builds upon repeated generation of random variables rather than the recalculation of functionals based on various individual data sets.

3.3 Wild bootstrap

A third resampling method arises from the fact that the limiting distribution of U_n may be represented in terms of martingales: It can be shown that

$$U_n(t) = \sum_{k=1}^{K}\sum_{i=1}^{n} \left(\int_0^t H_{k1i}(s,t)\,\mathrm{d}M_{ki}(s) + \int_0^\tau H_{k2i}(s,t)\,\mathrm{d}M_{ki}(s)\right) + o_p(1),$$

for functions H_{k1i} and H_{k2i} as defined in Sect. 1.2 of the supplementary material and $M_{ki}(t) = N_{ki}(t) - \int_0^t Y_i(s)\,\mathrm{d}\Lambda_k(s \mid A_i, \boldsymbol{Z_i}), k \in \{1,\ldots,K\}, i \in \{1,\ldots,n\}$ (Rühl and Friedrich 2023). Note that M_{ki} is a martingale relative to the history $(\mathscr{F}_t)_{t\geq 0}$ that is generated by the data observed until a given time, i.e., $\mathbb{E}(\mathrm{d}M_{ki}(t) \mid \mathscr{F}_{t-}) = 0$ and

$$\mathrm{Var}(\mathrm{d}M_{ki}(t) \mid \mathscr{F}_{t-}) = Y_i(t)\,\mathrm{d}\Lambda_k(t \mid A_i, \boldsymbol{Z_i}).$$

Provided that Aalen's multiplicative intensity model (Aalen 1978) applies, the characterization of the variance equals the conditional expectation of $\mathrm{d}N_{ki}(t)$ given the past $\mathscr{F}_{t-}$. This motivates the general idea of the wild bootstrap: By replacing $\mathrm{d}M_{ki}(t)$ with the product of $\mathrm{d}N_{ki}(t)$ and suitable random multipliers $G_i^{WB}, k \in \{1,\ldots,K\}, i \in \{1,\ldots,n\}$, we can approximate the asymptotic distribution of U_n. The initial method described by Lin et al. (1993) only covers standard normal multipliers, but was later extended to more general resampling schemes (cf. Beyersmann et al. 2013; Dobler et al. 2017). In Rühl and Friedrich (2023), we followed ideas of Cheng et al. (1998); Beyersmann et al. (2013) and Dobler et al. (2017) to formally prove that, conditional on the data, the wild bootstrap estimator of U_n,

$$\hat{U}_n(t) = \sum_{k=1}^{K}\sum_{i=1}^{n} \left(\hat{H}_{k1i}(T_i \wedge C_i, t)\, N_{ki}(t) G_i^{WB} + \hat{H}_{k2i}(T_i \wedge C_i, t)\, N_{ki}(\tau) G_i^{WB}\right),$$

converges weakly to the same process as U_n on $\mathcal{D}[0,\tau]$. (Here, the estimates $\hat{H}_{k1i}$ and $\hat{H}_{k2i}$ are calculated by plugging appropriate sample estimates into the definition of H_{k1i} and H_{k2i}.)

Remark 1 The following choices of multipliers G_i^{WB} fulfill the necessary conditions for the wild bootstrap (cf. Dobler et al., 2017):

- $G_i^{WB} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, i.e., independent standard normal multipliers (according to the original resampling approach by Lin et al., 1993);
- $G_i^{WB} \overset{\text{i.i.d.}}{\sim} \mathcal{P}ois(1) - 1$, that is, independent and centered unit Poisson multipliers (in line with the proposition of Beyersmann et al.,2013);
- $G_i^{WB} \sim \mathcal{B}in\left(Y(T_i \wedge C_i), \frac{1}{Y(T_i \wedge C_i)}\right) - 1$ with $Y(t) = \sum_{i=1}^{n} Y_i(t)$ and $(G_{i_1}^{WB} \perp\!\!\!\perp G_{i_2}^{WB}) \mid \mathscr{F}_\tau$ for $i_1 \neq i_2$, i.e., conditionally independent, centered binomial multipliers.

This version of the wild bootstrap is equivalent to the so-called weird bootstrap described in Andersen, Borgan, Gill, and Keiding (1993, Subsect. IV.1.4), as Dobler et al. (2017) illustrate.

For a set of multiplier realizations $\{(G_1^{WB;(b)}, \ldots, G_n^{WB;(b)})^T\}_{b\in\{1,\ldots,B\}}$, one obtains the asymptotic $(1-\alpha)$ confidence interval

$$\left[\widehat{ATE}(t) - \frac{1}{\sqrt{n}} q_{1-\alpha}^{WB}(t),\ \widehat{ATE}(t) + \frac{1}{\sqrt{n}} q_{1-\alpha}^{WB}(t)\right],$$

with $(1-\alpha)$ quantile $q_{1-\alpha}^{WB}(t)$ of $\{|\hat{U}_n^{(b)}(t)|\}_{b\in\{1,\ldots,B\}}$. Similarly, an asymptotic simultaneous $(1-\alpha)$ confidence band over the interval $[t_1, t_2]$ is specified by

$$\left[\widehat{ATE}(t) - \frac{1}{\sqrt{n}} q_{1-\alpha}^{WB}\sqrt{\hat{v}^{WB}(t)},\ \widehat{ATE}(t) + \frac{1}{\sqrt{n}} q_{1-\alpha}^{WB}\sqrt{\hat{v}^{WB}(t)}\right],$$

considering the empirical variance estimator $\hat{v}^{WB}(t)$ of $\{\hat{U}_n^{(b)}(t)\}_{b\in\{1,\ldots,B\}}$ and the $(1-\alpha)$ quantile $q_{1-\alpha}^{WB}$ of

$$\left\{\sup_{t\in[t_1,t_2]}\left|\frac{\hat{U}_n^{(b)}(t)}{\sqrt{\hat{v}^{WB}(t)}}\right|\right\}_{b\in\{1,\ldots,B\}}.$$

The described bootstrap, just like the approach based on the influence function, takes only a fraction of the time required by the classical bootstrap. In addition, martingale-based analysis approaches for time-to-event data are built upon the condition of independent right-censoring and do not rely on a strict i.i.d. setup (Andersen et al., 1993, Subsect. III.2.2). Therefore, they are less sensitive to dependencies inherent to the data, where Efron's approach is known to fail (Rühl et al. 2022; see also Friedrich et al.,1981; Singh, 2017).

4 Simulation study

In order to compare the performance of the resampling approaches described in Sect. 3, we simulated competing risks data following the same scheme as in Ozenne et al. (2020), and constructed confidence intervals and bands using the proposed methods.

4.1 Data generation

The generated data comprised twelve independent covariates, namely, $Z_1, \ldots, Z_6$ following a mean-zero normal distribution and $Z_7, \ldots, Z_{12}$ being Bernoulli distributed with parameter 0.5. Each covariate affected the treatment probability, the event time distributions of two competing failure causes and a conditionally independent censoring time in an individual manner (see Table 1 and Fig. S1 in the supplementary material for a directed acyclic graph). The treatment indicator A was for instance derived from a logistic regression model with linear predictor $\alpha_0 + \log(2)\cdot(Z_1 - Z_2 + Z_6 + Z_7 - Z_8 + Z_{12})$. Here, the intercept α_0 controls the overall frequency of treatment. Apart from that, we simulated the event time based on a multistate model with Weibull hazards $\lambda_d(t) = 0.02\, t \exp(\beta_{dA} A + \boldsymbol{\beta}_{d\boldsymbol{Z}}^T \boldsymbol{Z})$ for $\boldsymbol{Z} = (Z_1, \ldots, Z_{12})$ and corresponding parameters β_{dA} and $\boldsymbol{\beta}_{d\boldsymbol{Z}}$, $d \in \{1, 2\}$ (cf. Beyersmann, Latouche et al. 2009). The censoring time was generated independently with hazard $\lambda_C(t) = \frac{2}{\gamma}\, t \exp(\boldsymbol{\beta}_{C\boldsymbol{Z}}^T \boldsymbol{Z})$, where γ determines the intensity of censoring.

This general simulation scheme served as a basis for a variety of scenarios, each implemented with sample sizes of $n \in \{50, 75, 100, 200, 300\}$ and treatment effects according to parameter $\beta_{1A} \in \{-2, 0, 2\}$. By default, about half of the observations were assigned to be treated, and the event of interest was observed in a third, half or two thirds of the subjects until time $t = 9$, corresponding to the case where $\beta_{1A} = -2,\ 0,\ 2$, respectively. The frequency of censoring amounted to 17%, 14% or 11% by $t = 9$, whereas the competing event affected 41%, 31% or 21% of the subjects.
Among the examined scenarios were settings with varying degrees of censoring (namely, 0%, 14% and 30% in the case without treatment effect, i.e., $\beta_{1A} = 0$), treatment frequencies of 22% as well as 86% and non-unit variances (0.25 and 4, respectively) of the normally distributed covariates $Z1, \ldots, Z_6$. Besides, we considered a standard survival scenario without competing events that involved type II censoring with staggered entry in order to investigate a setting with independent, but not random censoring (Rühl et al. 2022). For an overview of the different scenarios, see Table 2.

Confidence intervals (at time points $t \in \{1, 3, 5, 7, 9\}$) and bands (over the time interval $[0, 9]$) for the average treatment effect were derived by applying Efron's bootstrap (EBS), the influence function approach (IF) and the wild bootstrap (WBS) to each generated data set, using 1000 resampling replications, respectively. The WBS was realized with standard normal, Poisson and binomial multipliers according to Remark 1. We then assessed the performance of the distinct methods by means of the associated 95% coverage probabilities and the widths of the confidence ranges. The simulations were repeated 5000 times for each scenario to keep the Monte Carlo standard error for the coverage below 0.75%.

We approximated the true average treatment effect in the mentioned scenarios empirically, as the analytic form of $ATE(t)$ is hard to evaluate in the presence of multiple covariates. For that purpose, we simulated 1000 data sets with sample size $n = 100,000$ as previously described, but with random treatment assignment independent of the covariates

Table 1 Effects of the covariates on the treatment probability, event and censoring times

Covariate	Odds ratio w.r.t. Treatment probability[1]	Hazard ratio w.r.t. event of interest[2]	Hazard ratio w.r.t. competing event[3]	Hazard ratio w.r.t. censoring[4]
A	–	$\exp(\beta_{1A})$, $\beta_{1A} \in \{-2, 0, 2\}$	$\exp(\beta_{2A})$ = 1.0	$\exp(\beta_{CA})$ = 1.0
Z_1, Z_7	$\exp(\alpha_1) = \exp(\alpha_7)$ = 2.0	$\exp(\beta_{1,1}) = \exp(\beta_{1,7})$ = 2.0	$\exp(\beta_{2,1}) = \exp(\beta_{2,7})$ = 0.5	$\exp(\beta_{C,1}) = \exp(\beta_{C,7})$ = 0.5
Z_2, Z_8	$\exp(\alpha_2) = \exp(\alpha_8)$ = 0.5	$\exp(\beta_{1,2}) = \exp(\beta_{1,8})$ = 1.0	$\exp(\beta_{2,2}) = \exp(\beta_{2,8})$ = 1.0	$\exp(\beta_{C,2}) = \exp(\beta_{C,8})$ = 1.0
Z_3, Z_9	$\exp(\alpha_3) = \exp(\alpha_9)$ = 1.0	$\exp(\beta_{1,3}) = \exp(\beta_{1,9})$ = 2.0	$\exp(\beta_{2,3}) = \exp(\beta_{2,9})$ = 1.0	$\exp(\beta_{C,3}) = \exp(\beta_{C,9})$ = 1.0
Z_4, Z_{10}	$\exp(\alpha_4) = \exp(\alpha_{10})$ = 1.0	$\exp(\beta_{1,4}) = \exp(\beta_{1,10})$ = 1.0	$\exp(\beta_{2,4}) = \exp(\beta_{2,10})$ = 1.0	$\exp(\beta_{C,4}) = \exp(\beta_{C,10})$ = 2.0
Z_5, Z_{11}	$\exp(\alpha_5) = \exp(\alpha_{11})$ = 1.0	$\exp(\beta_{1,5}) = \exp(\beta_{1,11})$ = 1.0	$\exp(\beta_{2,5}) = \exp(\beta_{2,11})$ = 2.0	$\exp(\beta_{C,5}) = \exp(\beta_{C,11})$ = 1.0
Z_6, Z_{12}	$\exp(\alpha_6) = \exp(\alpha_{12})$ = 2.0	$\exp(\beta_{1,6}) = \exp(\beta_{1,12})$ = 2.0	$\exp(\beta_{2,6}) = \exp(\beta_{2,12})$ = 2.0	$\exp(\beta_{C,6}) = \exp(\beta_{C,12})$ = 0.5

1 $P(A = 1) = \text{expit}\left(\alpha_0 + \boldsymbol{\alpha_Z}^T \boldsymbol{Z}\right)$, with $\boldsymbol{\alpha_Z} = (\alpha_1, \ldots, \alpha_{12})^T$
2 $\lambda_1(t) = 0.02\, t \exp\left(\beta_{1A} A + \boldsymbol{\beta_{1Z}}^T \boldsymbol{Z}\right)$, with $\boldsymbol{\beta_{1Z}} = (\beta_{1,1}, \ldots, \beta_{1,12})^T$
3 $\lambda_2(t) = 0.02\, t \exp\left(\beta_{2A} A + \boldsymbol{\beta_{2Z}}^T \boldsymbol{Z}\right)$, with $\boldsymbol{\beta_{2Z}} = (\beta_{2,1}, \ldots, \beta_{2,12})^T$
4 $\lambda_C(t) = \frac{2}{\gamma}\, t \exp\left(\beta_{CA} A + \boldsymbol{\beta_{CZ}}^T \boldsymbol{Z}\right)$, with $\boldsymbol{\beta_{CZ}} = (\beta_{C,1}, \ldots, \beta_{C,12})^T$

Table 2 Overview of the simulation scenarios

Scenario	% Censored at $t = 9$ [1]			% Type 1 events at $t = 9$ [1]			% Treated	Var(Z_1)
	$\beta_{1A} = -2$	$\beta_{1A} = 0$	$\beta_{1A} = 2$	$\beta_{1A} = -2$	$\beta_{1A} = 0$	$\beta_{1A} = 2$		
No censoring	0.0	0.0	0.0	35.7	56.1	70.3	56.4	1.00
Light censoring	16.7	14.0	11.0	32.2	51.5	66.2	56.4	1.00
Heavy censoring	35.3	29.7	23.0	27.0	44.5	60.1	56.4	1.00
Low treatment probability	14.9	14.0	13.1	43.7	51.5	56.6	22.3	1.00
High treatment probability	18.2	14.0	8.3	23.5	51.5	75.6	85.8	1.00
Low variance of the covariates	13.7	10.7	7.3	32.4	55.2	72.0	57.4	0.25
High variance of the covariates	22.0	20.2	17.9	32.6	45.6	56.4	54.6	4.00
Type II censoring	49.7	39.2	25.0	50.0	49.5	48.4	56.4	1.00

[1] For the scenario with type II censoring, the percentages of censoring and type 1 events are determined at $t = 10$, $t = 5$, and $t = 2.5$, for $\beta_{1A} = -2,\ 0,\ 2$, respectively

and no censoring. For each of these data sets, the difference $\hat{F}_1(t \mid A = 1) - \hat{F}_1(t \mid A = 0)$ was determined, and our final estimate of the true average treatment effect is the median of the 1000 resulting values. Because of the large sample sizes considered, this approximation should be fairly close to the true value. Figure 1 depicts the approximated average treatment effect except for the scenarios with non-unit variance of the covariates $Z_1, \ldots, Z_6$ and those with type II censoring.

4.2 Results

The WBS attained coverage probabilities of the pointwise confidence intervals that were, in total, the closest to the target level of 95%. The mean absolute deviation across all scenarios, sample sizes and time points was 2.42% for the WBS vs. 2.49% and 2.61% for the IF and the EBS, respectively. (See Sects. 2.2 and 2.4 in the supplementary material for the coverage probabilities in the scenarios not presented here as well as the corresponding Monte Carlo standard errors.) Throughout nearly all settings, the confidence intervals obtained by the EBS yielded coverages above those derived from the different WBS versions, whereas the IF intervals included the true average treatment effect the least frequently. Figure 2 illustrates this ranking in the case with low-level censoring and a positive average treatment effect (i.e., $\beta_{1A} = 2$, referring here and in the following to the sign of the causal risk

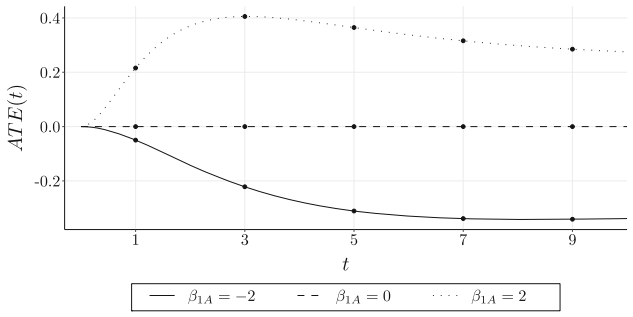

Fig. 1 Approximation of the true average treatment effect

difference; that is, a positive average treatment effect indicates that the potential cumulative incidence under treatment is higher than that under no treatment).We observed similar outcomes in the other scenarios that involved treatment effects according to $\beta_{1A} \in \{0, 2\}$ (see Figures S3, S5, S7, S8, S10, S13, S15, and S16 in the supplementary material), even though the performance of the resampling methods varied for early analysis time points (see e.g., Fig. 3).

An exception was the setting with widely dispersed covariates: Here, all methods provided rather conservative confidence intervals, and as a consequence, the IF approach achieved the most accurate coverages (see Figs. S18 and S19 in the supplementary material). The same effect was also encountered in the scenarios with negative treatment effect ($\beta_{1A} = -2$, see e.g., Fig. 4), once again excluding the setting with high variance of the covariates (where the EBS performed best for larger sample sizes, see Fig. S17 in the supplementary material). A common feature of all the schemes that yielded coverages along the lines of Fig. 4 is that the proportion of observed type 1 events was lower than in the scenarios with $\beta_{1A} \in \{0, 2\}$. This is due to the prevalence of the competing event, and the IF approach seems to be slightly more suitable to cope with that condition than the bootstrap methods.

On the other hand, the IF yielded fairly low coverage probabilities in several settings without treatment effect (see Fig. 3 and Figs. S2, S5, S7, S10, S15, and S21 in the supporting material). This issue remains with increasing sample sizes. Ozenne et al. (2020) encountered a similar pattern and considered a non-robust version of the influence function-based variance, which performed somewhat better.

The WBS generally reached its full potential towards later time points, when a sufficient amount of data was available. This became apparent in the scenario with type II censoring and a positive average treatment effect: Because of the absence of any competing events, we evaluated the confidence intervals at earlier times $t \in \{0.5, 1, 1.5, 2, 2.5\}$, and the WBS did not reach coverages as close to 95% as those obtained by the IF and the EBS until $t = 2$ (see Fig. S22 in the supporting material). For an explanation of this observation, note that the wild bootstrap process $\hat{U}_n(t)$ is based on the products $N_{ki}(t)\, G_i^{WB}$, for $i \in \{1, \ldots, n\}, k \in \{1, \ldots, K\}$. At early time points, the counting processes N_{ki} jump only rarely, and chances are that the few corresponding multipliers G_i^{WB} do not reflect the target distribution very well. Towards later times, a higher number of multipliers is taken into account, though, so the distribution of $\mathrm{d}N_{ki}(\cdot)\, G_i^{WB}$ will be closer to that of the martingale increments.

Against our expectations, the simulations revealed no significant superiority of the martingale-based methods in case of type II censoring with staggered entry, despite non-random censoring. It appears as if the dependence within the data was too weak for the sample sizes considered (cf. Rühl et al., 2022).

The coverage probabilities of the time-simultaneous confidence bands followed a similar trend as was observed for the pointwise intervals (see Sects. 2.3 and 2.4 in the supplementary material): While the highest and lowest coverages in almost all scenarios with positive or no average treatment effect were attained by the EBS and IF, respectively, there were only small differences in most of the settings with $\beta_{1A} = -2$. However, the EBS bands were especially accurate given positive average treatment effects ($\beta_{1A} = 2$, see e.g., Fig. 5). On average, the mean absolute discrepancy between the simulated coverages and the nominal level of 95% was 4.75% in comparison to 5.53% and 5.70% for the WBS and the IF approach, respectively.

Our results imply further that the choice of the multiplier for the WBS does not have any significant impact. Since the confidence intervals derived using the approaches of Lin et al. (1993) and Beyersmann et al. (2013) were occasionally wider than those resulting from the weird bootstrap, the latter method yielded lower coverages. Which of the multipliers provided the most accurate outcomes varied depending on the situation, however.

Other than that, the IF produced narrower intervals than any of the WBS versions, and in case of a negative average treatment effect, either approach lead to considerably greater variation in the interval width by comparison with the scenarios where $\beta_{1A} \in \{0, 2\}$. Interestingly, this effect did not apply to the EBS. The extent of the EBS-based intervals ranged between or above the remaining widths, apart from the settings with $\beta_{1A} = -2$. As the sample sizes increased, however, all resampling methods lead to nearly equally wide confidence intervals (cf. Fig. 6).

The widths of the confidence bands furthermore related to one another in the same way as their pointwise counterparts.

Due to the small sample sizes we considered, the number of observed events did occasionally not suffice to achieve convergence when the cause-specific Cox models were fitted. This is why some of the coverage probabilities are based on less than 5000 iterations for the influence function approach as well as the wild bootstrap, and less than 1000 bootstrap

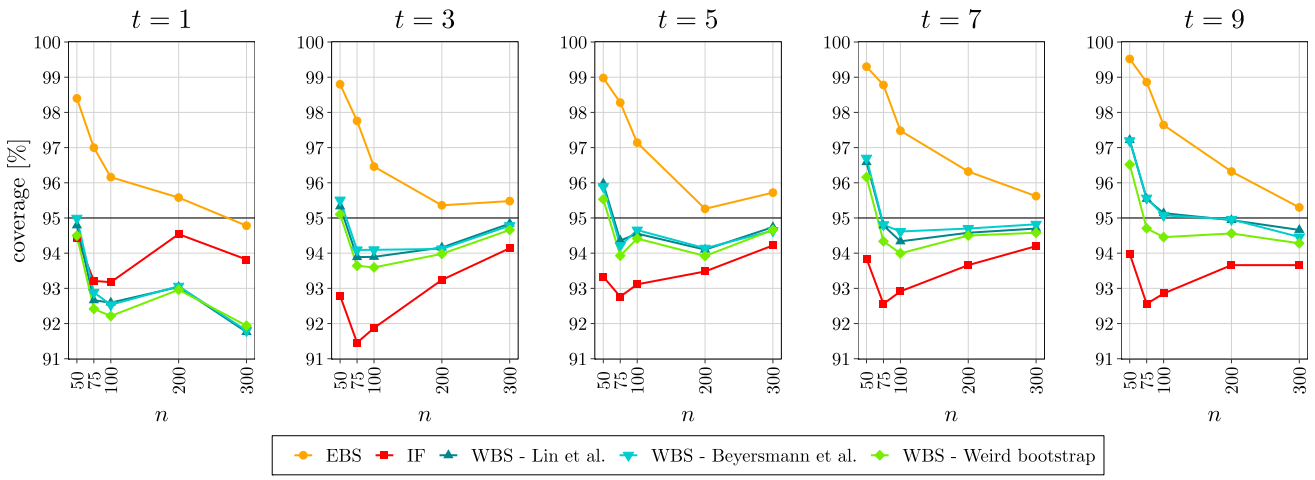

Fig. 2 Coverage of the confidence intervals in the scenario with light censoring (11% censored observations) and a positive average treatment effect ($\beta_{1A} = 2$)

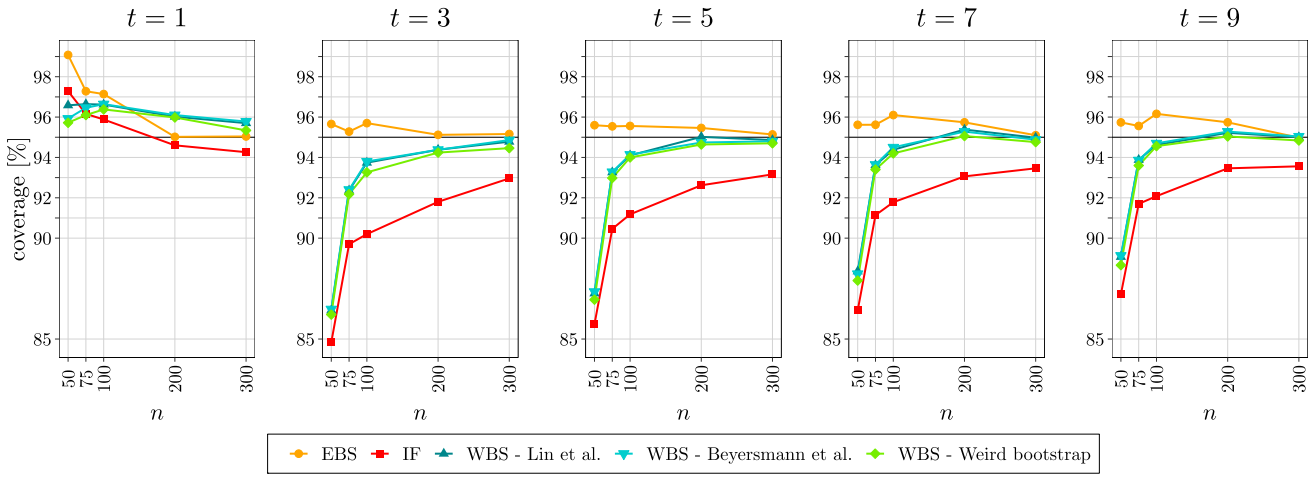

Fig. 3 Coverage of the confidence intervals in the scenario with high treatment probability (86% treated observations) and no treatment effect ($\beta_{1A} = 0$)

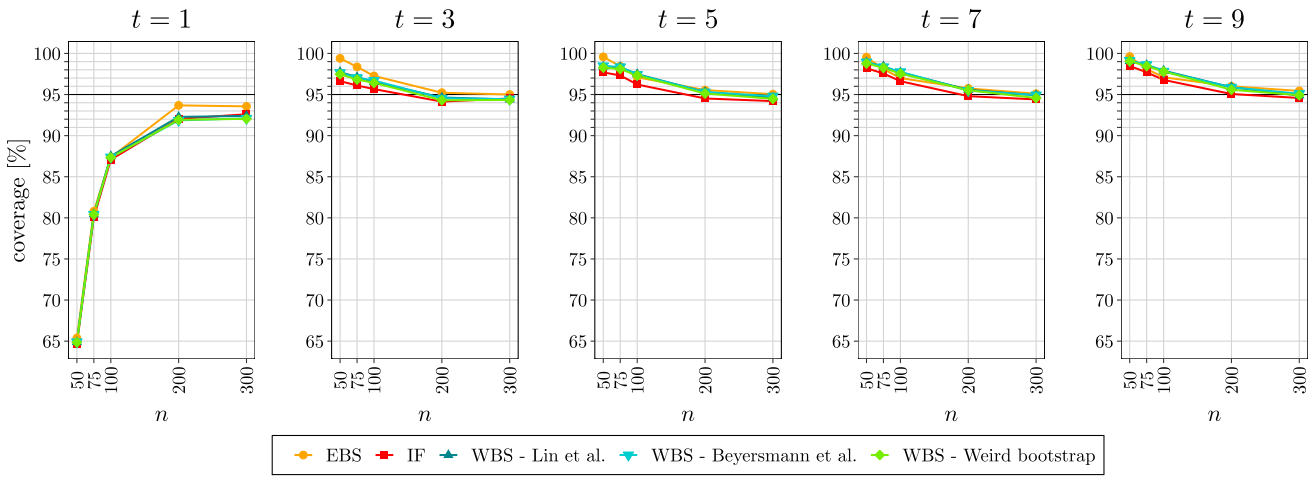

Fig. 4 Coverage of the confidence intervals in the scenario with no censoring (0% censored observations) and a negative average treatment effect ($\beta_{1A} = -2$)

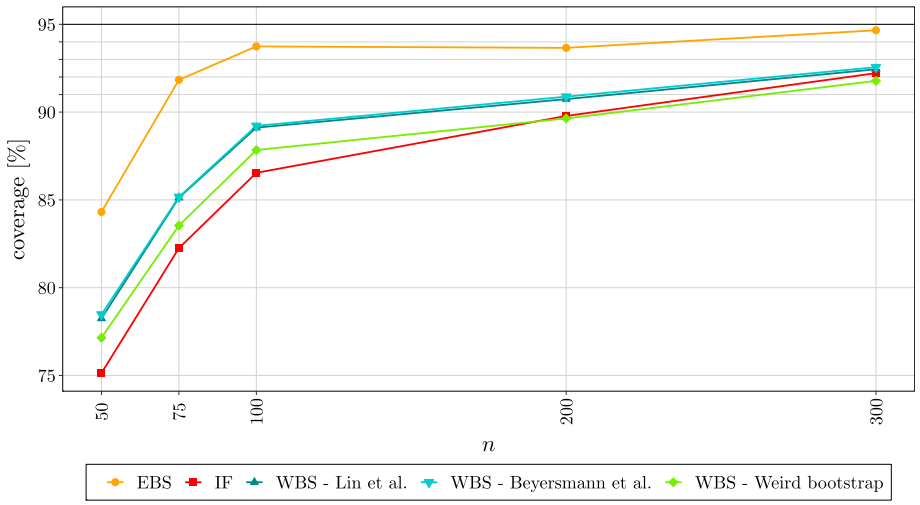

Fig. 5 Coverage of the confidence bands in the scenario with high treatment probability (86% treated observations) and a positive average treatment effect ($\beta_{1A} = 2$)

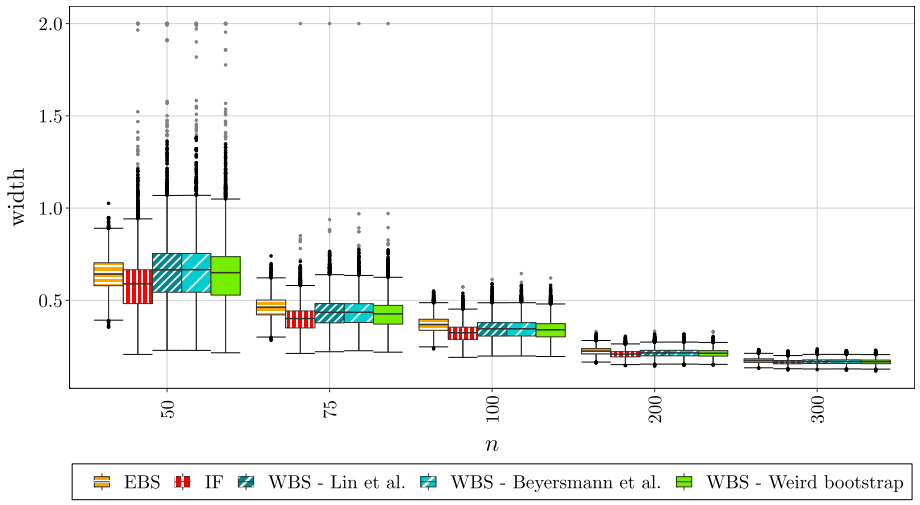

Fig. 6 Width of the confidence intervals at time $t = 5$ in the scenario with no censoring (0% censored observations) and a positive average treatment effect ($\beta_{1A} = 2$); note the spacing of the x-axis

samples when using Efron's approach. (The frequency of the convergence issues is shown in Table S10 in the supplementary material.) Results for the settings with $n = 50$ and $\beta_{1A} = 2$ should hence be analyzed with care.

Eventually, a last note is in order about the computation times of the distinct methods: The IF and EBS approaches have been implemented in the function '*ate*' of the `R` (R Core Team, 2021) package `riskRegression` by Gerds and Kattan (2021) (see Sect. 2.1.2 in the supplementary material for more information on the software we used). The calculations are sped up significantly by interfacing `C++` code for the IF method and parallelizing the computation of the bootstrap replicates for the EBS. We extracted and adapted the parts of the code that were relevant for our purposes. In addition, `C++` was also integrated to implement the WBS. The simulations were run on a high-performance computing cluster that operates on 2.4 GHz Intel® processors with 128 GB RAM, where we used 16 cores for parallel computations. Fig. 7 summarizes the resulting execution times for each resampling method.

Clearly, the EBS is several times slower than the multiplier-based methods and therefore, the IF approach as well as the WBS can in practice be implemented with a higher number of resampling repetitions, so that the accuracy of the resulting confidence regions is expected to be higher.

5 Real data application

To illustrate the performance of the resampling approaches when applied to real-world study data, we considered records of the long-term disease progression among patients with early-stage Hodgkin's lymphoma (i.e., stage I or II) (Pintilie 2006). These data are available within the `R` package `randomForestSRC` (data '*hd*', Ishwaran and Kogalur 2022) and comprise information on 865 subjects who were treated at the Princess Margaret Hospital in Toronto between 1968 and 1986, either with radiation alone ($n = 616$) or a combination of radiation and chemotherapy ($n = 249$). We studied the time (in years) from diagnosis until the com-

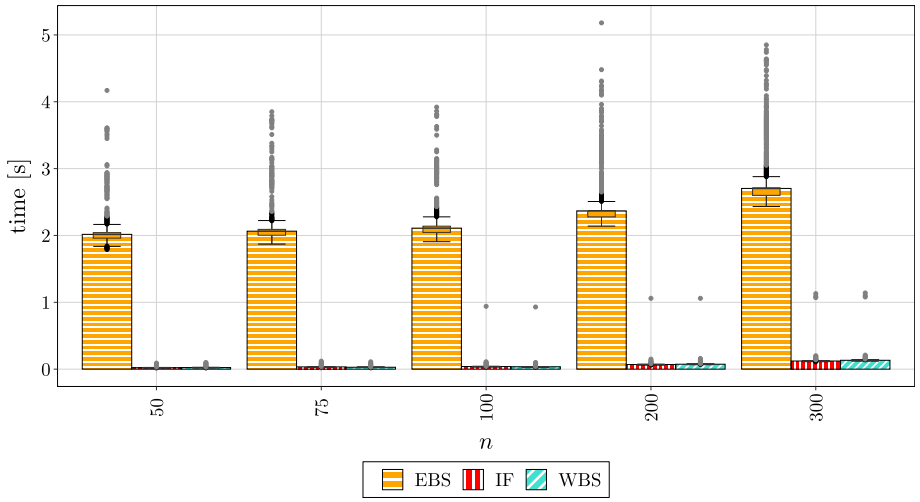

Fig. 7 Computation times in the scenario with no censoring (0% censored observations) and a positive average treatment effect ($\beta_{1A} = 2$). The height of the bars illustrates the mean computation time; note the spacing of the x-axis

Table 3 Summary of the Hodgkin's disease data

Covariate	Treatment Radiation alone ($n = 616$)	Radiation and chemotherapy ($n = 249$)
Age, mean (SD)	35.93 (16.37)	33.77 (12.86)
Sex: male	331 (53.73%)	132 (53.01%)
Lymphoma stage: I	266 (43.18%)	30 (12.05%)
Mediastinum involvement		
None	382 (62.01%)	82 (32.93%)
Small	211 (34.25%)	77 (30.92%)
Large	23 (3.73%)	90 (36.14%)
Extranodal disease	29 (4.70%)	50 (20.08%)

peting events of relapse and death, respectively. Random values of very small extent (i.e., normally distributed variables with mean zero and variance 10^{-6}) were added to the event times in order to break any ties in the data that emerged due to rounding. Covariates recorded include age, sex, clinical stage of the lymphoma, size of mediastinum involvement and whether the disease was extranodal (see Table 3 for a summary of the data). For our analysis, we assume that these variables are sufficient for confounding adjustment, and that the positivity and consistency conditions are met w.r.t. the two therapies. Moreover, tests on the scaled Schoenfeld residuals of the Cox models for both causes did not suggest any violations of the proportional hazards assumption apart from the variable age in the relapse model (Grambsch and Therneau 1994; see Figs. S46 and S47 in the supplementary material). The estimated coefficient in a corresponding model with time-dependent covariate is nearly constant over time, though. We thus use simple Cox models (with time-constant covariates) to derive the average treatment effect.

Our analysis suggests that after 30 years, the risk of relapse would be reduced by 17.89% in a hypothetical setting where every subject had been treated with both radiation and chemotherapy as compared to the case where everyone had received radiation therapy only (see Fig. S48 in the supplementary material). Simultaneously, the risk of death would be raised by 9.49% between these scenarios (see Fig. 8). Note how the ATE concerning relapse drops rather sharply within the first 5 years, whereas the ATE w.r.t. death increases gradually over the entire 30-year interval. In conclusion, treatment with the combined therapy seems to effectively prevent relapse in the studied population, but since we consider competing causes, a decrease in relapse events will leave more subjects who die without prior relapse.

In Fig. 8, it can be seen that all resampling methods lead to fairly similar confidence intervals concerning the effect on death. Yet the EBS confidence bands are notably wider than those derived from the remaining approaches.

On the other hand, relapse events are observed more than twice as often as deaths, which is why the corresponding confidence intervals and bands are closer to each other.

6 Discussion

The article at hand compares three resampling methods for the derivation of confidence intervals and bands for the aver-

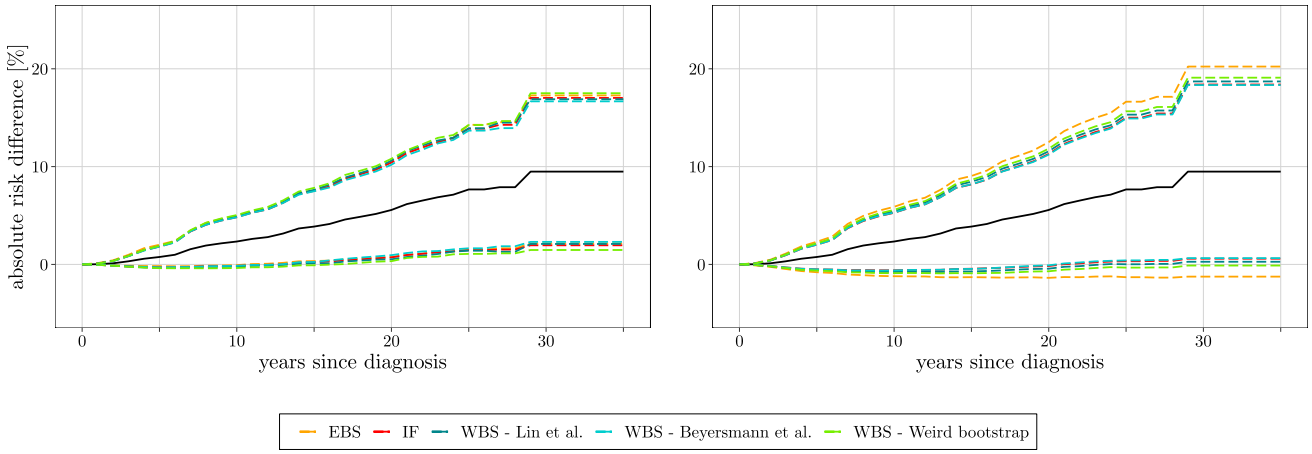

Fig. 8 Confidence intervals (left) and bands (rights) for the average treatment effect on the risk of death

age treatment effect in competing risks settings (although the influence function-based confidence intervals, strictly speaking, do not rely on resampling). As our simulations show, the wild bootstrap yields correct coverage levels for pointwise confidence intervals in the presence of rather small data sets, provided that sufficient events have been observed until the considered time point. This applies regardless of the type of multiplier that is implemented (i.e., standard normal, centered Poisson, or weird bootstrap multipliers). The theory behind the wild bootstrap relies on martingales and therefore accommodates counting processes, which are naturally used to represent time-to-event data. As a consequence, it is straightforward to tackle common issues in survival analysis, such as e.g., left-truncation. (Note the controversy about left-truncation in causal contexts, though, cf. Hernán, 2015; Vandenbroucke and Pearce 2015.) If competing events prevail (like it was the case in the scenarios with $\beta_{1A} = -2$ in our simulation study), one may prefer the influence function approach (or a non-robust version, as proposed by Ozenne et al. 2020, if the treatment is unlikely to have any effect), and if earlier time points are examined, the classical bootstrap seems to be a reasonable choice. The latter also achieves very accurate coverages with respect to time-simultaneous confidence bands. As the amount of available data increases, the differences between the distinct resampling approaches fade away. Efron's simple bootstrap, which is most commonly used in practice, requires considerable computation time, however. What is more, dependencies might cause issues with this resampling method (Singh 1981; Friedrich et al. 2017; Rühl et al. 2022), even though our simulations did not disclose any major bias in this context.

The three covered approaches were additionally compared given real data on the long-term risk of relapse and death among patients with early-stage Hodgkin's disease (Pintilie 2006). While the outcomes are generally quite similar, Efron's bootstrap generated somewhat wider confidence bands for the average treatment effect on the risk of death.

It should be noted that for consistent estimation of the average treatment effect, the model for the cumulative incidence function must be correctly specified. Instead of the cause-specific Cox model used here, one might employ alternatives such as the nonparametric additive hazards model proposed by Aalen (1980) (cf. Ryalen et al. 2018), or the Fine-Gray regression model for $F_1(t \mid a, z)$ adopting the subdistribution approach (see Rudolph et al. 2020 or the more technical discourse by Young et al., 2020 for a discussion on cause-specific vs. subdistribution measures in causal frameworks). In the latter case, however, additional considerations on the associated stochastic process are necessary to make inferences on $\widehat{ATE}$.

We did not address estimators based on inverse probability of treatment weighting (IPTW, which requires correct specification of a treatment model rather than the outcome model) or the doubly-robust version combining both the g-formula and IPTW. This is because one would need to derive the asymptotic distributions of the corresponding processes to justify the application of any resampling methods, which is beyond the scope of this work. Only the representation of the processes based on the influence function has been determined already, see Ozenne et al. (2020) for more details.

In order to handle complex conditions that are often observed in real-world trials with time-varying treatments, a possible subject of future work is the extension of the investigated resampling methods to settings that involve time-dependent confounding. The standard time-dependent Cox analysis has been shown to yield incorrect results in such settings (Hernán et al. 2000), which is why it is important to incorporate appropriate models (see e.g. Keogh et al. 2023).

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11222-024-10420-w.

Acknowledgements Support by the Deutsche Forschungsgemeinschaft (DFG Grant FR 4121/2-1) is gratefully acknowledged. The authors also appreciate the helpful feedback from B. Ozenne and the constructive comments by two anonymous reviewers.

Funding Open Access funding enabled and organized by Projekt DEAL.

Data availability The data that support the findings of this study are openly available in the `R` package `randomForestSRC` at https://cran.r-project.org/web/packages/randomForestSRC (Ishwaran and Kogalur 2022).

Code Availability An `R` package for the computation of the average treatment together with the proposed confidence intervals and bands, as well as the code to reproduce the results of the simulation study and the real data analysis is available on github (https://github.com/jruehl/ATESurvival).

Declarations

Conflict of interest The authors have no competing interests to declare.

References

Aalen, O.O.: Nonparametric inference for a family of counting processes. Annals of Statistics **6**, 701–726 (1978)

Aalen, O.O.: A model for non-parametric regression analysis of counting processes. W. Klonecki, A. Kozek, & J. Rosiński (Eds.), Mathematical statistics and probability theory (pp. 1-25). New York: Springer (1980)

Aalen, O.O., Cook, R.J., Røysland, K.: Does Cox analysis of a randomized survival study yield a causal treatment effect? Lifetime Data Analysis **21**, 579–593 (2015)

Andersen, P.K., Borgan, ø., Gill, R.D., Keiding, N.: Statistical models based on counting processes (1st ed.). Springer, New York (1993)

Benichou, J., Gail, M.H.: Estimates of absolute cause-specific risk in cohort studies. Biometrics **46**(3), 813–826 (1990)

Beyersmann, J., Di Termini, S., Pauly, M.: Weak convergence of the wild bootstrap for the Aalen-Johansen estimator of the cumulative incidence function of a competing risk. Scandinavian Journal of Statistics **40**(3), 387–402 (2013)

Beyersmann, J., Latouche, A., Buchholz, A., Schumacher, M.: Simulating competing risks data in survival analysis. Statistics in Medicine **28**, 956–971 (2009)

Breslow, N.E.: Contribution to discussion of paper by DR Cox. Journal of the Royal Statistical Society, Series B **34**, 216–217 (1972)

Butt, J.H., De Backer, O., Olesen, J.B., Gerds, T.A., Havers-Borgersen, E., Gislason, G.H., Fosbøl, E.L.: Vitamin K antagonists vs. direct oral anticoagulants after transcatheter aortic valve implantation in atrial fibrillation. European Heart Journal Cardiovascular Pharmacotherapy **7**(1), 11–19 (2021)

Chauhan, L., Pattee, J., Ford, J., Thomas, C., Lesteberg, K., Richards, E., Beckham, J.D.: A multicenter, prospective, observational, cohort-controlled study of clinical outcomes following coronavirus disease 2019 (COVID-19) convalescent plasma therapy in hospitalized patients with COVID-19. Clinical Infectious Diseases **75**(1), e466–e472 (2022)

Cheng, S.C., Fine, J.P., Wei, L.J.: Prediction of cumulative incidence function under the proportional hazards model. Biometrics **54**(1), 219–228 (1998)

Dobler, D., Beyersmann, J., Pauly, M.: Non-strange weird resampling for complex survival data. Biometrika **104**(3), 699–711 (2017)

Efron, B.: Censored data and the bootstrap. Journal of the American Statistical Association **76**(374), 312–319 (1981)

FDA: Adjusting for covariates in randomized clinical trials for drugs and biological products. Draft guidance for industry (2023)

Friedrich, S., Brunner, E., Pauly, M.: Permuting longitudinal data in spite of the dependencies. Journal of Multivariate Analysis **153**, 255–265 (2017)

Gerds, T.A., Kattan, M.W.: Medical risk prediction models: With ties to machine learning (1st ed.). Chapman and Hall/CRC (2021)

Grambsch, P.M., Therneau, T.M.: Proportional hazards tests and diagnostics based on weighted residuals. Biometrika **81**(3), 515–526 (1994)

Hernán, M.A.: The hazards of hazard ratios. Epidemiology **21**(1), 13–15 (2010)

Hernán, M.A.: Counterpoint: epidemiology to guide decision-making: moving away from practice-free research. American Journal of Epidemiology **182**(10), 834–839 (2015)

Hernán, M.A., Brumback, B., Robins, J.M.: Marginal structural models to estimate the causal effect of Zidovudine on the survival of HIV-positive men. Epidemiology **11**(5), 561–570 (2000)

Hernán, M.A., Robins, J.M.: Causal inference: What if. Chapman & Hall/CRC, Boca Raton (2020)

Ishwaran, H., Kogalur, U.B.: Fast unified random forests for survival, regression, and classification (rf-src) [Computer software manual]. manual. Retrieved from https://cran.rproject.org/package=randomForestSRC (R package version 3.1.1) (2022)

Keogh, R.H., Gran, J.M., Seaman, S.R., Davies, G., Vansteelandt, S.: Causal inference in survival analysis using longitudinal observational data: Sequential trials and marginal structural models. Statistics in Medicine **42**(13), 2191–2225 (2023)

Lin, D.Y., Wei, L.J., Ying, Z.: Checking the Cox model with cumulative sums of martingale-based residuals. Biometrika **80**(3), 557–572 (1993)

Martinussen, T., Vansteelandt, S.: On collapsibility and confounding bias in Cox and Aalen regression models. Lifetime Data Analysis **19**(3), 279–296 (2013)

Neumann, A., Billionnet, C.: Covariate adjustment of cumulative incidence functions for competing risks data using inverse probability of treatment weighting. Computer Methods and Programs in Biomedicine **129**, 63–70 (2016)

Nørgaard, M., Ehrenstein, V., Vandenbroucke, J.P.: Confounding in observational studies based on large health care databases: problems and potential solutions -a primer for the clinician. Clinical Epidemiology **9**, 185–193 (2017)

Ozenne, B.M.H., Scheike, T.H., Staerk, L., Gerds, T.A.: On the estimation of average treatment effects with right-censored time to event outcome and competing risks. Biometrical Journal **62**(3), 751–763 (2020)

Ozenne, B.M.H., Sørensen, A.L., Scheike, T.H., Torp-Pedersen, C., Gerds, T.A.: riskRegression: predicting the risk of an event using Cox regression models. The R Journal **9**(2), 440–460 (2017)

Philipps, W., Fietz, A.-K., Meixner, K., Bluhmki, T., Meister, R., Schaefer, C., Padberg, S.: Pregnancy outcome after first-trimester

exposure to fosfomycin for the treatment of urinary tract infection: an observational cohort study. Infection **48**, 57–64 (2020)
Pintilie, M.: Competing risks: A practical perspective. John Wiley & Sons (2006)
R Core Team (2021). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.Rproject.org/Rubin,
Rubin, D.B.: Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology, 66 (5), 688–701 (1974)
Rudolph, J.E., Lesko, C.R., Naimi, A.I.: Causal inference in the face of competing events. Current Epidemiology Reports, 7 (3), 125-131 (2020)
Ryalen, P.C., Stensrud, M.J., Fosså, S., Røysland, K.: Causal inference in continuous time: an example on prostate cancer therapy. Biostatistics, 21(1), 172-185 (2020)
Ryalen, P.C., Stensrud, M.J., Røysland, K.: Transforming cumulative hazard estimates. Biometrika **105**(4), 905–916 (2018)
Rühl, J., Beyersmann, J., Friedrich, S.: General independent censoring in event-driven trials with staggered entry. Biometrics **79**, 1737–1748 (2022)
Rühl, J., Friedrich, S.: Asymptotic properties of resampling-based processes for the average treatment effect in observational studies with competing risks. arXiv:2306.02970 [math-STAT] (2023)
Scheike, T.H., Zhang, M.-J.: Flexible competing risks regression modeling and goodness-of-fit. Lifetime Data Analysis **14**, 464–483 (2008)
Singh, K.: On the asymptotic accuracy of Efron's bootstrap. The Annals of Statistics **9**(6), 1187–1195 (1981)
Stensrud, M.J., Young, J.G., Didelez, V., Robins, J.M., Hernán, M.A.: Separable effects for causal inference in the presence of competing events. Journal of the American Statistical Association **117**(537), 175–183 (2020)
Vandenbroucke, J., Pearce, N.: Point: incident exposures, prevalent exposures, and causal inference: does limiting studies to persons who are followed from first exposure onward damage epidemiology? American Journal of Epidemiology **182**(10), 826–833 (2015)
Yang, W., Zilov, A., Soewondo, P., Bech, O.M., Sekkal, F., Home, P.D.: Observational studies: going beyond the boundaries of randomized controlled trials. Diabetes Research and Clinical Practice **88**, 3–9 (2010)
Young, J.G., Stensrud, M.J., Tchetgen Tchetgen, E.J., Hernán, M.A.: A causal framework for classical statistical estimands in failure-time settings with competing events. Statistics in Medicine **39**, 1199–1236 (2020)