Resampling-Based Inference for Causal Effect Estimates in Time-to-Event Data

Dissertation

zur Erlangung des akademischen Grades

Dr. rer. nat.

eingereicht an der

Mathematisch-Naturwissenschaftlich-Technischen Fakultät

der Universität Augsburg

von

Jasmin Rühl

Augsburg, Oktober 2024

Erstgutachter:	**Prof. Dr. Sarah Friedrich** Mathematical Statistics and Artificial Intelligence in Medicine, Universität Augsburg
Zweitgutachter:	**Prof. Dr. Torben Martinussen** Section of Biostatistics, University of Copenhagen

Datum der mündlichen Prüfung: 13.02.2025

Abstract

Event-driven trials with staggered entry are frequently encountered in clinical practice. However, the associated data involve dependencies, and it is not obvious whether independent censoring – which is an essential condition for the validity of the common survival methodology – applies in this context. The corresponding proof is subject to the restriction that the calendar times are masked, as information on the order of the events may otherwise derange the underlying intensities. Using simulations, we show that violations of said constraint can entail erroneous results. Furthermore, we analyse simulated data by means of Efron's classic bootstrap. This resampling method is based on the assumption of random censoring, and thus, leads to bias, whereas the outcomes obtained by the wild bootstrap are correct. When analysing event-driven trials with staggered entry, one should therefore both exclude calendar time information and resort to martingale-based techniques in order to draw valid conclusions.

The second focal point of this work is on methods for statistical inference about causal effect estimators. We define the treatment effect in terms of the cumulative incidence, which allows for the incorporation of competing risks.

A corresponding effect estimator can be determined using the g-formula. The distribution of the associated stochastic process is rather complex, however, and in practice, researchers mostly resort to resampling by Efron's nonparametric bootstrap in order to construct confidence intervals and bands. We show that the classical bootstrap, a resampling method based on the influence function, as well as the wild bootstrap can be used to approximate the asymptotic distribution of the stochastic process of interest. Moreover, we conduct a simulation study to compare confidence regions derived by means of the respective resampling techniques. It becomes apparent that the wild bootstrap generally yields the most accurate results, unless the event of interest is observed only rarely. In that case, the approach based on the influence function attains valid confidence regions. For time-simultaneous confidence bands, the classical bootstrap should be considered.

Apart from the g-formula, propensity score matching provides another possibility for estimating the causal treatment effect. There exists only little research, however, regarding suitable variance estimators that take into account the particular structure of the matched data. We adapt a double-resampling technique as well as a cluster-based variance estimator to the situation at hand and, once more, perform simulations comparing the corresponding confidence regions. Both approaches yield appropriate, yet somewhat conservative outcomes, and will thus be further improved as part of future investigations.

Zusammenfassung

Ereignisgesteuerte Studien mit zeitversetztem Eintritt werden in der klinischen Praxis häufig eingesetzt. Die zugehörigen Daten weisen jedoch Abhängigkeiten auf und es ist nicht offensichtlich, ob in diesem Zusammenhang von unabhängiger Zensierung – einer wesentlichen Voraussetzung für die Gültigkeit der gängigen Überlebenszeitmethodik – ausgegangen werden kann. Der entsprechende Beweis setzt voraus, dass keine Kalenderzeiten in die Analyse miteingehen, da Informationen zur Reihenfolge der Ereignisse andernfalls die zugrundeliegenden Intensitäten stören können. Mit Hilfe von Simulationen zeigen wir, dass Verletzungen der genannten Einschränkung falsche Ergebnisse erzeugen. Außerdem analysieren wir simulierte Daten unter Verwendung von Efrons klassischem Bootstrap. Diese Resampling-Methode basiert auf der Annahme zufälliger Zensierung und führt deshalb zu Verzerrungen, wohingegen der Wild-Bootstrap in unseren Untersuchungen korrekte Ergebnisse liefert. Bei der Auswertung ereignisgesteuerter Studien mit zeitversetztem Eintritt sollte man daher einerseits Kalenderzeitinformationen ausschließen und andererseits auf Martingal-basierte Methoden zurückgreifen, um gültige Rückschlüsse ziehen zu können.

Den zweiten Schwerpunkt dieser Arbeit bilden Methoden zur statistischen Inferenz für kausale Effektschätzer. Wir definieren den Behandlungseffekt in Bezug auf die kumulative Inzidenz, sodass auch konkurrierende Risiken berücksichtigt werden können.

Ein entsprechender Schätzer lässt sich mittels der g-Formel bestimmen. Die Verteilung des zugehörigen stochastischen Prozesses ist jedoch relativ komplex, weshalb man sich in der Praxis zur Herleitung von Konfidenzintervallen und -bändern meist mit Resampling durch Efrons nichtparametrischen Bootstrap behilft. Wir weisen nach, dass der klassische Bootstrap, eine Resampling-Methode basierend auf der Einflussfunktion, sowie der Wild-Bootstrap verwendet werden können, um die asymptotische Verteilung des untersuchten stochastischen Prozesses zu approximieren. Außerdem führen wir eine Simulationsstudie durch, um Konfidenzregionen zu vergleichen, die mit Hilfe der erwähnten Resampling-Techniken bestimmt werden. Wie sich zeigt, führt der Wild-Bootstrap im Allgemeinen zu den treffendsten Ergebnissen, es sei denn, das interessierende Ereignis wird nur selten beobachtet. In diesem Fall erzielt die Methode, die auf der Einflussfunktion beruht, gültige Konfidenzregionen. Der klassische Bootstrap empfiehlt sich für zeitsimultane Konfidenzbänder.

Neben der g-Formel bietet zudem Propensity Score Matching eine Alternative für die Schätzung des kausalen Behandlungseffektes. Es liegen jedoch kaum Untersuchungen zu geeigneten Varianzschätzern vor, welche die besondere Struktur der gematchten Daten einbeziehen. Wir passen eine Doppel-Resampling-Technik sowie einen clusterbasierten Varianzschätzer an die gegebene Situation an und führen erneut Simulationen zum Vergleich der entsprechenden Konfidenzregionen durch. Beide Methoden liefern gültige, aber etwas zu konservative Ergebnisse und sollen deshalb im Rahmen zukünftiger Untersuchungen weiter verbessert werden.

Acknowledgements

At this point, I would like to thank all persons who supported me during the completion of this dissertation.

First and foremost, I am grateful to my supervisor Prof. Dr. Sarah Friedrich for introducing me to the topic of my research as well as for her consistent guidance in numerous aspects, including (but not limited to) the provision of constructive feedback that contributed significantly to the present work.

My gratitude also goes to Prof. Dr. Jan Beyersmann, who sparked my interest in time-to-event analysis during my years of study and who encouraged me to pursue a PhD.

I further want to thank Prof. Dr. Torben Martinussen for agreeing to act as second reviewer of this thesis.

Many thanks go to my former and current colleagues at the Chair of Mathematical Statistics and Artificial Intelligence in Medicine. They helped to create a pleasant working environment not only at the university, but also during various conferences.

I am moreover thankful to Brice Ozenne and his colleagues at the University of Copenhagen for the opportunity of a research visit. In this context, I would also like to express my gratitude for the funding by the DFG.

Last but not least, this dissertation has benefited greatly from the support of my family and friends, who helped to take my mind off my thesis every now and then.

Publications

Parts of the material introduced in this thesis have already been published in peer-reviewed scientific journals:

- Results from Rühl, Beyersmann, and Friedrich (2023) are described in Chapter 3. In addition, some of the findings in Chapter 3 overlap with my master thesis on the same subject.
- The theoretical considerations presented in Rühl and Friedrich (2024a) are covered in Subsection 4.1.1.
- Subsection 4.1.2 is based on the simulation study in Rühl and Friedrich (2024b). The outcomes of the real data application in the same manuscript are further reported in Subsection 4.1.3.
- Besides, Chapters 1 and 5 address aspects from all of the mentioned references.

I am the first author of these articles, respectively, and was in charge of their preparation, including the implementation of proofs, simulations, and analyses. The corresponding co-authors supervised the work and reviewed first drafts as well as revisions.

References:

Rühl, J., Beyersmann, J., and Friedrich, S. (2023). "General independent censoring in event-driven trials with staggered entry." In: *Biometrics* 79.3, pp. 1737–1748.

Rühl, J. and Friedrich, S. (2024a). "Asymptotic properties of resampling-based processes for the average treatment effect in observational studies with competing risks." In: *Scandinavian Journal of Statistics*. DOI: 10.1111/sjos.12714, pp. 1–27.

Rühl, J. and Friedrich, S. (2024b). "Resampling-based confidence intervals and bands for the average treatment effect in observational studies with competing risks." In: *Statistics and Computing* 34.101.

Contents

List of figures

List of tables

List of abbreviations

Acronyms are listed in alphabetical order.

ATE	Average treatment effect
Càdlàg	Continue à droite, limite à gauche (right-continuous with left limits)
CB	Confidence band
CDF	Cumulative distribution function
CI	Confidence interval
CIF	Cumulative incidence function
DAG	Directed acyclic graph
DR	Double-resampling
EBS	Efron's nonparametric bootstrap
HR	Hazard ratio
IF	Influence function
I.i.d.	Independent and identically distributed
IPT	Inverse probability of treatment
MCSE	Monte Carlo standard error
PS	Propensity score
RMSE	Root mean square error
TTE	Time-to-event
WBS	Wild bootstrap

List of symbols

Mathematical terms and symbols are listed in the order they appear in the thesis. Note that vectors and matrices are denoted by bold symbols.

Denotation

τ	Terminal time
T	Survival time
$S(t)$	Survival function at time t
$\alpha(t)$	Hazard function at time t
$A(t)$	Cumulative hazard function at time t, with Nelson-Aalen estimator $\hat{A}(t)$
n	Sample size
$N^c(t)\,/N(t)$	Counting process at time t based on completely observed/ censored data
$\mathscr{F}_t^c/\mathscr{F}_t\,/\mathscr{G}_t$	History until time t given completely observed/censored/ both completely observed & censored data
$\lambda^{\mathscr{F}^c}(t)\,/\,\lambda^{\mathscr{F}}(t)\,/\,\lambda^{\mathscr{G}}(t)$	Intensity process at time t w.r.t. $\left(\mathscr{F}_t^c\right)/(\mathscr{F}_t)/(\mathscr{G}_t)$
$Y^c(t)\,/\,Y(t)$	At-risk process for T at time t based on completely observed/ censored data
C	Censoring time
D	Censoring/event type indicator
$Y^o(t)$	Censoring process at time t
$M(t)$	$(\mathscr{F}_t)$-martingale w.r.t. the Doob-Meyer decomposition of $N(t)$
$W(t)$	Brownian motion at time t
$\mathbf{Z}$	Baseline covariate vector, including treatment indicator Z_A and covariates $\mathbf{Z}_L$
p	Number of baseline covariates
$\alpha_0(t)$	Baseline hazard at time t
$A_0(t)$	Cumulative baseline hazard at time t, with Breslow estimator $\hat{A}_0(t)$
$\boldsymbol{\beta}_0$	Cox model parameter vector, with Cox estimator $\hat{\boldsymbol{\beta}}$

$S^{(0)}(\boldsymbol{\beta}, t)$, $\mathbf{S}^{(1)}(\boldsymbol{\beta}, t)$, $\mathbf{S}^{(2)}(\boldsymbol{\beta}, t)$	Functions used to describe statistics based on the Cox partial likelihood, with expectations $s^{(0)}(\boldsymbol{\beta}, t)$, $\mathbf{s}^{(1)}(\boldsymbol{\beta}, t)$, $\mathbf{s}^{(2)}(\boldsymbol{\beta}, t)$ (see p. 11f)
$\mathbf{E}(\boldsymbol{\beta}, t)$	Function used to describe statistics based on the Cox partial likelihood, with large sample limit $\mathbf{e}(\boldsymbol{\beta}, t)$ (see p. 11f)
$\boldsymbol{\Sigma}$	Standardized information matrix for the Cox estimator $\hat{\boldsymbol{\beta}}$
K	Number of competing events
$F_k(t)$	Cumulative incidence of cause k at time t
$PS(\mathbf{l})$	Propensity score given covariate vector $\mathbf{l}$
$ATE(t)$	Average treatment effect w.r.t. $F_1(t)$, with estimator $\widehat{ATE}(t)$
B	Number of bootstrap samples
Q / R	Time of entry/event in calendar time scale
m	Number of events to be observed under type II censoring
G	Random multiplier
$U_n(t)$	Stochastic process characterising the distribution of the average treatment effect at time t
$\xi(t_1, t_2)$	Covariance function of the limiting distribution of $\big(U_n(t)\big)$

Mathematical symbols

$\mathbb{1}\{A\}$	Indicator function of A
$\sigma(A)$	σ-algebra generated by A
$A(t-)$	Left-hand limit $\lim_{s \nearrow t} A(s)$ of A at t
$\mathrm{d}A(t)$	Increment $A((t+\mathrm{d}t)-) - A(t-)$ of A over $[t, t+\mathrm{d}t)$
$a \wedge b$	Minimum of a and b
$A \perp\!\!\!\perp B$	Stochastic independence of A and B
$A_n \xrightarrow{P} A$	Convergence of A_n to A in probability
$A_n \xrightarrow{\mathscr{D}} A$	Convergence in distribution/weak convergence of A_n to A
$D[a, b]$	Skorokhod space of real-valued, càdlàg functions on $[a, b]$
$\boldsymbol{a}^{\otimes 0}$ / $\boldsymbol{a}^{\otimes 1}$ / $\boldsymbol{a}^{\otimes 2}$	$\boldsymbol{a}^{\otimes 0} = 1$ / $\boldsymbol{a}^{\otimes 1} = \boldsymbol{a}$ / $\boldsymbol{a}^{\otimes 2} = \boldsymbol{a}\boldsymbol{a}^T$, for column vector $\boldsymbol{a}$
$A \times B$	Cartesian product of A and B
$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	Normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$
$O^{A=a}$ (or O^a)	Potential outcome O under intervention $A = a$
$\mathcal{M}(n, \mathbf{p})$	Multinomial distribution w.r.t. n trials and probability vector $\mathbf{p}$
$\hat{\theta}^*$	Bootstrap replication of $\hat{\theta}$
$\partial\theta_G$	Hadamard derivative of θ at G
$IF_\theta(x)$	Influence function of θ at x
$o_P(a_n)$	$o_P(a_n)/a_n \xrightarrow[n\to\infty]{P} 0$, for a sequence a_n
$Exp(\lambda)$	Exponential distribution with rate parameter $\boldsymbol{\mu}$
$\mathcal{U}(a, b)$	Uniform distribution over the interval $[a, b]$
$q_{\mathcal{D}}(x)$	x quantile of the distribution $\mathcal{D}$
$\mathcal{W}b(k, \lambda)$	Weibull distribution with shape and scale parameters k and λ
$\langle A, B \rangle(t)$	Predictable covariation process of A and B at time t
$A_n \xrightarrow{a.s.} A$	Almost sure convergence of A_n to A
$\mathcal{P}oi(\lambda)$	Poisson distribution with rate parameter λ
$\mathcal{B}in(n, p)$	Binomial distribution w.r.t. n trials and probability p
$O_P(a_n)$	$\lim_{M\to\infty} \limsup_{n\to\infty} P(\lvert O_P(a_n)/a_n\rvert > M) = 0$, for a sequence a_n

1. Introduction

In this thesis, we apply considerations relating to the fields of survival analysis and causal inference with the aim to enable statistical reasoning based on resampling.

The chapter at hand highlights the relevance of our investigations in view of the current state of research, describes our objectives, and finally, gives a brief overview of the following chapters.

1.1. Motivation

Time-to-event (TTE) data account for a significant proportion of the data considered in clinical trials, especially in the field of oncology. Approaches that can handle right-censoring are therefore well-established, and yet, it appears that the underlying requirements w.r.t. censoring processes are not fully understood. Inconsistent definitions and unnecessarily deviating assumptions further contribute to the confusion (cf. O'Quigley, 2008, p. 122; Kleinbaum and Klein, 2012, p. 38). Andersen (2005) has provided an overview of the various censoring mechanisms encountered in practice, and Overgaard and Hansen (2021) discussed assumptions that are frequently imposed on them. Particular focus is placed on independent right-censoring (Andersen, Borgan, et al., 1993, p. 139), which, in short, ensures that the intensity w.r.t. the event of interest is not affected by the additional information conveyed through the censoring process. This condition is a necessary prerequisite for the application of the common analysis methods.

We concentrate on the specific case of event-driven trials for now, where the aim is to achieve a pre-specified number of observed events. Such study designs usually occur in combination with staggered study entry (cf. Elisei et al., 2013; McLaughlin et al., 2015; Sitbon et al., 2015; Husain et al., 2019; Baden et al., 2021). Aalen, Borgan, and Gjessing (2008, p. 59) showed that scenarios with simple type II censoring actually do satisfy the condition of independent censoring. However, successive entry times induce additional randomness due to the projection of the data onto the study time scale, which leads to uncertainty regarding the assumption of independent censoring, and accordingly, the validity of the usual analysis methodology.

It is clear, on the other hand, that event-driven trials involve dependent data, since administrative censoring is determined by the timing of the events. One should therefore refrain from using methods based on the assumption of random censoring, such as the nonparametric bootstrap proposed by Efron (1979) (cf. Singh, 1981; Friedrich, Brunner, and Pauly, 2017; Hrba et al., 2022). Whether this restriction is in fact taken into account

in situations where resampling is the only feasible method to draw statistical conclusions is questionable, though.

Beyond the context of classical TTE analysis, the combination of survival methodology with tools for causal inference has gained increasing importance in recent years. The resulting approaches can for example be used to identify causal treatment effects on TTE endpoints in observational studies, where the confounding factors are not distributed equally across treatment groups. We focus on effects quantified by the cumulative incidence function (CIF) in this work, as such characterizations permit both the analysis of competing risks settings as well as the circumvention of the issues associated with hazard ratios (HRs) in causal contexts (Hernán, 2010; Martinussen and Vansteelandt, 2013; Aalen, Cook, and Røysland, 2015). Causal effects like the one described here are often considered in cardiovascular trials (cf. eg. Lamberts et al., 2014; Stærk et al., 2017).

A commonly employed strategy to identify the average treatment effect (ATE) in a given study sample is to use the g-formula (Robins, 1986). To be more specific, this involves computing the standardized expectations of the outcome given both treatment levels and then forming their contrast. The distribution of the stochastic process that corresponds to the g-formula estimator of the causal CIF is rather complex, which is why statistical inference is usually based on resampling. In practice, researchers almost exclusively use Efron's nonparametric bootstrap (Lamberts et al., 2014; Stærk et al., 2017). It has been mentioned before that this approach is not optimal in settings that involve dependencies, though, and investigations about potential alternatives are scarce: Ozenne, Scheike, et al. (2020) have applied a resampling approach based on the influence function (IF) in their experimental study (cf. Scheike and Zhang, 2008). It further stands to reason that approximations based on the martingale-based wild bootstrap yield decent outcomes (cf. Lin, Wei, and Ying, 1993; Beyersmann, Di Termini, and Pauly, 2013; Dobler, Beyersmann, and Pauly, 2017).

Another way to approximate the ATE is to perform matching with replacement on the propensity score (PS) (i.e. the conditional probability to be treated given the confounder values). Only few investigations have addressed appropriate ways to assess the variability of the associated estimator in the presence of TTE data, as the special set-up of the matched data is hard to factor in. What is more, Abadie and Imbens (2008) demonstrated that the classical bootstrap is not suitable for inference on estimators obtained by matching with replacement. The weighted bootstrap approach proposed by Otsu and Rai (2017) does solve this issue, but disregards the variance that results from the estimation of the PSs. Other suggestions for methods that give insights into the distribution of PS-matched effect estimators for TTE data merely refer to the causal HR (Austin and Cafri, 2020; Adusumilli, 2022; Wang et al., 2024).

1.2. Contribution of this thesis

In this thesis, we establish that the censoring process underlying event-driven trials with staggered entry fulfils the condition of independent censoring in the counting process sense. This way, we justify the use of the common survival methodology for the analysis of such trials.

It turns out that the validity of independent censoring in the given context is subject to the exclusion of all information related to the calendar times of the events. We use simulations to illustrate the consequences in case this constraint is violated, and thereby, highlight potential sources of bias that should be avoided in practice.

Another empirical study moreover demonstrates the limitations of techniques based on random censoring against the backdrop of event-driven trials with staggered entry. For this purpose, we compare the results obtained by means of the classical bootstrap to those provided by the wild bootstrap (cf. Efron, 1979; Lin, Wei, and Ying, 1993).

Lastly, real data covering immunotherapy in cancer patients are analysed with the aim of exemplifying the extent to which the incorrect application of the classical bootstrap may impact analysis outcomes in practice (cf Rittmeyer et al., 2017).

The second part of the dissertation focuses on the exploration of different resampling methods for statistical inference about the ATE.

We first derive a martingale representation of the stochastic process that characterizes the g-formula estimator of the ATE, and on the basis thereof, we demonstrate that the classical bootstrap, a multiplier-based resampling approach building on the influence function, as well as the wild bootstrap are all suitable for approximating the asymptotic distribution of the estimated ATE (cf. Efron, 1979; Scheike and Zhang, 2008; Lin, Wei, and Ying, 1993). The corresponding proofs are based on arguments similar to those used by Cheng, Fine, and Wei (1998), Akritas (1986), and Dobler, Beyersmann, and Pauly (2017).

Extensive simulations are further conducted to examine the performance of these methods under varying conditions. We derive pointwise as well as time-simultaneous confidence regions and assess the associated coverage probabilities. By means of comparisons between the resampling techniques, we establish guidelines for their use in distinct scenarios.

The application of the investigated methods is finally illustrated considering real study data on the progression of Hodgkin's disease after treatment with radiation (Pintilie, 2006).

In order to gain insight into potential approaches allowing for inferences about the PS-matched ATE estimator, we further adapt the double-resampling technique according to

Wang et al. (2024) as well as the clustered variance estimator proposed by Austin and Cafri (2020) to the case where the ATE is determined by the CIF.

The performance of the resulting procedures is explored using the same simulation set-up as already considered w.r.t. the resampling approaches relating to the g-formula estimator. We evaluate the adequacy of the new methods based on the coverage probabilities of the generated confidence regions and identify conditions under which they tend to deteriorate.

Eventually, the Hodgkin's disease data are re-analysed to test how the PS matching-based approaches perform when applied to real world data.

1.3. Outline

The remainder of this dissertation is structured as outlined hereafter.

In Chapter 2, we introduce the theoretical background necessary for developing the ensuing ideas, which includes topics related to survival analysis (Section 2.1), causal inference (Section 2.2), and resampling (Section 2.3). The notation used throughout the thesis is set up in this context.

Chapter 3 addresses the dependence structure of the data in event-driven trials with staggered entry. After stating the theorem that verifies the condition of independent censoring in such trials (Section 3.1), we present the associated simulation studies in Section 3.2, and subsequently describe the analysis of the real data example considered in this context (Section 3.3).

Chapter 4 next investigates resampling-based inference for the ATE. In Section 4.1, we concentrate on the g-formula estimator and start by proving the validity of the distinct resampling methods for statistical inference on this parameter. The simulations comparing the respective approaches is presented afterwards, and the section closes with the analysis of the Hodgkin's disease data. The focus in Section 4.2 is then put on inference about the PS-matched ATE estimator, which necessitates an introduction to suitable methods for this purpose. We highlight two such methods, study their performance by means of simulations, and discuss the results. Lastly, we repeat the analysis of the Hodgkin's disease data applying the new methods for PS-matched data.

The dissertation eventually concludes with a summary, followed by a discussion of our findings and a brief outlook on future research.

All statistical analyses in this thesis are performed using the open source software R (version 4.4.1, R Core Team, 2024). The corresponding code can be accessed through the GitHub repository (https://github.com/jruehl).

2. Preliminary background

In this chapter, we introduce the theory that lays the foundation for the remainder of the thesis, including concepts of survival analysis, causal inference, and different resampling techniques.

2.1. Survival analysis

The target of inference for survival analysis is the *survival time*, i.e. the duration of the time period between a pre-specified origin and the occurrence of an event of interest. In clinical trials, the temporal origin is usually defined by a subject's entry into the study (e.g. at the time of randomization). The event of interest could for instance be death, but contrary to its name, survival analysis is not restricted to the evaluation of life spans only.

Standard statistical methods are inappropriate for the analysis of TTE data because in general, part of the survival times are not fully recorded. For individuals who have not experienced the event of interest by the end of the observation period, it is merely known that the true survival time is at least as large as the duration of the time under observation. The corresponding, incomplete data points are said to be *right-censored*. We distinguish between different types of right-censoring with varying implications: *Simple type I censoring* arises in experiments that end at a predetermined, deterministic point of time, such that all participants whose event happens afterwards are censored. If the end of a trial is determined by a fixed number of events to be observed, we speak of *simple type II censoring*. Beside administrative factors, right-censoring may also result from subject withdrawal or loss to follow-up.

Left- and, more generally, *interval-censoring* are less common kinds of censoring, addressing survival times that are only known to fall below an observed threshold, or to range within a specific interval, respectively. Sometimes, the observation of survival times is subject to additional constraints: In the presence of *left-truncation*, an individual's inclusion into the study depends on whether their event occurs after a specific incident, which marks the beginning of the observation period. *Right-truncation* can be defined analogously, but this condition is less relevant in practice.

To account for censored data while still including all the given information, survival analysis relies on specific methods, which we will introduce in this section. The presented theory is largely based on the books by Aalen, Borgan, and Gjessing (2008) and Andersen, Borgan, et al. (1993).

2.1.1. Basic concepts & counting process representation of time-to-event data

In the following, we consider the probability space $(\Omega, \mathscr{F}, P)$ and focus on the study time interval $[0, \tau]$ that lasts until the terminal time τ.

Let the absolutely continuous, non-negative random variables $T_1, \ldots, T_n$ be independent survival times with *survival function* $S_i(t) = P(T_i > t)$ and *hazard*

$$\alpha_i(t) = \lim_{\Delta t \searrow 0} \frac{P\big(T_i \in [t, t + \Delta t) \mid T_i \geq t\big)}{\Delta t},$$

respectively, for $i \in \{1, \ldots, n\}$, $n \in \mathbb{N}$. Due to practical reasons, we will mostly adopt the alternative notation $\alpha_i(t)\,\mathrm{d}t = P\big(T_i \in [t, t + \mathrm{d}t) \mid T_i \geq t\big)$ on the basis of the differential $\mathrm{d}t$ throughout this dissertation. It is easy to see that the survival function and the hazard rate are linked by the relation $S_i(t) = \exp\big(-A_i(t)\big)$, $i \in \{1, \ldots, n\}$, with (finite) *cumulative hazard function*

$$A_i(t) = \int_0^t \alpha_i(u)\,\mathrm{d}u.$$

Aalen (1978) significantly enhanced the theory underlying survival analysis by representing TTE data via counting processes. To signify whether the event of interest has happened by time t for subject $i \in \{1, \ldots, n\}$, define the counting process

$$\big(N_i^c(t) = \mathbb{1}\{T_i \leq t\}\big)_{t \in [0,\tau]}.$$

We use the expression $\mathbb{1}\{\cdot\}$ to refer to the indicator function of the event in the argument. Note also that the superscript 'c' is applied to emphasize that 'complete', i.e. uncensored, data are observed.

Since the past usually affects the subsequent behaviour of a stochastic process, it is helpful to establish a formalization of the *history* through right-continuous filtrations $(\mathscr{F}_t)_{t \in [0,\tau]}$ (henceforth denoted by $(\mathscr{F}_t)$, and likewise for processes), that is, increasing families of sub-σ-algebras of $\mathscr{F}$ fulfilling $\mathscr{F}_u = \bigcap_{t>u} \mathscr{F}_t$. The pre-$t$ σ-algebra $\mathscr{F}_{t-}$, informally, reflects the information available at time t about events (or the absence thereof) over the past interval $[0, t)$. Our focus w.r.t. the multivariate process $\big(\mathbf{N}^c(t) = \big(N_1^c(t), \ldots, N_n^c(t)\big)\big)$ is on the so-called self-exciting filtration $\big(\mathscr{F}_t^c\big)$ that is generated by the counting process itself:

$$\mathscr{F}_t^c = \sigma\Big(\big(\mathbf{N}^c(u)\big)_{u \leq t}\Big).$$

We say that $\big(\mathbf{N}^c(t)\big)$ is adapted to $\big(\mathscr{F}_t^c\big)$, which means simply that $\mathbf{N}^c(t)$ is $\mathscr{F}_t^c$-measurable for every $t \in [0, \tau]$.

Taking the history into account, the behaviour of $\big(N_i^c(t)\big)$ is characterized by its *intensity process* $\big(\lambda_i^{\mathscr{F}^c}(t)\big)$, $i \in \{1, \dots, n\}$, via

$$\lambda_i^{\mathscr{F}^c}(t)\,\mathrm{d}t = P\big(\mathrm{d}N_i^c(t) = 1 \mid \mathscr{F}_{t-}^c\big)\,.$$

The term $\mathrm{d}N_i^c(t)$ above refers to the (binary) increment of the counting process over the infinitesimal interval $[t, t + \mathrm{d}t)$, and hence, it follows that $\lambda_i^{\mathscr{F}^c}(t)$ is equal to the product of the hazard $\alpha_i(t)$ and $Y_i^c(t) = \mathbb{1}\{T_i \geq t\}$, the value of the *at-risk-process* at time t.

Next, we introduce the non-negative, possibly random censoring times $C_1, \dots, C_n$ in order to extend the described set-up by right-censoring. This leads to observed data of the form

$$\big((T_i \wedge C_i,\ D_i)\big)_{i \in \{1,\dots,n\}}\,.$$

The first expression in the tuple above represents the minimum of T_i and C_i, i.e. the time under observation, whereas the second entry, D_i denotes the censoring status $\mathbb{1}\{T_i \leq C_i\}$.

As some of the survival times are masked, we need to consider an adjusted version $\big(N_i(t)\big)$ of $\big(N_i^c(t)\big)$, $i \in \{1, \dots, n\}$, that jumps at observed survival times only:

$$N_i(t) = \int_0^t Y_i^o(u)\,\mathrm{d}N_i^c(u) = \mathbb{1}\{T_i \wedge C_i \leq t,\ D_i = 1\}. \tag{2.1}$$

Here, $Y_i^o(t) = \mathbb{1}\{C_i \geq t\}$ refers to the value of the left-continuous *censoring process* for subject i at time t. The summation over the entire cohort,

$$N(t) = \sum_{i=1}^{n} N_i(t),$$

yields the aggregated counting process $\big(N(t)\big)$, assuming $T_i \neq T_j\ \forall i \neq j$ or, in words, no *ties* (to which we adhere subsequently).

One can further formalize the observable past under right-censoring by the multivariate processes $\Big(\mathbf{N}(t) = \big(N_1(t), \dots, N_n(t)\big)\Big)$ and $\Big(\mathbf{Y}(t) = \big(Y_1(t), \dots, Y_n(t)\big)\Big)$, where the at-risk indicator

$$Y_i(t) = Y_i^c(t)\,Y_i^o(t) = \mathbb{1}\{T_i \wedge C_i \geq t\} \tag{2.2}$$

implies whether an event of subject $i \in \{1, \dots, n\}$ at time $t \in [0, \tau]$ is both possible and detectable. A more specific definition of the history is as follows:

$$\Big(\mathscr{F}_t = \sigma\Big(\big(\mathbf{N}(u), \mathbf{Y}(u)\big)_{u \leq t}\Big)\Big).$$

This filtration, essentially, reflects the knowledge about preceding events that is usually available in real-life studies, and it gives rise to the intensity $\big(\lambda_i^{\mathscr{F}}(t)\big)$ of $\big(N_i(t)\big)$, which is – similarly to the uncensored case – determined by

$$\lambda_i^{\mathscr{F}}(t)\,\mathrm{d}t = P\big(\mathrm{d}N_i(t) = 1 \mid \mathscr{F}_{t-}\big)\,.$$

Because the counting process $\big(N_i(t)\big)$ is a (local) submartingale w.r.t. the history $(\mathscr{F}_t)$, we obtain the following key result: One can split $N_i(t)$ uniquely into the sum of $\int_0^t \lambda_i^{\mathscr{F}}(u)\,\mathrm{d}u$ and

$$M_i(t) = N_i(t) - \int_0^t \lambda_i^{\mathscr{F}}(u)\,\mathrm{d}u$$

according to the (localized) Doob-Meyer decomposition (P.-A. Meyer, 1962; P.-A. Meyer, 1963), where $\big(\int_0^t \lambda_i^{\mathscr{F}}(u)\,\mathrm{d}u\big)$ is a predictable compensator and $\big(M_i(t)\big)$ defines a purely random, local square-integrable $(\mathscr{F}_t)$-martingale. By predictability of the compensator, it is meant that $\big(\lambda_i^{\mathscr{F}}(t)\big)$ is measurable w.r.t. the σ-algebra generated by all left-continuous, adapted processes.

The counting process representation at hand permits to exploit practical results of martingale theory for the analysis of survival data, provided that a vital condition explained in the subsequent paragraph is met.

To that end, consider a third, enlarged filtration, which combines the history $\big(\mathscr{F}_t^c\big)$ observed in the absence of censoring with details on the individual censoring processes, i.e.

$$\mathscr{G}_t = \sigma\Big(\mathscr{F}_t^c \,\cup\, \sigma\Big(\big(Y_1^o(u), \ldots, Y_n^o(u)\big)\Big)_{u\le t}\Big).$$

The definition above implies that $\big(\mathscr{F}_t^c\big)$ is nested within $(\mathscr{G}_t)$, and according to Equations (2.1) as well as (2.2), so is the history $(\mathscr{F}_t)$ that is observed under censoring. Thus, while $\big(N_i^c(t)\big)$ and $\big(N_i(t)\big)$ are $\big(\mathscr{F}_t^c\big)$- and $(\mathscr{F}_t)$-adapted, respectively, but not vice versa, both counting processes are adapted to $(\mathscr{G}_t)$. Similarly, $\big(Y_i^c(t)\big)$, $\big(Y_i^o(t)\big)$, as well as $\big(Y_i(t)\big)$ are $(\mathscr{G}_t)$-predictable for $i \in \{1, \ldots, n\}$. We now define a third intensity $\big(\lambda_i^{\mathscr{G}}(t)\big)$ w.r.t. $(\mathscr{G}_t)$ as follows:

$$\lambda_i^{\mathscr{G}}(t)\,\mathrm{d}t = P\big(\mathrm{d}N_i^c(t) = 1 \mid \mathscr{G}_{t-}\big)\,.$$

Definition 2.1 (Independent censoring)**:**
Censoring is said to be independent if

$$\lambda_i^{\mathscr{G}}(t) = \lambda_i^{\mathscr{F}^c}(t) \tag{2.3}$$

for all $i \in \{1, \ldots, n\}$ *and for all* $t \in [0, \tau]$.

The condition of *independent censoring* in Definition 2.1 ensures that the censoring process does not comprise any information that changes the intensity of $\big(N_i^c(t)\big)$. This assumption leads to a crucial finding: By the law of total expectation, we obtain the equality $\lambda_i^{\mathscr{F}}(t)\,\mathrm{d}t = \mathbb{E}\big(Y_i^o(t)\,\lambda_i^{\mathscr{G}}(t)\,\mathrm{d}t \mid \mathscr{F}_{t-}\big)$, and under independent censoring, it follows that $\lambda_i^{\mathscr{F}}(t)\,\mathrm{d}t = Y_i(t)\,\mathbb{E}\big(\alpha_i(t)\,\mathrm{d}t \mid \mathscr{F}_{t-}\big)$ since $\lambda_i^{\mathscr{F}^c}(t) = Y_i^c(t)\,\alpha_i(t)$. Throughout this thesis, we will assume that the hazard is $(\mathscr{F}_t)$-predictable so that $\lambda_i^{\mathscr{F}}(t) = Y_i(t)\,\alpha_i(t)$. One can therefore regard the course of events among the subjects at risk as representative of what would have happened without censoring.

Random censoring, i.e. stochastic independence $T_i \perp\!\!\!\perp C_i$ of the event and censoring times, is a fairly restrictive example of independent censoring. Besides, simple type I and type II censoring (where $\mathscr{F}_t^c = \mathscr{G}_t\ \forall t \in [0,\tau]$) as well as random left-truncation (with the 'censoring' process characterized by truncation) also fulfil Condition (2.3).

It should be noted that the term 'independent censoring' is not defined consistently in the literature (cf. O'Quigley, 2008, p. 122, Kleinbaum and Klein, 2012, p. 38; different concepts of independent censoring are discussed by Martinussen and Scheike, 2006, p. 52-57). We will adhere to the notion proposed by Andersen, Borgan, et al. (1993, Definition III.2.1) here. For further details on the properties of censoring processes, see also Andersen (2005).

Suppose now that the hazard is deterministic with $\alpha_i(t) = \alpha(t)\,\forall i \in \{1,\dots,n\}, t \in [0,\tau]$, and that censoring is independent. Then, $\big(N(t)\big)$ satisfies 'Aalen's multiplicative intensity model' (Aalen, 1978), that is,

$$\lambda^{\mathscr{F}}(t) = Y(t)\,\alpha(t),$$

with aggregated intensity process $\big(\lambda^{\mathscr{F}}(t)\big)$ defined by $\lambda^{\mathscr{F}}(t) = P\big(\mathrm{d}N(t) = 1 \mid \mathscr{F}_{t-}\big)$, and $\big(Y(t)\big)$ denoting the total number of subjects at risk at time $t \in [0,\tau]$, i.e.

$$Y(t) = \sum_{i=1}^{n} Y_i(t).$$

We consequently obtain a uniformly consistent estimator of the cumulative hazard $A(t)$, namely the *Nelson-Aalen estimator*

$$\hat{A}(t) = \int_0^t \frac{\mathbb{1}\{Y(u) > 0\}}{Y(u)}\,\mathrm{d}N(u)$$

(with the convention $0/0 \coloneqq 0$; Nelson, 1969; Nelson, 1972; Aalen, 1978) if we assume that $\inf_{u\in[0,t]} Y(u) \xrightarrow{P} \infty$ as $n \to \infty$ for $t \in [0,\tau]$.

Further large sample properties of the Nelson-Aalen estimator and many other quantities in survival analysis can be derived by means of the following version of Rebolledo's martingale central limit theorem (1980):

Theorem 2.1 (Central limit theorem for stochastic integrals w.r.t. counting processes)**:** *For each $n \in \mathbb{N}$ and $j \in \{1, \ldots, k\}$, with $k \in \mathbb{N}$ fixed, let $\big(N_j^{(n)}(t)\big)$ be a counting process defined on $[0, \tau]$ with intensity $\big(\lambda_j^{(n)}(t)\big)$, so that $N_j^{(n)}(t) = \int_0^t \lambda_j^{(n)}(u)\,\mathrm{d}u + M_j^{(n)}(t)$. Furthermore, let $\big(H_j^{(n)}(t)\big)$ be a locally bounded, predictable process.*
The conditions

(i) $\sum_{j=1}^{k} \int_0^t \big(H_j^{(n)}(u)\big)^2 \lambda_j^{(n)}(u)\,\mathrm{d}u \xrightarrow[n\to\infty]{P} V(t)\ \forall t \in [0, \tau]$,
for a deterministic, continuous, strictly increasing function V on $[0, \tau]$, with $V(0) = 0$

and

(ii) $\sum_{j=1}^{k} \int_0^t \big(H_j^{(n)}(u)\big)^2 \mathbb{1}\{\big|H_j^{(n)}(u)\big| > \epsilon\} \lambda_j^{(n)}(u)\,\mathrm{d}u \xrightarrow[n\to\infty]{P} 0\ \forall t \in [0, \tau],\ \epsilon > 0$

imply that on the Skorokhod space $D[0, \tau]$,

$$\sum_{j=1}^{k} \int H_j^{(n)}(u)\,\mathrm{d}M_j^{(n)}(u) \xrightarrow[n\to\infty]{\mathscr{D}} U,$$

where $\big(U(t)\big)$ is a Gaussian martingale defined by $U(t) = W\big(V(t)\big)$ and $\big(W(t)\big)$ denotes the Brownian motion.

The expression '$X_n \xrightarrow{\mathscr{D}} X$ on $D[0, \tau]$' in Theorem 2.1 indicates weak convergence of a sequence of processes $\big(X_n(t)\big)$ to $\big(X(t)\big)$, with the sample paths of $\big(X_n(t)\big)$ as well as $\big(X(t)\big)$ mapping $[0, \tau]$ to $\mathbb{R}$ and being right-continuous with left limits (càdlàg) (cf. Fleming and Harrington, 2005, Appendix B). Moreover, the Brownian motion $\big(W(t)\big)$ is a continuous process that is zero at time zero, has almost surely continuous sample paths as well as independent, normally distributed increments with mean zero and variance equal to the size of the respective increment.

Applying Theorem 2.1 to the Nelson-Aalen estimator, one finds that $\sqrt{n}\big(\hat{A}(\cdot) - A(\cdot)\big)$ converges weakly to a Gaussian martingale on $D[0, \tau]$ if there is a non-negative function y defined on $[0, \tau]$, with $\inf_{u\in[0,\tau]} y(u) > 0$, such that α/y is integrable over $[0, \tau]$ and $\sup_{u\in[0,\tau]} \big|Y(u)/n - y(u)\big| \xrightarrow{P} 0$ as $n \to \infty$. An estimator for the variance of $\hat{A}(t)$ is further obtained as

$$\widehat{\mathrm{Var}}\big(\hat{A}(t)\big) = \int_0^t \frac{\mathbb{1}\{Y(u) > 0\}}{\big(Y(u)\big)^2}\,\mathrm{d}N(u).$$

2.1.2. Cox proportional hazards regression

For the purpose of regression modelling, let $\mathbf{Z}_i \in \mathbb{R}^p$ denote a bounded vector of p baseline covariates for individual i, $i \in \{1, \ldots, n\}$, so that the survival times $T_1, \ldots, T_n$ are independent given $(\mathbf{Z}_1, \ldots, \mathbf{Z}_n)$. Consider the hazard $\alpha(t \mid \mathbf{z})$ that is defined conditional on the covariate values. We extend the histories $\left(\mathscr{F}_t^c\right)$ as well as $(\mathscr{F}_t)$ (and to that effect, $(\mathscr{G}_t)$) by information on $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ accordingly, and assume that Condition (2.3) applies, i.e. censoring may depend on the measured covariates. Under the premise of time-constant covariate effects on the multiplicative scale, it is convenient to study the following semiparametric model for the hazard at time $t \in [0, \tau]$ (Cox, 1972):

$$\alpha(t \mid \mathbf{z}) = \alpha_0(t) \exp\left(\boldsymbol{\beta}_0^T \mathbf{z}\right). \tag{2.4}$$

Model (2.4) characterizes $\alpha(t \mid \mathbf{z})$ by a non-negative, integrable baseline hazard $\alpha_0(t)$ and a vector $\boldsymbol{\beta}_0 \in \mathbb{R}^p$ of regression parameters. It follows that for $j \in \{1, \ldots, p\}$,

$$\frac{\alpha\big(t \mid (z_1, \ldots, z_{j-1}, z_j+1, z_{j+1}, \ldots, z_p)\big)}{\alpha\big(t \mid (z_1, \ldots, z_{j-1}, z_j, z_{j+1}, \ldots, z_p)\big)} = \exp\big(\beta_{0j}\big)\,,$$

i.e. the j^{th} element β_{0j} of $\boldsymbol{\beta}_0$ describes the logarithm of the HR between two observations whose covariate vectors only differ by one unit in component j.

The starting point for the estimation of the regression coefficients $\beta_{01}, \ldots, \beta_{0p}$ is the partial likelihood

$$\mathcal{L}(\boldsymbol{\beta}) = \prod_{t \in [0,\tau]} \prod_{i=1}^{n} \left(\frac{Y_i(t) \exp\big(\boldsymbol{\beta}^T \mathbf{Z}_i\big)}{\sum_{j=1}^{n} Y_j(t) \exp\big(\boldsymbol{\beta}^T \mathbf{Z}_j\big)} \right)^{\Delta N_i(t)}.$$

We define

$$\mathbf{S}^{(r)}(\boldsymbol{\beta}, t) = \frac{1}{n} \sum_{i=1}^{n} Y_i(t) \exp\Big(\boldsymbol{\beta}^T \mathbf{Z}_i\Big) \mathbf{Z}_i^{\otimes r},$$

for $\boldsymbol{z}^{\otimes 0} = 1$, $\boldsymbol{z}^{\otimes 1} = \boldsymbol{z}$, as well as $\boldsymbol{z}^{\otimes 2} = \boldsymbol{z}\boldsymbol{z}^T$, $r \in \{0, 1, 2\}$, and

$$\mathbf{E}(\boldsymbol{\beta}, t) = \frac{\mathbf{S}^{(1)}(\boldsymbol{\beta}, t)}{S^{(0)}(\boldsymbol{\beta}, t)}$$

as functions of a vector $\boldsymbol{\beta} \in \mathbb{R}^p$ and time $t \in [0, \tau]$. The vector of score functions w.r.t. $\mathcal{L}(\boldsymbol{\beta})$ is therefore given by

$$\frac{\partial}{\partial \boldsymbol{\beta}} \log \mathcal{L}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \int_0^{\tau} \big(\mathbf{Z}_i - \mathbf{E}(\boldsymbol{\beta}, t)\big)\, \mathrm{d}N_i(t).$$

Note that this expression reduces to $\sum_{i=1}^{n} \int_0^\tau \big(\mathbf{Z}_i - \mathbf{E}(\boldsymbol{\beta}_0, t)\big)\, \mathrm{d}M_i(t)$ if $\boldsymbol{\beta} = \boldsymbol{\beta}_0$, which is due to the Doob-Meyer decomposition and the definitions of $S^{(0)}$, $\mathbf{S}^{(1)}$, and $\mathbf{E}$ above.

We will for now suppose that the observations $\big((T_i \wedge C_i, D_i, \mathbf{Z}_i)\big)_{i \in \{1,\ldots,n\}}$ are independent and identically distributed (i.i.d.), with $T_i \perp\!\!\!\perp C_i \mid \mathbf{Z}_i$ and strictly positive probability $P(Y_i(t) = 1)\ \forall i \in \{1, \ldots, n\},\ t \in [0, \tau]$. Let

$$\mathbf{s}^{(r)}(\boldsymbol{\beta}, t) = \mathbb{E}\Big(\mathbf{S}^{(r)}(\boldsymbol{\beta}, t)\Big),$$

for $r \in \{0, 1, 2\}$, and

$$\mathbf{e}(\boldsymbol{\beta}, t) = \frac{\mathbf{s}^{(1)}(\boldsymbol{\beta}, t)}{s^{(0)}(\boldsymbol{\beta}, t)}.$$

One can show that there is a compact neighbourhood $\mathcal{B}$ of the true parameter vector $\boldsymbol{\beta}_0$ so that for $r \in \{0, 1, 2\}$ and $n \to \infty$, the function $\mathbf{S}^{(r)}$ converges uniformly to $\mathbf{s}^{(r)}$ in probability on the Cartesian product $\mathcal{B} \times [0, \tau]$. Besides, the limits $\mathbf{s}^{(r)}$ are continuous functions of $\boldsymbol{\beta} \in \mathcal{B}$ uniformly in $t \in [0, \tau]$, and on $\mathcal{B} \times [0, \tau]$, they are bounded, with $s^{(0)}$ being additionally bounded away from zero (Fleming and Harrington, 2005, Theorem 8.4.1). These findings now permit to conclude that the score function evaluated at $\boldsymbol{\beta}_0$ is asymptotically normally distributed, or more specifically,

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \int_0^\tau \big(\mathbf{Z}_i - \mathbf{E}(\boldsymbol{\beta}_0, t)\big)\, \mathrm{d}M_i(t) \xrightarrow{\mathscr{D}} \mathcal{N}(0, \boldsymbol{\Sigma}), \tag{2.5}$$

provided that the standardized information matrix $\boldsymbol{\Sigma} = -\mathbb{E}\Big(\frac{1}{n} \cdot \frac{\partial^2}{\partial \boldsymbol{\beta}^2} \log \mathcal{L}(\boldsymbol{\beta})|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}\Big)$, with

$$\boldsymbol{\Sigma} = \int_0^\tau \left(\frac{\mathbf{s}^{(2)}(\boldsymbol{\beta}_0, t)}{s^{(0)}(\boldsymbol{\beta}_0, t)} - \big(\mathbf{e}(\boldsymbol{\beta}_0, t)\big)^{\otimes 2} \right) s^{(0)}(\boldsymbol{\beta}_0, t)\, \alpha_0(t)\, \mathrm{d}t, \tag{2.6}$$

is positive definite (Andersen and Gill, 1982). The proof relies on the martingale central limit theorem presented at the end of the previous section.

Andersen, Borgan, et al. (1993, Theorem VII.2.1) showed that as $n \to \infty$, the probability that there is a unique vector $\hat{\boldsymbol{\beta}}$ maximizing the partial likelihood $\mathcal{L}(\boldsymbol{\beta})$ tends to one, and if it exists, this vector is consistent. The asymptotic normality of the score statistic in Equation (2.5) further entails that $\sqrt{n}\big(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\big)$ is asymptotically Gaussian with mean zero and covariance matrix $\boldsymbol{\Sigma}^{-1}$. In practice, one usually resorts to iterative methods like the Newton-Raphson algorithm to approximate the value of the Cox estimator $\hat{\boldsymbol{\beta}}$.

It is worth mentioning that the results above can also be obtained under relaxed assumptions (cf. Andersen, Borgan, et al., 1993, Theorem VII.2.1 and VII.2.2), but we will later rely on several findings that are based on the independence constraints considered here. For example, Tsiatis (1981) showed that if the data are i.i.d., $T_i \perp\!\!\!\perp C_i \mid \mathbf{Z}_i$, and

$P\big(Y_i(t) = 1\big) > 0 \; \forall i \in \{1, \ldots, n\}$, $t \in [0, \tau]$, then the Cox estimator converges almost surely to the true parameter vector.

Apart from $\boldsymbol{\beta}_0$, another unknown component in Model (2.4) is the baseline hazard $\alpha_0(t)$, i.e. the hazard if all predictors are equal to zero. While we do not need to know this function to make inference about the HR, an estimate of $\alpha_0(t)$ is yet necessary to examine other functionals of the hazard rate subject to the Cox model. Breslow (1972) proposed the subsequent estimator of the cumulative baseline hazard $A_0(t) = \int_0^t \alpha_0(u) \, \mathrm{d}u$ for a given vector $\hat{\boldsymbol{\beta}}$:

$$\hat{A}_0(t) = \int_0^t \frac{\mathbb{1}\{Y(u) > 0\}}{n \, S^{(0)}\big(\hat{\boldsymbol{\beta}}, u\big)} \, \mathrm{d}N(u).$$

(Note the parallel to the Nelson-Aalen estimator.) The strong uniform consistency of $\hat{A}_0$ can be demonstrated under the same assumptions as above (Lopuhaä and Nane, 2013), and consequently, $\hat{A}(t \mid \mathbf{z}) = \hat{A}_0(t) \exp\big(\hat{\boldsymbol{\beta}}^T \mathbf{z}\big)$ approximates the cumulative hazard for an individual with covariate vector $\mathbf{z}$. It will be convenient later on to represent $\hat{A}(t \mid \mathbf{z})$ by means of martingales in order to investigate large sample properties of hazard-based processes. Lin, Fleming, and Wei (1994) proceeded similarly as Andersen, Borgan, et al. (1993, proof of Theorem VII.2.3), by exploiting the weak convergence in Equation (2.5), to show that $\sqrt{n}\Big(\hat{A}(t \mid \mathbf{z}) - A(t \mid \mathbf{z})\Big)$ is asymptotically equivalent to

$$\frac{1}{\sqrt{n}} \int_0^t \frac{\exp\big(\boldsymbol{\beta}_0^T \mathbf{z}\big)}{S^{(0)}(\boldsymbol{\beta}_0, u)} \, \mathrm{d}M(u) + \frac{1}{\sqrt{n}} \big(\mathbf{h}(t \mid \mathbf{z})\big)^T \, \boldsymbol{\Sigma}^{-1} \sum_{i=1}^{n} \int_0^\tau \big(\mathbf{Z}_i - \mathbf{E}(\boldsymbol{\beta}_0, u)\big) \, \mathrm{d}M_i(u),$$

with

$$\mathbf{h}(t \mid \mathbf{z}) = \int_0^t \big(\mathbf{z} - \mathbf{e}(\boldsymbol{\beta}_0, u)\big) \, \mathrm{d}A(u \mid \mathbf{z}).$$

As noted earlier, the Cox model depends on the rather strong assumption of proportional hazards. One possibility to check whether covariate effects are in fact time-constant is to test for a correlation between the scaled Schoenfeld residuals and the (possibly transformed) failure times. Under proportional hazards, we expect the correlation to be zero (Grambsch and Therneau, 1994). Besides, a graphical display of the estimated coefficients in a model with time-varying effects should not reveal any pattern over time. If there is evidence that the proportional hazards assumption is violated for some of the covariates, one might alternatively consider an additive hazards model (Aalen, 1989) or a combination in the form of an additive-multiplicative Cox-Aalen model (Scheike and Zhang, 2002).

2.1.3. Competing risks

So far, the focus of our considerations has been on the setting where subjects may only experience one single type of event. A more general framework (which includes the previously described situation as a special case) builds upon competing risks models. This concept is suitable if the occurrence of a certain event is prevented by prior incidents of other, 'competing' events.

Suppose that the data take the form $\big((T_i \wedge C_i, D_i)\big)_{i\in\{1,\dots,n\}}$, where T_i now denotes the time until the first event (regardless of its cause) and the indicator $D_i \in \{0, 1, \dots, K\}$ specifies the corresponding event type. (A value of 0 marks censored observations.) The cause-specific hazard rate of cause $k \in \{1, \dots, K\}$ is defined by

$$\alpha_k(t)\,\mathrm{d}t = P\big(T \in [t, t+\mathrm{d}t),\, D{=}k \mid T \geq t\big)\,.$$

We capture the number of observed type k events until time $t \in [0, \tau]$ as the sum over $N_{ki}(t) = \mathbb{1}\{T_i \wedge C_i \leq t,\, D_i{=}k\}$ for $i \in \{1, \dots, n\}$, which results in the counting process $\big(N_k(t)\big)$. Extending the history $(\mathscr{F}_t)$ by the past values of all cause-specific counting processes, the intensity $\big(\lambda_k^{\mathscr{F}}(t)\big)$ is determined analogously to the standard setting. This gives rise to the martingale $\big(M_k(t)\big)$, with $M_k(t) = N_k(t) - \int_0^t \lambda_k^{\mathscr{F}}(u)\,\mathrm{d}u$. An estimator of the cause-specific cumulative hazard $A_k(t) = \int_0^t \alpha_k(u)\,\mathrm{d}u$ is further obtained as in Subsection 2.1.1: Under the assumption that the multiplicative intensity model holds, we approximate $A_k(t)$ by $\hat{A}_k(t) = \int_0^t \mathbb{1}\{Y(u) > 0\}/Y(u)\,\mathrm{d}N_k(u)$, where $Y(t)$ now excludes subjects with *any* event or censoring prior to t. Similarly, one may fit a cause k specific Cox regression model with given baseline covariates, while treating events of type $\tilde{k} \neq k$ as censored. Large-sample inference is then based on a multivariate version of the martingale central limit theorem.

Even though the approach above allows for consistent estimation of the cause-specific cumulative hazard, it is in general not appropriate to simply censor individuals with competing events and adopt the relations that apply in the standard survival setting. To see why, note that the survival probability $S(t) = \exp\big(-\sum_{k=1}^K A_k(t)\big)$ depends on all K cause-specific hazards if there are multiple event types. The one-to-one correspondence between $\alpha_k(t)$ and the risk of experiencing a type k event until t, as quantified by the CIF

$$F_k(t) = P(T \leq t,\, D = k) = \int_0^t S(u-)\,\alpha_k(u)\,\mathrm{d}u,$$

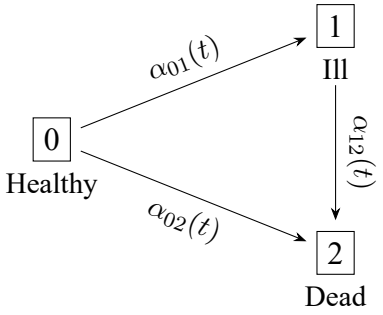
is therefore no longer valid, and the naive estimator $1 - \exp\big(-\hat{A}_k(t)\big)$ will generally suffer from upward bias. In particular, one may obtain values greater than 1 by summing over all causes (which is inconsistent with the equivalence $\sum_{k=1}^K F_k(t) = P(T \leq t)$).

The risk according to this biased estimator pertains to a hypothetical setting where competing causes are eliminated so that everyone will experience a type k event eventually, but in reality, subjects who failed from other events are no longer exposed to the cause of interest (Andersen, Geskus, et al., 2012). For the analysis of competing risks data, one should rather evaluate all cause-specific hazards (and, ideally, CIFs) to gain a full understanding of the covariate effects on each type of event and the interactions between them.

Fine and Gray (1999) suggested to model a 'subdistribution hazard' analogously to Equation (2.4), but keeping subjects with competing events at risk for the cause of interest. This way, the one-to-one correspondence between subdistribution hazard and CIF is maintained. The regression coefficients in the subdistribution model relate to $F_k(t)$ instead of $\alpha_k(t)$, $k \in \{1, \ldots, K\}$, and thus quantify the effect of the covariates on the CIF, but lack an intuitive interpretation other than that. (Note that associations between covariates and competing events also contribute to the model parameters for the cause of interest.) In addition, the sum over all estimated CIFs may exceed the value of 1 when separate Fine and Gray models are fitted for each cause (Austin, Steyerberg, and Putter, 2021).

More complex situations than those discussed so far can be investigated, too, by considering multistate models with transient states and imposing the time-inhomogeneous Markov property. In fact, standard survival and competing risks scenarios are only special cases of multistate models, and the corresponding analysis methods can be extended rather straightforwardly (see e.g. Putter, Fiocco, and Geskus, 2007, Section 4; Beyersmann, Allignol, and Schumacher, 2012, Part III). An example of a more general multistate model is the illness-death model without recovery (see Figure 2.1). Here, individuals might move to the absorbing state either directly or after a sojourn in an intermediate state.

Figure 2.1: Illness-death model without recovery.



2.2. Causal inference

Regardless of whether survival or another endpoint is examined, clinical trials mostly aim to identify causal relationships between some kind of intervention and the outcome of interest. Knowledge of the underlying causal connections helps to better understand the mechanisms by which the intervention operates on the outcome, possibly facilitating treatment decisions in the future. However, standard statistical methods measure association rather than causation, and their naive application bears the risk of erroneously identifying a causal link between variables that is really just due to confounding. A condition that yet allows to approximate the causal effect using common estimators is randomization. By arbitrary allocation of participants to treatment groups, any confounders – whether they are known to the researcher or not – are expected to be distributed equally across the groups, so that association and causation between treatment and outcome may be regarded as equivalent. Randomization is not always possible, though: It would for instance be unethical to investigate the effect of a harmful substance by deliberately administering it to part of the subjects. Besides, the exposure of interest is often beyond the researchers' control. One has to rely on observational studies in such cases, which means that there is a high chance of unsolicited factors being associated with both the exposure and the outcome, and side-by-side comparisons between treatment groups will consequently yield biased estimators of the (direct) treatment effect.

The framework of causal inference provides tools to eliminate such bias under certain identifiability conditions. In addition, there are ways to improve the analysis of randomized trials, e.g. by accounting for non-compliance.

A variety of causal techniques has been proposed during recent years (see e.g. Hernán and Robins, 2020), but for the sake of brevity, our focus lies only on those relevant within this thesis.

We will adhere to the counterfactual approach to causal inference in the following (Rubin, 1974): In summary, one aims to compare the potential outcomes that would have occurred under different exposures, but otherwise identical conditions. It is only possible to observe one of these potential outcomes for each individual, though (i.e. the one corresponding to the exposure they were actually subject to, as opposed to the potential outcomes reflecting settings that are 'counter to the fact'). Due to this 'fundamental problem of causal inference', one tries to estimate the mean potential outcomes among the whole study population. This leads to the subsequent definition of the *ATE* on the additive scale, considering the outcome O and comparing exposures a and a':

$$ATE = \mathbb{E}\big(O^{Z_A=a}\big) - \mathbb{E}\big(O^{Z_A=a'}\big).$$

The expression $O^{Z_A=a}$, or O^a in short, represents the potential outcome if the exposure status Z_A was set to a. (As of now, let Z_A be the first entry of the covariate vector $\mathbf{Z}$, followed by other measured covariates $\mathbf{Z}_L$.) We will assume throughout that $Z_A \in \{0, 1\}$ is a binary indicator for a single-point treatment, so that the ATE is given by $\mathbb{E}(O^1) - \mathbb{E}(O^0)$.

2.2.1. Identifiability conditions, standardization & propensity score matching

It has already been hinted that several assumptions need to hold to allow for the identification of the ATE. These assumptions are briefly explained below (see e.g. Hernán and Robins, 2020, Section I.3, for more details):

(Conditional) exchangeability ensures that there are no unmeasured confounders. Given the covariate vector $\mathbf{Z}_L$, the mean outcome among the treated subjects is therefore equal to the mean counterfactual outcome among the untreated subjects under treatment, and vice versa. More formally, conditional exchangeability implies that

$$O^a \perp\!\!\!\perp Z_A \mid \mathbf{Z}_L,\ a \in \{0, 1\}. \tag{2.7}$$

If the independence above applies marginally, we speak of unconditional exchangeability.

Positivity holds if it is both possible to be treated or untreated, respectively, conditional on every covariate combination that may occur in the target population:

$$P\big(Z_A = 1 \mid \mathbf{Z}_L = \mathbf{l}\big) \in (0, 1)\ \forall \mathbf{l} : P(\mathbf{Z}_L = \mathbf{l}) > 0. \tag{2.8}$$

This assumption is necessary to enable comparisons between treatment groups while taking confounders into account.

Lastly, *consistency* warrants that the potential outcome under treatment coincides with the observed outcome among treated subjects, just like the potential outcome under no treatment matches the observed outcome among untreated individuals. Formally, we write

$$O^a = O \text{ if } Z_A = a,\ a \in \{0, 1\}. \tag{2.9}$$

It seems obvious that this condition is met, but consistency depends on the precise definition of the potential outcomes via (sufficiently) well-defined treatment levels that correspond to the treatment levels actually pursued, and if there are any discrepancies between the potential and actual treatment levels, they must not affect the outcome. Consistency is also related to the condition of *no interference*, which ensures that an individual's treatment (or the absence thereof) does not act on another individual's potential outcome (as it may be the case e.g. in studies of infectious diseases).

Exchangeability, positivity, and consistency are referred to as *identifiability conditions*. In (ideal) randomized trials, we expect them to be fulfilled, with exchangeability and positivity applying even marginally. As a consequence, association measures obtained from randomized samples are consistent estimates of causal effect measures. If an observational study suffices the identifiability conditions, it may be treated as if it was conditionally randomized given the covariates $\mathbf{Z}_L$, which means that the causal effect can be derived using methods that will be introduced hereafter. An empirical verification of the identifiability conditions is generally not possible, though, and to assess their plausibility, one needs to rely on expert knowledge about potential confounders. It is recommended to explicitly state any assumptions made during the causal analysis, so that the outcome is open to critical scrutiny. There are certain situations that permit causal inference under violations of the identifiability conditions (Pearl, 1995; Angrist, Imbens, and Rubin, 1996), but the corresponding approaches are based on alternative, fairly strong assumptions.

Suppose that Conditions (2.7), (2.8), as well as (2.9) are fulfilled. Then one finds that for $a \in \{0, 1\}$,

$$\begin{aligned}
\mathbb{E}(O^a) \;&=\; \int \mathbb{E}\big(O^a \mid \mathbf{Z}_L = \mathbf{l}\big)\; \mathrm{d}F_{\mathbf{Z}_L}(\mathbf{l}) \\
&\overset{\substack{(2.7),\\(2.8)}}{=} \int \mathbb{E}\big(O^a \mid Z_A = a, \mathbf{Z}_L = \mathbf{l}\big)\; \mathrm{d}F_{\mathbf{Z}_L}(\mathbf{l}) \\
&\overset{(2.9)}{=} \int \mathbb{E}\big(O \mid Z_A = a, \mathbf{Z}_L = \mathbf{l}\big)\; \mathrm{d}F_{\mathbf{Z}_L}(\mathbf{l}),
\end{aligned}$$

where $F_{\mathbf{Z}_L}$ is the joint cumulative distribution function (CDF) of the covariate vector $\mathbf{Z}_L$ and the integral is taken over its domain. This equivalence is an instance of the so-called *g-formula* (Robins, 1986), and it gives rise to the *direct standardization* estimator

$$\widehat{ATE}_{ds} = \frac{1}{n} \sum_{i=1}^{n} \Big(\widehat{\mathbb{E}}\big(O \mid Z_A = 1, \mathbf{Z}_L = \mathbf{Z}_{Li}\big) - \widehat{\mathbb{E}}\big(O \mid Z_A = 0, \mathbf{Z}_L = \mathbf{Z}_{Li}\big) \Big),$$

letting $\mathbf{Z}_{Li}$ denote the vector of covariate values for subject $i \in \{1, \ldots, n\}$. If the conditional expectations in the expression above are estimated nonparametrically, the same estimate of the ATE may be obtained by *inverse probability of treatment (IPT) weighting*, i.e. by assigning the weight $1/\hat{P}(Z_A = Z_{Ai} \mid \mathbf{Z}_L = \mathbf{Z}_{Li})$ – or, for reduced variability, the stabilized weight $\hat{P}(Z_A = Z_{Ai})/\hat{P}(Z_A = Z_{Ai} \mid \mathbf{Z}_L = \mathbf{Z}_{Li})$ – to individual $i \in \{1, \ldots, n\}$ whose actual treatment indicator is Z_{Ai}, respectively, and subtracting the (weighted) mean of the outcomes among the controls from that among the treated. The pseudo-population generated by IPT weighting suffices for unconditional exchangeability, given

that exchangeability conditional on $\mathbf{Z}_L$ holds in the original sample, because the weights balance the confounders in $\mathbf{Z}_L$ over the treatment groups. It thus follows that the difference of the weighted means of the outcomes is consistent for the ATE (Hernán and Robins, 2020, Section I.2).

In case of high-dimensional data or continuous covariates $\mathbf{Z}_L$, we have to estimate $\mathbb{E}\big(O \mid Z_A = a, \mathbf{Z}_L = \mathbf{l}\big)$ and $P\big(Z_A = a \mid \mathbf{Z}_L = \mathbf{l}\big)$ based on models (that is, a so-called 'Q model' for the expected outcome and, usually, a logistic regression model for the treatment probability). The corresponding standardized and IPT-weighted estimates of the ATE will most likely deviate from each other although both aim to approximate the g-formula specified above. This is because they rely on different model assumptions, and model misspecification is unavoidable up to a certain degree. One should ideally obtain similar results for both estimators, however. An alternative option are doubly-robust estimators of the ATE that are consistent if either of the two models is correctly specified (Robins, Rotnitzky, and Zhao, 1994; van der Laan and Rubin, 2006).

The treatment model applied to calculate the IPT weights also comes into play with another method for the estimation of the ATE. Let

$$PS(\mathbf{Z}_{Li}) = P\big(Z_A = 1 \mid \mathbf{Z}_L = \mathbf{Z}_{Li}\big)$$

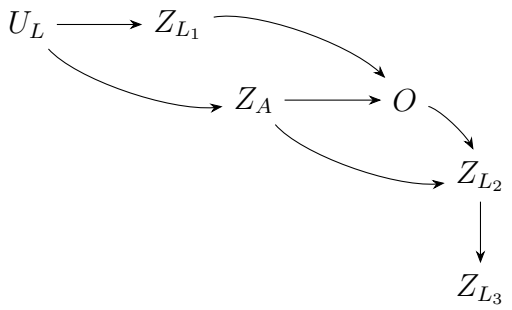
denote the *PS* of subject $i \in \{1, \ldots, n\}$, i.e. the probability of treatment given i's covariate values. As mentioned before, estimates of the PSs can be obtained by logistic regression. It has proven helpful to compare histograms of $\widehat{PS}(\mathbf{Z}_{Li})$ among treated and untreated individuals, since substantial disparities in the supports indicate positivity violations. On the other hand, equally distributed PS in both treatment groups suggest that there is no confounding by the covariates in $\mathbf{Z}_L$. The estimation of the PSs is further worthwhile because the covariates are compressed into a single value that maintains identifiability: By the law of total expectation, exchangeability and positivity conditional on $\mathbf{Z}_L$ remain valid conditional on the PS (Rosenbaum and Rubin, 1983), and thus, the outcomes of treated and untreated individuals can be directly compared given $PS(\mathbf{Z}_{Li})$. Note that the PS should be defined w.r.t. covariates that are expected to ensure exchangeability rather than accurate prediction of the treatment group, or else, the resulting estimate of the ATE may suffer from bias or high variance.

Matching on the PS creates a new sample that includes pairs (or, with many-to-one matching, larger sets) of subjects with similar PSs but different treatment levels. This way, the measured covariates are equally distributed across treatment groups, and after assigning a match to each individual, we estimate the ATE as the difference between the mean outcomes of treated and untreated individuals in the matched population (sup-

posing that positivity holds). Exact matches w.r.t. continuous variables are unlikely, though, so that it is necessary to define a measure of proximity between PSs. One usually considers the absolute difference between the estimated PSs. For the estimation of the ATE in the entire population, nearest-neighbour matching with replacement is commonly employed as it generates sufficient eligible matches for each (treated or untreated) individual (Abadie and Imbens, 2006; Otsu and Rai, 2017). It is possible to match without replacement if the aim is to identify the average treatment effect in the treatment group with fewer individuals. While this facilitates variance estimation, the effect estimate may be biased, though, due to the risk of incomplete matching (Austin and Cafri, 2020).

Causal analyses often benefit from a conceptualization of the underlying assumptions by means of *directed acyclic graphs (DAGs)*. A visual display of the potential causal relations between variables may reveal hidden sources of bias and thus, direct researchers towards suitable evaluation strategies.

Figure 2.2: Directed acyclic graph.



DAGs consist of nodes representing random variables and directed edges between them. As implied by the name 'DAG', the included edges must not form cycles. We connect nodes by edges to depict conjectured causal effects, whereas the lack of an edge that links two nodes encodes the assumption of no (direct) causal relation between them. The DAG in Figure 2.2 suggests for example that variable Z_A has a causal effect on variable O, and in that case, Z_A is referred to as a 'parent' of O. If we follow the edges in a DAG – possibly against the indicated direction – from one node to another, such that any node along the way is passed only once, we obtain a 'path' between the first and the last node. There are multiple paths from Z_A to O in Figure 2.2, namely the direct one ($Z_A \rightarrow O$) as well as the indirect routes via $Z_A \leftarrow U_L \rightarrow Z_{L_1} \rightarrow O$ and $Z_A \rightarrow Z_{L_2} \leftarrow O$. The nodes that can be reached by a directed path from a given variable are called its 'descendants': Consider e.g. variable O with descendants Z_{L_2} and Z_{L_3}. By these definitions, we may eventually express the main premise underlying DAGs: The causal Markov assumption states that a variable in a DAG is independent of all other variables, except for its descendants, if we condition on its parents. Consequently, any common parents of two variables – whether they are measured or not – need to be included in the DAG (Hernán and Robins, 2020, Section I.6).

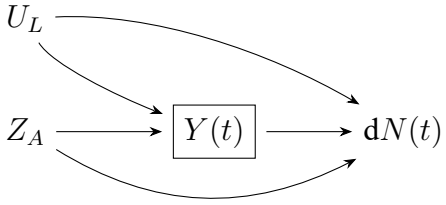
The use of DAGs for the identification of causal effects is now attributed to the fact that such graphs do not only disclose causation but also association: Two nodes in a DAG are associated if there is an edge between them, or if they share common causes. In order to determine the causal effect of Z_A on O in the setting illustrated in Figure 2.2, we therefore need to eliminate the spurious association between Z_A and O that is due to the (unmeasured) common cause U_L. More generally, Pearl (1988) and Pearl (1995) developed a framework for the probabilistic theory underlying DAGs, which implied that the direct effect between two variables is identified by 'blocking' all non-direct paths between them. A path can be blocked if one conditions on a variable that builds either a 'chain' or a 'fork' on that path. Here, chains are structures of the form '$\rightarrow Z_L \rightarrow$' or '$\leftarrow Z_L \leftarrow$', whereas forks correspond to the pattern '$\leftarrow Z_L \rightarrow$'. If a path includes a sub-structure like '$\rightarrow Z_L \leftarrow$', it is blocked inherently, and one should refrain from unblocking it by conditioning on the 'collider' Z_L or any descendants thereof. The direct effect of Z_A on O in our example may thus be determined by conditioning on the variable Z_{L_1}, which forms a chain on the path $Z_A \leftarrow U_L \rightarrow Z_{L_1} \rightarrow O$, but one may neither condition on the collider Z_{L_2} nor its descendant Z_{L_3}. For a more thorough introduction to causal DAGs, we refer to Greenland, Pearl, and Robins (1999) and Hernán and Robins (2020, Section I.6).

2.2.2. Causal inference for time-to-event data

When causal inference is to be drawn for TTE data, one should be aware of certain issues related to the use of the HR as effect measure. Martinussen and Vansteelandt (2013) pointed out that, unlike risk differences and risk ratios, HRs are non-collapsible, i.e. the marginal HR may differ from the conditional HR given an independent risk factor for the event of interest, even if there is no confounding. This means that the causal HR cannot be expressed as a (weighted) average of the HRs across strata. In a trial that satisfies the identifiability conditions and has unlimited sample size, the treatment effect as measured by the HR may still vary if one adjusts for a risk factor that is not associated with the event of interest at all. Besides, identification of potential confounders via HRs is prone to errors (Daniel, Zhang, and Farewell, 2021).

Hernán (2010) further warned about the so-called 'built-in selection bias' of the HR, which is illustrated by the DAG in Figure 2.3: The hazard rate at time t is defined conditional on survival up to t, and since the at-risk status $Y(t)$ is a collider on the path $Z_A \rightarrow Y(t) \leftarrow U_L \rightarrow \mathrm{d}N(t)$, a spurious association is introduced be-

Figure 2.3: Selection bias of the HR.



tween Z_A and $\mathrm{d}N(t)$. In other words, exchangeability will fade away over time among the subjects at risk if treatment does actually affect survival, because a higher proportion of 'frail' individuals is removed from the risk set in the group with lower chances of survival (see also Aalen, Cook, and Røysland, 2015; Post, van den Heuvel, and Putter, 2024). One could however argue in favour of the HR by pointing out that the interest is generally in the effect of treatment among the complete study population over the entire study period (as quantified by the HR) instead of the effect of treatment at a single time point t among subjects at risk at t. (Note that Figure 2.3 only represents a 'snapshot' of the causal relations at a given time point.) Nevertheless, HRs cannot reflect any changes in the ATE over time.

While Cox regression may be used as a tool to obtain meaningful causal effect estimates, the interpretation of the causal treatment effect should not be based directly on the HR, taking all the mentioned concerns into account (Martinussen and Stensrud, 2023). Aalen, Cook, and Røysland (2015) and Martinussen and Vansteelandt (2013) proposed to consider the difference between the hazards in an additive model. Because of the intuitive understanding of the risk, we define the ATE at time $t \in [0, \tau]$ w.r.t. the CIF, which means that competing risks settings are covered on top:

$$ATE(t) = F_1^{Z_A=1}(t) - F_1^{Z_A=0}(t). \tag{2.10}$$

Here, $F_k^a(t) = P(T^a \leq t, D^a = k)$ denotes the potential CIF for cause $k \in \{1, \ldots, K\}$, $a \in \{0, 1\}$, and the interest is w.l.o.g. on type 1 events. The g-formula can for instance be applied to approximate $ATE(t)$, provided that the identifiability conditions (as formulated w.r.t. the potential outcome $O^a = \mathbb{1}\{T^a \leq t, D^a = 1\}$) are met (see Section 4.1).

In the presence of competing risks, the ATE defined in Equation (2.10) summarizes the direct influence of the therapy on the time to the event of interest as well as any impact that is due to advancing or preventing competing events. Thus, $ATE(t)$ measures the total effect on the investigated cause. In contrast, the direct effect refers to the influence of the therapy in a hypothetical setting where the occurrence of any competing event is rendered impossible (Young et al., 2020). Direct effects may improve our understanding of the mechanisms by which the treatment acts on the event of interest, but they are often of minor interest in clinical trials, since interventions that eliminate competing events are generally not realistic. Stensrud et al. (2022) and Martinussen and Stensrud (2023) suggested possible ways to separate direct from indirect effects on the basis of untestable assumptions.

2.3. Resampling

One of the central tasks of statistical analyses is to evaluate the accuracy of a given estimate. For this purpose, we consider pointwise confidence intervals (CIs) and – in case of time-dependent estimators – time-simultaneous confidence bands (CBs): It is expected that $(1-\alpha)\cdot 100\%$ of the $(1-\alpha)\cdot 100\%$ CIs computed in a series of trials conducted under equivalent conditions will cover the true point estimand. Similarly, $(1-\alpha)\cdot 100\%$ of the CBs at level $(1-\alpha)$ should include the true estimand over an entire, pre-specified time interval. The derivation of valid CIs and CBs can be difficult, however, since many estimators have rather complex distributions and the central limit theorem only provides a poor approximation in case of small sample sizes.

As a remedy, one may resort to *resampling*, that is, repeated generation of subsamples from the given data sample. These subsamples can then be used to infer statistical properties of the estimator at hand. Resampling requires hardly any distributional assumptions, and is used for various applications apart from the construction of confidence regions, e.g. for hypothesis tests or in the context of machine learning. Depending on how the subsamples are created, there are different approaches to resampling, including permutation and bootstrapping (Chernick, 2012). We will focus on the latter hereafter.

2.3.1. Efron's nonparametric bootstrap

The resampling method that is most commonly applied to derive CIs is the nonparametric bootstrap introduced by Efron (1979), subsequently abbreviated as EBS. For an i.i.d. sample $(X_1, \ldots, X_n))$ of n observations and a statistic $\hat{\theta}$ that approximates the parameter θ, the idea is to draw many times with replacement (say $B = 1,000$ times), creating B subsamples $\left(X_1^{*(b)}, \ldots, X_n^{*(b)}\right)$ of size n ($b \in \{1, \ldots, B\}$), and to determine the statistic of interest in each of these bootstrap samples. One thus generates B bootstrap statistics $\hat{\theta}^{*(1)}, \ldots, \hat{\theta}^{*(B)}$ based on the empirical distribution given the original data. In fact, this approach is equivalent to repeatedly computing the weighted statistic after assigning multinomial $\mathcal{M}\left(n, (1/n, \ldots, 1/n)^T\right)$-distributed weights. Provided that the distribution of the bootstrap statistics conditional on the data approximates the sampling distribution, we may eventually construct an asymptotic $(1-\alpha)\cdot 100\%$ CI for θ by using the $\alpha/2$ and $1-\alpha/2$ sample quantiles of the bootstrap statistics as limits.

The *functional delta method* yields a condition that ensures the validity of the classical bootstrap for a particular statistic (van der Vaart and Wellner, 1996, Chapter 3.9; Kosorok, 2008, Chapter 12). Before presenting the corresponding theorem, we need to introduce a suitable concept of a directional derivative: A mapping $\theta : \mathbb{D} \to \mathbb{E}$ between normed spaces $\mathbb{D}$ and $\mathbb{E}$ is said to be *Hadamard differentiable* at $G \in \mathbb{D}$ tangentially to

$\mathbb{D}_0 \subseteq \mathbb{D}$ if there exists a continuous linear map $\partial\theta_G : \mathbb{D}_0 \to \mathbb{E}$ such that

$$\sup_{H \in C} \left\| \frac{\theta(G + \epsilon H) - \theta(G)}{\epsilon} - \partial\theta_G(H) \right\|_{\mathbb{E}} \underset{\epsilon \to 0}{\longrightarrow} 0$$

for all compact sets $C \subset \mathbb{D}_0$. The derivative $\partial\theta_G$ according to this definition satisfies a chain rule, which can be exploited to determine the Hadamard derivative of complex expressions.

Theorem 2.2 (Functional delta method; cf. e.g. Kosorok, 2008, Theorem 2.8)**:**
Let $\mathbb{D}$ and $\mathbb{E}$ be normed spaces, and let $\theta : \mathbb{D} \to \mathbb{E}$ be Hadamard differentiable at $G \in \mathbb{D}$ tangentially to $\mathbb{D}_0 \subseteq \mathbb{D}$. Suppose that $a_n (G_n - G) \xrightarrow{\mathscr{D}} H$ as $n \to \infty$ for some sequence of constants $a_n \to \infty$, a sequence G_n taking its values in $\mathbb{D}$, and a tight process H with values in $\mathbb{D}_0$. Then,

$$a_n \left(\theta(G_n) - \theta(G) \right) \underset{n \to \infty}{\xrightarrow{\mathscr{D}}} \partial\theta_G(H).$$

Van der Vaart and Wellner (1996, Theorem 3.9.11) and Kosorok (2008, Theorem 2.9) concluded by means of Donsker's theorem that the EBS is valid if we consider plug-in estimators based on Hadamard differentiable statistical functionals, with finite variance of the derivative evaluated at the CDF. Examples include e.g. the sample mean and variance. There are actually several cases where the bootstrap approximation converges at a faster rate than the normal approximation; consider e.g. the mean of continuous variables (Singh, 1981). For a more detailed insight into bootstrap theory, we refer to van der Vaart (1998, Chapter 23).

The EBS does not rely on any parametric assumptions, but the underlying data generally need to be i.i.d. Singh (1981), Friedrich, Brunner, and Pauly (2017), and Hrba et al. (2022) pointed out situations where dependencies lead to poor approximation. Another drawback of this bootstrap method is the computational effort, which entails excessive execution times for large datasets.

2.3.2. Resampling based on the influence function

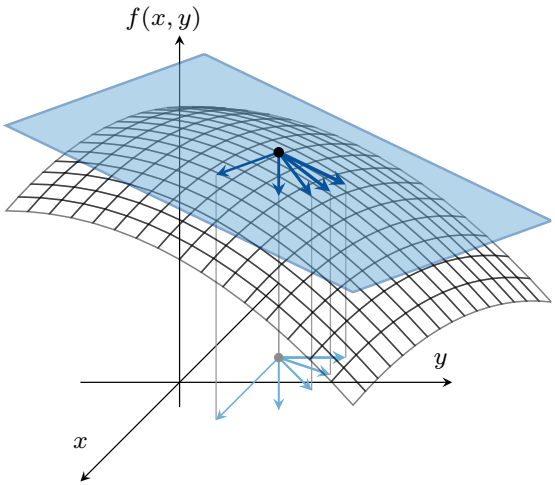
In order to establish a second resampling approach, we define the *IF* of a Hadamard differentiable statistical functional $\theta = \theta(F)$, with F denoting the CDF, by

$$IF_\theta(x) = \partial\theta_F \left(\mathbb{1}_{[x,\infty)} - F \right).$$

Note that $\mathbb{1}_{[x,\infty)}(\cdot)$ is just another representation of the expression $\mathbb{1}\{\cdot \geq x\}$. IFs were first mentioned explicitly by Hampel (1974), and pursuant to the definition as (general-

ized) directional derivative, they quantify how the functional of interest is affected by adding a small point mass to the underlying distribution (see Figure 2.4, Zepeda-Tello et al., 2022). For an i.i.d. sample $(X_1, \ldots, X_n)$ with true CDF F and empirical distribution function $\hat{F}$, the empirical IF $\widehat{IF}_\theta(X_i) = IF_\theta(x)\big|_{F=\hat{F}}$ accordingly measures the 'influence' of observation X_i on the plug-in estimator $\hat{\theta} = \theta(\hat{F})$ ($i \in \{1, \ldots, n\}$).

Figure 2.4: Illustration of the directional derivative: Considering functions on multidimensional spaces, the directional derivative reflects the slope of the tangent plane w.r.t. a specific direction. The Hadamard derivative is a generalization of the directional derivative to infinite-dimensional spaces. It ensures that the derivative towards direction z coincides with the limit of the derivatives towards any sequence of directions that converges to z. The graphic is based on Figure 1(c) in Zepeda-Tello et al. (2022; licensed under CC BY 4.0).



An additional application of the IF is variance approximation. Van der Vaart (1998, Chapter 20) used the von Mises expansion, a 'functional' analogue of the Taylor expansion, to show that

$$\sqrt{n}\Big(\theta(\hat{F}) - \theta(F)\Big) = \sqrt{n}\,\partial\theta_F\Big(\hat{F} - F\Big) + o_P(1),$$

considering once more an i.i.d. sample $(X_1, \ldots, X_n)$ with true and empirical CDFs F and $\hat{F}$, respectively. This representation can be further advanced by the linearity of the Hadamard derivative: One obtains

$$\begin{aligned}\sqrt{n}\Big(\theta(\hat{F}) - \theta(F)\Big) &= \partial\theta_F\bigg(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big(\mathbb{1}_{[X_i,\infty)} - F\Big)\bigg) + o_P(1) \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} IF_\theta(X_i) + o_P(1)\end{aligned} \tag{2.11}$$

as well as $\mathbb{E}\big(IF_\theta\left(X_1\right)\big) = 0$ so that it is tempting to invoke the central limit theorem. The functional delta method (see Theorem 2.2), together with Donsker's theorem, suggests in fact that $\sqrt{n}\Big(\theta(\hat{F}) - \theta(F)\Big)$ converges in distribution to a normal distribution with mean zero and variance equal to $\mathrm{Var}\big(IF_\theta\left(X_1\right)\big)$ if the second moment of $IF_\theta\left(X_1\right)$ is finite. We may hence approximate the variance of $\hat{\theta}$ by

$$\widehat{\mathrm{Var}}\big(\hat{\theta}\big) = \frac{1}{n^2}\sum_{i=1}^{n}\Big(\widehat{IF}_\theta(X_i)\Big)^2.$$

The estimator above allows to construct a CI for θ without resampling. A not so common, yet effective resampling approach to derive time-simultaneous CBs for stochastic processes has been described by Scheike and Zhang (2008). They consider multiple time points over an interval of interest. For each of these time points, they compute the IFs w.r.t. all observations, multiply them with independent standard normal variables and take the sum over the population. This procedure is repeated a large number of times in order to imitate the distribution of $\sqrt{n}\Big(\theta(t;\hat{F}) - \theta(t;F)\Big)$ over the examined interval (cf. Equation (2.11)).

Resampling as described above is less time-intensive than the execution of the classical bootstrap, since it is not necessary to recalculate the functional of interest in multiple data sets. The derivation of the IF can be facilitated by means of Gateaux derivatives and standard differentiation rules; guidance has been given e.g. by Kennedy (2022). Still, the theory underlying von Mises calculus relies on an i.i.d. set-up, and therefore, approximations hinging on dependent data may be inconsistent.

2.3.3. Martingale-based resampling

The third resampling approach we consider in this thesis traces back to a method initially described by Wu (1986). The author suggests to emulate the heteroscedastic residuals in a linear regression model based on their first and second moments, by reweighing the estimated residuals using random multipliers. The same idea can also be applied to the martingale residuals in a counting process model, which makes it particularly appealing w.r.t. TTE data (see Subsection 2.1.1). Here, we focus on the so-called wild bootstrap (WBS) as proposed by Lin, Wei, and Ying (1993).

It is easy to see that under the multiplicative intensity model, the variance of the martingale increments $\mathrm{d}M_i(t)$ given the past $\mathscr{F}_{t-}$ is (nearly) equal to the expectation of the counting process increments $\mathrm{d}N_i$ given $\mathscr{F}_{t-}$ ($i \in \{1,\ldots,n\}$). Following the scheme of Wu (1986), we may thus approximate $\mathrm{d}M_i$ by the product of $\mathrm{d}N_i$ with suitable random multipliers. Choices include independent standard normally distributed random vari-

ables (as considered by Lin, Wei, and Ying, 1993) as well as independent and centred unit Poisson variables (Beyersmann, Di Termini, and Pauly, 2013). Generally, the mean and variance of the multipliers need to converge in probability to 0 and 1, respectively, in order to ensure that the first and second moments of the resampled and the true counting processes are asymptotically equal. A more thorough delineation of the conditions imposed on the multipliers as well as a proof of the validity of this resampling approach when applied to the CIF can be found in Dobler, Beyersmann, and Pauly (2017).

The martingale-based WBS and its variants may be used in a multitude of settings (Lin, 1997; Bluhmki et al., 2019; Ditzhaus and Pauly, 2019). As the underlying theory is built upon the condition of independent censoring (rather than random censoring), the described resampling approach is expected to yield valid outcomes for data that deviate from a strict i.i.d. structure (as long as the censoring process fulfils Condition 2.3). What is more, it can be implemented faster than the EBS for similar reasons as those mentioned in Section 2.3.2.

3. Independent censoring in event-driven trials with staggered entry

The focus of this chapter is on a study design that is specific to clinical studies with time-to-event endpoints. Due to right-censoring, the number of observed events – termed the effective sample size – generally falls short of the actual sample size. It is common to plan clinical trials according to a predetermined target number of events in order to account for this discrepancy. This leads to so-called *event-driven* trials. Studies subject to simple type II censoring are for instance event-driven, and while this censoring scheme is non-random and entails dependent data, it still satisfies the condition of independent censoring (see Subsection 2.1.1).

The participants of clinical trials are usually recruited at different time points throughout an enrolment period, though. If an event-driven set-up is combined with *staggered patient entry*, additional randomness emerges by the projection of the event times onto the study time scale. As a consequence, considerations about the dependence structure of the data are complicated.

The subsequent section establishes the validity of independent censoring in event-driven trials with staggered entry. Afterwards, we conduct simulations to showcase potential issues related to the trial set-up at hand, and demonstrate the impact of these issues by examining real data on immunotherapy for non-small cell lung cancer.

The contents of this chapter have already been published with *Biometrics* (Rühl, Beyersmann, and Friedrich, 2023), and figures as well as tables shown hereafter have been adopted from this source.

3.1. Validity of independent censoring

To illustrate the aforementioned dependence under (simple) type II censoring, consider the setting with sample size $n = 2$ and a follow-up period that lasts until the first event, so that $C_i = T_1 \wedge T_2$, $i \in \{1, 2\}$. The observed data $\left(T_i \wedge C_i,\ \mathbb{1}\{T_i \leq C_i\}\right)_{i \in \{1,\ldots,n\}}$ can thus be represented as

$$\Big(\left(T_1 \wedge T_2,\ \mathbb{1}\{T_1 \leq T_2\}\right),\ \left(T_1 \wedge T_2,\ \mathbb{1}\{T_2 \leq T_1\}\right)\Big).$$

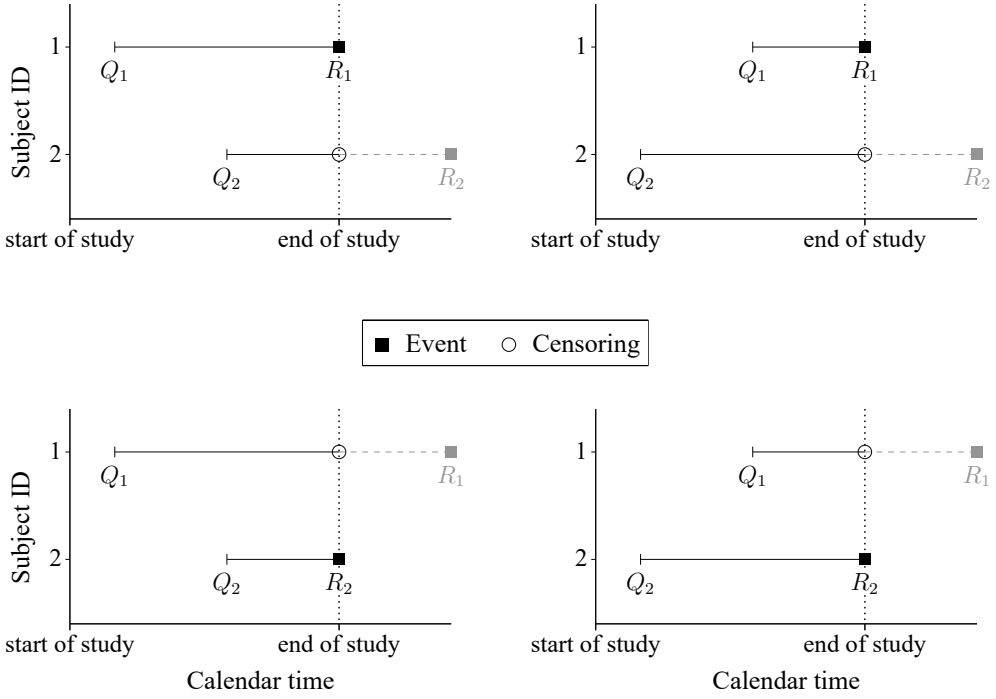
It is clear from the characterization above that the data of the two subjects depend on each other.

If participants enter the study in a successive manner, one needs to look at two different time scales: The censoring times are determined by the chronological sequence of the events, as defined on the calendar time scale, whereas TTE analysis is based on the study time scale with a common origin $t = 0$ for all subjects. Let Q_i and R_i denote the entry and event times of subject i in calendar time scale, respectively. Then, it holds for $n = 2$ that $T_i = R_i - Q_i$ as well as $C_i = R_1 \wedge R_2 - Q_i$, $i \in \{1, 2\}$, and the observed data are therefore of the form

$$\Big(\big(R_1 \wedge R_2 - Q_1,\ \mathbb{1}\{R_1 \leq R_2\} \big),\ \big(R_1 \wedge R_2 - Q_2,\ \mathbb{1}\{R_2 \leq R_1\} \big) \Big).$$

Figure 3.1 depicts the possible scenarios w.r.t. the order of study entries and events, still considering $n = 2$. Regardless of the sample size, the last observed event generally determines the available information on the censored units in the given trial set-up. With less events recorded until the end of follow-up, the number of censored observations increases so that more data points hinge on the time of the last captured event. The dependence within the study data will consequently be stronger.

Figure 3.1: Possible study scenarios in event-driven trials with staggered entry and $n = 2$.



Most analysis techniques for TTE data (implicitly) rely on the assumption of independent censoring to generate unbiased outcomes (see Definition 2.1). Such techniques are also used to examine event-driven trials with staggered entry, which are rather frequent

in clinical practice (see e.g. Elisei et al., 2013; McLaughlin et al., 2015; Sitbon et al., 2015; Husain et al., 2019; Baden et al., 2021). We demonstrate that the condition in Definition 2.1 does in fact hold in event-driven trials with staggered entry in order to ensure that inferences are valid for this particular study design.

For simplicity, our focus is on a standard survival setting without competing risks. Recall now the filtrations $\left(\mathscr{F}_t^c\right)$, $(\mathscr{F}_t)$, and $(\mathscr{G}_t)$ from Subsection 2.1.1. While $\left(\mathscr{F}_t^c\right)$ refers to the history in a (hypothetical) world without censoring, and thus relates to the parameters of interest, $(\mathscr{F}_t)$ denotes the observable, censored history that is actually studied for the analysis. The 'joint' filtration $(\mathscr{G}_t)$ combines both the information from $\left(\mathscr{F}_t^c\right)$ as well as details on the censoring process.

Theorem 3.1 (Independent censoring in event-driven trials with staggered entry)**:**
Consider an event-driven trial with n participants. Suppose that the survival times $T_1, \dots, T_n$ are independent and let Q_i and R_i denote the calendar times of study entry and the event of interest for subject i, respectively, such that $Q_i < R_i \; \forall i \in \{1, \dots, n\}$. Let further $R_{(1)} < R_{(2)} < \cdots < R_{(n)}$ be the ordered event times (w.r.t. calendar time), assuming no ties. We define the observation period to end at time $R_{(m)}$, for a fixed number $m \in \mathbb{N}$ with $0 < m < n$. Then, it holds that $\lambda_i^{\mathscr{G}}(t) = \lambda_i^{\mathscr{F}^c}(t)$ for all $i \in \{1, \dots, n\}$ and for all $t > 0$, i.e. the condition of independent censoring is fulfilled.

The statement above suggests that the usual analysis methods (including Nelson-Aalen estimator and Cox regression) can be applied to examine event-driven trials with staggered entry.

We defer the proof of Theorem 3.1 to Section A.1 in Appendix A. What should be noted is that one must not condition on the calendar times Q_i and R_i, $i \in \{1, \dots, n\}$, when analysing the data, or else, the additional information obtained through the censoring process may allow to predict whether a subject experiences the event in the next instant, so that the intensities w.r.t. $(\mathscr{G}_t)$ and $\left(\mathscr{F}_t^c\right)$ will differ.

3.2. Simulation studies

We conducted two simulation studies in order to explore the impact of the dependence structure inherent to the data in event-driven trials with staggered entry when it comes to practical investigations. The first experiment is based on the final remark in the preceding section: The idea was to condition on the calendar time of study entry – contrary to the concomitant issue of non-independent censoring insinuated by the proof of Theorem 3.1 – and to compare the outcomes to those obtained by an analysis that is oblivious to the calendar times. A second simulation study further aimed to examine the use of analysis methods that are based on the assumption of random instead of independent

censoring. Since the survival and censoring times cannot be expected to be independent in the given study set-up, such analysis methods may lead to bias.

3.2.1. Impact of conditioning on calendar times

To perform the first experiment, we simulated event-driven trials with staggered entry according to the scenarios listed in Table 3.1. The study design was determined by the parameters $(n, m) \in \{(600, 300), (300, 150), (50, 25), (50, 10), (26, 13)\}$, where n denotes the sample size and m is the number of events to be observed. Our focus on rather small sample sizes is due to the fact that the dependence within the data increases with smaller values of n and, in particular, m (see Section 3.1). The survival times followed either an exponential or a Weibull distribution, respectively, with the distribution parameters specified to achieve a target HR out of $\{0.8, 1, 1.25\}$ between randomly allocated, equal-sized groups.

For each data set, a binary group indicator Z_{Ai} was assigned to subject $i \in \{1, \dots, n\}$, so that $\sum_{i=1}^{n} \mathbb{1}\{Z_{Ai} = 0\} = \sum_{i=1}^{n} \mathbb{1}\{Z_{Ai} = 1\} = n/2$. We generated survival times T_i corresponding to the respective group-specific target distribution and added them to uniformly distributed entry times Q_i. This yielded the calendar times R_i of the events. The time from entry to censoring was then determined by the m^{th} largest value of R_i, or more specifically, $C_i = R_{(m)} - Q_i$. One last variable, $Z_{Ni} = \sum_{j=1}^{n} \mathbb{1}\{Q_j < Q_i\}$, was furthermore derived on the basis of the entry times. As a result, we obtained a data set of the form $\big(T_i \wedge C_i, \mathbb{1}\{T_i \leq C_i\}, Q_i, Z_{Ai}, Z_{Ni}\big)_{\,i \in \{1,\dots,n\}}$, which served as the foundation for the subsequent analysis.

Table 3.1 also includes three randomly censored scenarios that were implemented for comparison. By contrasting the analysis outcomes of the event-driven and the randomly censored scenarios, we hoped to be able to assess if any peculiarities are in fact attributable to the interaction between staggered study entry and type II censoring. The data for the randomly censored scenarios were generated in a similar fashion as described above, except that the censoring times were sampled from an exponential distribution and the subject-level data were simulated individually until we acquired m uncensored events. (This meant that the intended sample size n could be attained only approximately, with the accuracy depending on the distribution of the censoring times.)

Following the simulation step, the generated data were analysed by means of three different Cox regression models. First, we fitted a model that quantifies the HR between the groups defined by the indicator Z_A, without controlling for any other covariates. Since we did not condition on calendar times, Theorem 3.1 implies that the outcomes should be unbiased. The corresponding model, hereinafter termed the ‘Standard Model’, was thus used as a reference.

Table 3.1: Simulation scenarios considered w.r.t. the impact of conditioning on calendar times in event-driven trials with staggered entry.

Censoring	HR	Distribution of T ($Z_A = 0$ / $Z_A = 1$)	n	m	Distribution of Q
event-driven	1.00	$Exp(1)$ / $Exp(1)$	600	300	$\mathcal{U}\big(0, q_{Exp(1)}(0.5)\big)$[a]
			300	150	$\mathcal{U}\big(0, q_{Exp(1)}(0.5)\big)$[a]
			50	25	$\mathcal{U}\big(0, q_{Exp(1)}(0.5)\big)$[a]
			50	10	$\mathcal{U}\big(0, q_{Exp(1)}(0.2)\big)$[a]
			26	13	$\mathcal{U}\big(0, q_{Exp(1)}(0.5)\big)$[a]
		$\mathcal{W}b(0.5, 1)$ / $\mathcal{W}b(0.5, 1)$	600	300	$\mathcal{U}\big(0, q_{\mathcal{W}b(0.5,1)}(0.5)\big)$[a]
			300	150	$\mathcal{U}\big(0, q_{\mathcal{W}b(0.5,1)}(0.5)\big)$[a]
			50	25	$\mathcal{U}\big(0, q_{\mathcal{W}b(0.5,1)}(0.5)\big)$[a]
			50	10	$\mathcal{U}\big(0, q_{\mathcal{W}b(0.5,1)}(0.2)\big)$[a]
			26	13	$\mathcal{U}\big(0, q_{\mathcal{W}b(0.5,1)}(0.5)\big)$[a]
	0.80	$Exp(1)$ / $Exp(0.8)$	600	300	$\mathcal{U}\big(0, q_{Exp(1)}(0.5)\big)$[a]
			300	150	$\mathcal{U}\big(0, q_{Exp(1)}(0.5)\big)$[a]
			50	25	$\mathcal{U}\big(0, q_{Exp(1)}(0.5)\big)$[a]
			50	10	$\mathcal{U}\big(0, q_{Exp(1)}(0.2)\big)$[a]
			26	13	$\mathcal{U}\big(0, q_{Exp(1)}(0.5)\big)$[a]
		$\mathcal{W}b(0.5, 1)$ / $\mathcal{W}b\big(0.5, \frac{1}{0.8^2}\big)$	600	300	$\mathcal{U}\big(0, q_{\mathcal{W}b(0.5,1)}(0.5)\big)$[a]
			300	150	$\mathcal{U}\big(0, q_{\mathcal{W}b(0.5,1)}(0.5)\big)$[a]
			50	25	$\mathcal{U}\big(0, q_{\mathcal{W}b(0.5,1)}(0.5)\big)$[a]
			50	10	$\mathcal{U}\big(0, q_{\mathcal{W}b(0.5,1)}(0.2)\big)$[a]
			26	13	$\mathcal{U}\big(0, q_{\mathcal{W}b(0.5,1)}(0.5)\big)$[a]
	1.25	$Exp(1)$ / $Exp(1.25)$	600	300	$\mathcal{U}\big(0, q_{Exp(1)}(0.5)\big)$[a]
			300	150	$\mathcal{U}\big(0, q_{Exp(1)}(0.5)\big)$[a]
			50	25	$\mathcal{U}\big(0, q_{Exp(1)}(0.5)\big)$[a]
			50	10	$\mathcal{U}\big(0, q_{Exp(1)}(0.2)\big)$[a]
			26	13	$\mathcal{U}\big(0, q_{Exp(1)}(0.5)\big)$[a]
		$\mathcal{W}b(0.5, 1)$ / $\mathcal{W}b\big(0.5, \frac{1}{1.25^2}\big)$	600	300	$\mathcal{U}\big(0, q_{\mathcal{W}b(0.5,1)}(0.5)\big)$[a]
			300	150	$\mathcal{U}\big(0, q_{\mathcal{W}b(0.5,1)}(0.5)\big)$[a]
			50	25	$\mathcal{U}\big(0, q_{\mathcal{W}b(0.5,1)}(0.5)\big)$[a]
			50	10	$\mathcal{U}\big(0, q_{\mathcal{W}b(0.5,1)}(0.2)\big)$[a]
			26	13	$\mathcal{U}\big(0, q_{\mathcal{W}b(0.5,1)}(0.5)\big)$[a]
random ($C \sim Exp(1)$)[b]	1.00	$Exp(1)$ / $Exp(1)$	≈ 50	25	$\mathcal{U}\big(0, q_{Exp(1)}(0.5)\big)$[a]
random ($C \sim Exp(4)$)[b]	1.00	$Exp(1)$ / $Exp(1)$	≈ 50	10	$\mathcal{U}\big(0, q_{Exp(1)}(0.2)\big)$[a]
random ($C \sim Exp(1)$)[b]	1.00	$Exp(1)$ / $Exp(1)$	≈ 26	13	$\mathcal{U}\big(0, q_{Exp(1)}(0.5)\big)$[a]

[a] The m/n quantile of the survival time distribution was set as upper limit of the entry times in order to keep the probability of entries after the m^{th} event low. ($q_{\mathcal{D}}(x)$ denotes the x quantile of the distribution $\mathcal{D}$.)
[b] The distribution parameters of the censoring times were chosen so that $P(T \leq C) = \frac{m}{n}$.

Two less conventional Cox models were further considered to illustrate the effect of conditioning on calendar times: In addition to the group indicator, 'Model 1' also included the entry time Q as a second covariate. 'Model 2' moreover covered a third predictor on top of Z_A and Q, namely the variable Z_N, which represents the number of study participants that have already been recruited at a subject's admission time. The rationale behind this covariate choice was to mimic the calendar time information that is conveyed through counting processes more directly. We hoped that the greater level of detail on the order of the subject entries would distinctly distort the intensities in Model 2.

It is important to note that we do not propose the use of calendar time models when analysing event-driven trials with staggered entry, but rather investigate Models 1 and 2 to examine the potential bias that may arise due to the violation of independent censoring.

The process of generating data and performing the Cox regression analyses was repeated 100,000 times for each of the scenarios specified in Table 3.1. Some of the iterations in the settings with small values of m involved data with all (or all but one) of the observed events occurring among subjects from the same group. The resulting HRs were estimated extremely high or low, which is why we disregarded such data and assessed the accuracy of the three Cox models based on the summarized outcomes of the remaining iterations. As our outcomes showed, this only reduced the total magnitude of the deviations, but did not alter the interrelations between the biases observed in the distinct models.

Figure 3.2: Shadow plots of the Breslow estimators in the Weibull scenario with HR 1, $n = 50$, and $m = 10$. The shadow lines are restricted to a random sample of size 2,000 for greater clarity.

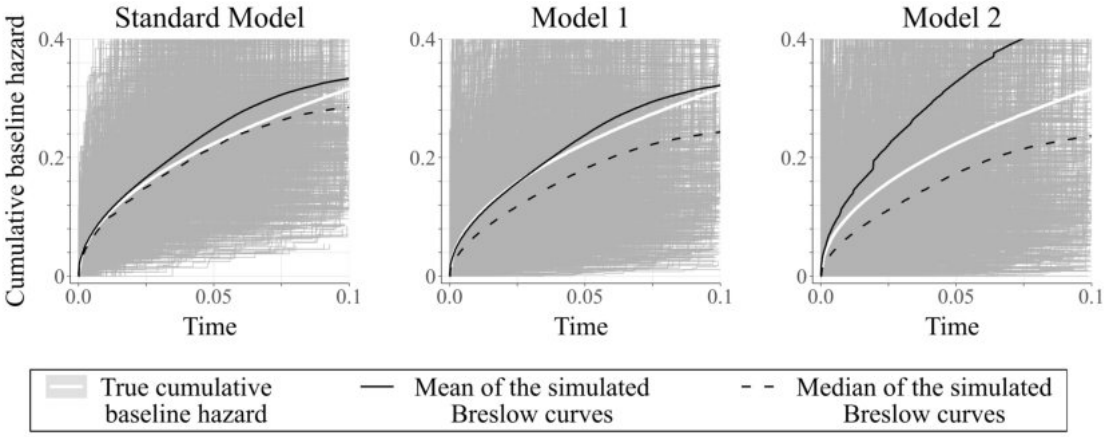


Figure 3.2 depicts the Breslow estimators of the cumulative baseline hazards in each model, considering the Weibull scenario with an underlying HR of 1 and parameters $n = 50$ as well as $m = 10$. Note that 1,034 iterations were excluded because the ob-

served events in the corresponding data were not distributed sufficiently well across the groups. One can see that the median of the Breslow estimators in the Standard Model nearly coincides with the true cumulative baseline hazard. Differences are visible only at late time points, when few events were observed, so that the amount of available data was rather low. It is, in contrast, clear that the medians of the estimated cumulative baseline hazards in Models 1 and 2 are too small at all times, while the mean curve in Model 2 suffers from severe upward bias.

The extent of these deviations is moreover quantified in Table 3.2. Looking at several individual time points over the study time interval, the mean bias in Model 1 is fairly similar to that found in the Standard Model. In Model 2, however, we encountered discrepancies that were multiple times as high, even with the Monte Carlo standard errors (MCSEs) taken into account. The median bias in Models 1 and 2 further amounts to more than 5 times its value in the Standard Model, which means that the calendar time models must have yielded distorted estimators for a large number of iterations. We also computed the root mean square errors (RMSEs) in order to evaluate the precision of the estimated cumulative baseline hazards. The results reflect the deficiency of Model 2 even more distinctly.

Table 3.2: Bias of the Breslow estimators in the Weibull scenario with HR 1, $n = 50$, and $m = 10$, considering selected time points.

Time	Measure of bias	Standard Model		Model 1		Model 2	
		Value	MCSE	Value	MCSE	Value	MCSE
0.03	mean bias	0.01094	0.00034	0.00138	0.00064	0.07112	0.02678
	median bias	-0.00859	0.00004[a]	-0.04163	0.00027[a]	-0.04837	0.00020[a]
	RMSE	0.10663	0.00193[a]	0.20163	0.04030[a]	8.42391	5.87212[a]
0.05	mean bias	0.02570	0.00054	0.01440	0.00083	0.10420	0.02868
	median bias	-0.00620	0.00073[a]	-0.04226	0.00009[a]	-0.05025	0.00071[a]
	RMSE	0.17141	0.00572[a]	0.26164	0.03001[a]	9.02352	4.96754[a]
0.07	mean bias	0.03267	0.00075	0.02099	0.00101	0.12607	0.03180
	median bias	-0.00917	0.00003[a]	-0.04829	0.00028[a]	-0.05604	0.00075[a]
	RMSE	0.23794	0.00758[a]	0.31989	0.02447[a]	10.00564	4.39934[a]

[a] The MCSEs are determined by means of the jackknife estimator (Efron and Stein, 1981).

Apart from the Breslow estimators, the estimated regression coefficients in each model were considered, too (see Table 3.3). We studied their performance in terms of the log-transformed rather than the actual HRs, since the distribution of the corresponding estimators is symmetric (see Subsection 2.1.2), and thus, their accuracy is easier to assess. As Table 3.3 shows, the mean bias of the estimated log-HR for covariate Z_A differs hardly between the three models. The median bias is likewise very small, but one can

still observe a slight increase proceeding from the Standard Model over Model 1 to Model 2. Such an increase is also present w.r.t. the RMSEs. Besides, the coverage probabilities of the CIs for the log-HR decrease by a small amount when passing from model to model.

A different picture emerged concerning the additional covariates included in the calendar time models. While the bias of the log-HR for predictor Z_N in Model 2 can be neglected, both Model 1 and 2 drastically overestimate the influence of the entry times Q. The mean and median biases as well as the RMSEs are very high, which implies that conditioning on the calendar times does in fact disturb the intensities. Nevertheless, the CIs for the log-HR of Q seem to be accurate.

Table 3.3: Bias of the estimated log-HRs in the Weibull scenario with HR 1, $n = 50$, and $m = 10$.

Covariate	Measure of bias	Standard Model		Model 1		Model 2	
		Value	MCSE	Value	MCSE	Value	MCSE
Z_A	mean bias	0.00235	0.00209	0.00263	0.00213	0.00227	0.00220
	median bias	0.00258	0.00000[a]	0.00290	0.00144[a]	0.00337	0.00000[a]
	RMSE	0.65654	0.00146[a]	0.66925	0.00149[a]	0.69106	0.00156[a]
	coverage	0.98294	0.00041	0.97954	0.00045	0.97413	0.00050
Q	mean bias			9.19421	0.09133	1.22950	0.80990
	median bias			8.25649	0.03593[a]	2.99396	0.16478[a]
	RMSE			30.16661	0.09055[a]	254.78666	0.83921[a]
	coverage			0.95076	0.00069	0.95095	0.00069
Z_N	mean bias					0.00809	0.00079
	median bias					0.00571	0.00016[a]
	RMSE					0.24807	0.00080[a]
	coverage					0.95100	0.00069

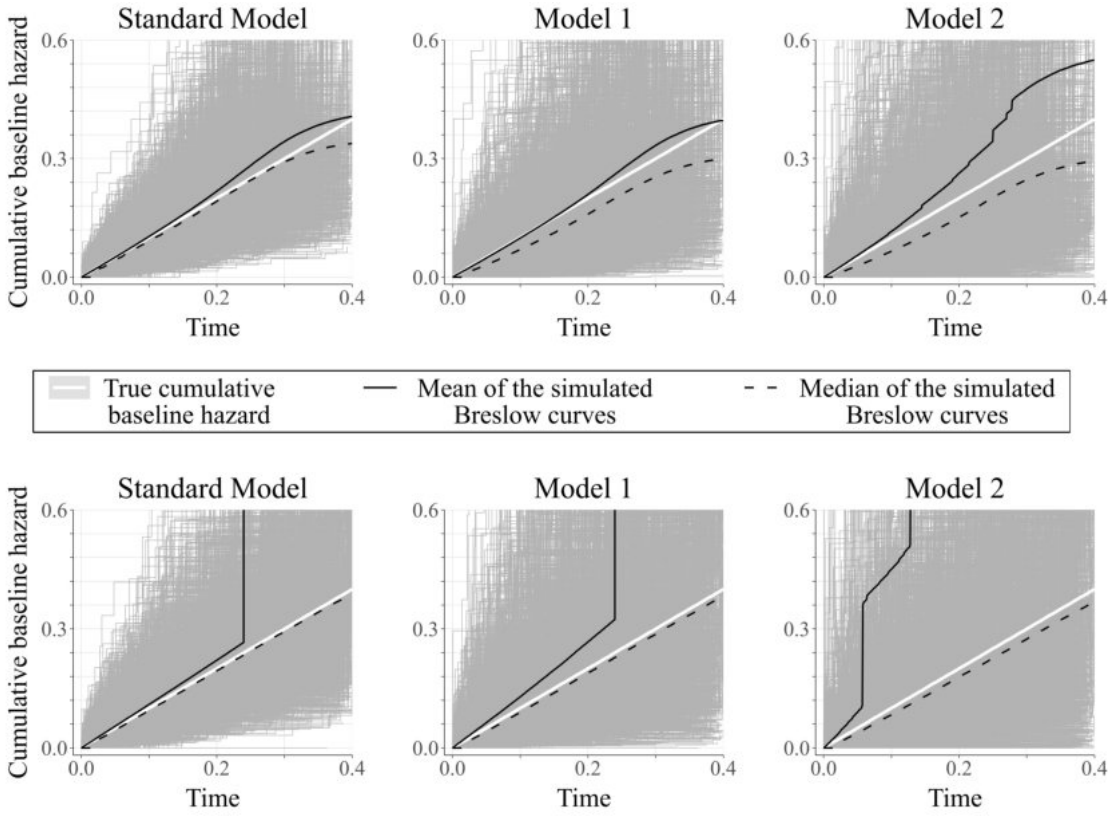
[a] The MCSEs are determined by means of the jackknife estimator (Efron and Stein, 1981).

Similar findings to those described above were also obtained with other sample sizes, event time distributions and HRs (see Section B.1 in Appendix B for the outcomes in the remaining scenarios). The bias of the estimators derived in the calendar time models was less pronounced for larger values of the parameters n and m and for greater proportions of m/n, though. This is not surprising bearing in mind that the data involved stronger dependencies in those cases.

As mentioned earlier, we further re-simulated some of the scenarios, but implemented random instead of event-driven censoring. The resulting Breslow estimators in the exponential scenario with HR 1 and parameters $n = 50$ as well as $m = 10$ are shown in the second row of Figure 3.3. The first row moreover depicts the corresponding outcomes

under event-driven censoring. Note that we excluded 2,133 and 1,032 iterations, respectively, because (nearly) all of the observed events occurred in one group. (Individual iterations with only two observed events in one of the treatment groups still led to very extreme results in the randomly censored scenarios; see the extreme slopes of the curves that illustrate the means of the Breslow estimators.) Even though the mean curves do not resemble the true cumulative baseline hazard function in either the event-driven or the randomly censored case, it is indicated by the medians that the estimators in the calendar time models are no longer biased if censoring is random. This confirms the combination of staggered study entry and type II censoring as the cause of the distortions.

Figure 3.3: Shadow plots of the Breslow estimators in the exponential (first row) and the randomly censored (second row) scenarios with HR 1, $n = 50$, and $m = 10$. The shadow lines are restricted to a random sample of size 2,000 for greater clarity.



All in all, our simulations demonstrated that conditioning on calendar times in event-driven trials with staggered entry does disturb the underlying intensities. The extra information on the sequence of the events that is provided if one also conditions on the number of recruited subjects deranges the intensities even more. With small sample sizes and few observed events until the end of follow-up, analyses are therefore prone

to bias, which may affect Cox regression estimates as well as Breslow estimates, and consequently, predictions of the survival probabilities.

3.2.2. Impact of using methods based on random censoring

Applying the considerations in Subsection 2.3.1 to TTE data, it turns out that the EBS should only be used under random censoring, or else, there will be non-i.i.d. data. The WBS, on the other hand, is valid if censoring is independent (see Subsection 2.3.3). Our second simulation study compared the outcomes obtained after employing these two resampling techniques in an event-driven setting with staggered entry, where censoring is independent but not random.

Preliminary simulations suggested that the interrelations between the data in a classical two-state survival model are too simple to reveal any noteworthy effects. This is why we performed our investigations based on the illness-death model without recovery that has been introduced in Subsection 2.1.3 (see Figure 2.1). The idea was that internal left-truncation caused by the transition into the intermediate state might reinforce potential effects because of the additional pressure exerted on statistical procedures.

To generate data conforming to the illness-death model, we followed the simulation process described by Nießl et al. (2021), but implemented staggered study entry and event-driven censoring. First, n uniformly distributed entry times were simulated over the interval between 0 and 60. The time until the transition to the subsequent state was then determined by exponentially distributed random values, considering a rate parameter of 0.04, which reflects the sum of the transition hazards from the initial state to illness, that is, $\alpha_{01}(t) = 0.01$, and to death, i.e. $\alpha_{02}(t) = 0.03$, for $t \in [0, \tau]$ (see Figure 2.1). The type of the event was randomly selected between illness and death, with probabilities $\alpha_{01}(t) \,/\, \big(\alpha_{01}(t) + \alpha_{02}(t)\big) = 0.25$ and $\alpha_{02}(t) \,/\, \big(\alpha_{01}(t) + \alpha_{02}(t)\big) = 0.75$, respectively (Beyersmann, Latouche, et al., 2009). Subjects who had transitioned to the illness state died after a waiting time that followed an exponential distribution with parameter 0.1. We finally specified the end of the follow-up period as the time at which the m^{th} death occurred, and transformed the data to the study-time scale afterwards.

As can be seen in Table 3.4, varying values of the parameters n and m were examined. Besides, the simulations also covered simultaneous study entry for purposes of comparison.

The next step was to apply both the EBS and the WBS to the generated data. We did so by deriving 95% CIs for the cumulative hazard $A_{12}(t)$, our focus being on the transition from illness to death because of the internal left-truncation mentioned earlier. See Bluhmki et al. (2019) for the theoretical justification of the WBS in this setting.

To improve small-sample performance, we studied log-transformed CIs according to the formula given by Andersen, Borgan, et al. (1993, Subsection IV.1.3.1.):

$$\left[\hat{A}_{12}(t)\exp\left(\frac{-q(0.975)\sqrt{\hat{\sigma}^2(t)}}{\hat{A}_{12}(t)}\right),\ \hat{A}_{12}(t)\exp\left(\frac{q(0.975)\sqrt{\hat{\sigma}^2(t)}}{\hat{A}_{12}(t)}\right)\right].$$

The quantile $q(0.975)$ and the variance estimator $\hat{\sigma}^2(t)$ were based on the respective bootstrap statistics, however. In case of the EBS, the empirical variance of $\left(\hat{A}_{12}^{(b)}(t)\right)_{b\in\{1,\ldots,B\}}$ was thus used for $\hat{\sigma}^2(t)$, with $\hat{A}_{12}^{(b)}(t)$ referring to the bootstrap estimator obtained in the b^{th} bootstrap sample. The term $q(0.975)$ further represented the empirical 0.975 quantile of

$$\left(\left(\hat{A}_{12}^{(b)}(t)-\hat{A}_{12}(t)\right)\Big/\sqrt{\hat{\sigma}^2(t)}\right)_{b\in\{1,\ldots,B\}}.$$

When the CIs were constructed by the WBS instead, the expression $\hat{\sigma}^2(t)$ described the empirical variance of the bootstrap processes

$$U^{(b)}(t)=\sum_{i=1}^{n}\int_0^t \frac{\mathbb{1}\{Y_1(u)>0\}}{Y_1(u)}\,\mathrm{d}N_{12i}(u)\,G_i^{(b)}$$

for $b\in\{1,\ldots,B\}$, while $q(0.975)$ characterized the empirical 0.975 quantile of the set $\left(U^{(b)}(t)/\sqrt{\hat{\sigma}^2(t)}\right)_{b\in\{1,\ldots,B\}}$. Note that the subscript of $Y_1(u)$ in the definition above indicates that we address the risk set for the transition from illness to death (i.e. subjects who have fallen ill before time u and are still alive just prior to u), and $\mathrm{d}N_{12i}(u)$ is the increment of the counting process that jumps if subject $i\in\{1,\ldots,n\}$ dies after illness. The term $G_i^{(b)}$ moreover labels the i^{th} component of a (standard normally distributed) vector of multipliers.

We generated $B=1,000$ bootstrap samples for both resampling approaches, respectively, and evaluated the accuracy of the CIs at time points $t\in\{16,18,20\}$. For that purpose, we took the coverage and the widths of the CIs into account.

Each scenario was simulated 1,000 times in order to keep the MCSE w.r.t. the coverage probabilities of the CIs below 1.6%. The mean outcomes are summarized in Table 3.4.

It is evident that for large sample sizes, the CIs derived by the WBS were somewhat narrower than their counterparts obtained by means of the EBS. At the same time, both resampling methods achieved coverage probabilities close to the nominal level of 95%. When the parameters n and m were below 200 and 100, respectively, the WBS yielded wider CIs that clearly outperformed the EBS in terms of coverage, though. This effect can once again be explained by the higher degree of dependence associated with smaller sample sizes.

Table 3.4: Coverage probabilities (in %) and mean widths of the bootstrapped CIs at selected time points.

n	m	Time	Staggered entry				Simultaneous entry			
			EBS		WBS		EBS		WBS	
			Coverage	Width	Coverage	Width	Coverage	Width	Coverage	Width
600	300	16	93.9	1.430	93.9	1.372	95.0	1.256	94.4	1.205
		18	93.9	1.492	93.3	1.432	95.0	1.294	94.7	1.244
		20	93.5	1.559	93.6	1.492	95.2	1.326	94.5	1.273
400	200	16	93.9	1.732	94.1	1.663	93.7	1.521	94.3	1.490
		18	93.6	1.811	94.2	1.729	92.8	1.567	94.3	1.535
		20	94.6	1.897	94.7	1.804	94.4	1.612	94.6	1.570
200	100	16	91.1	2.230	93.3	2.354	91.8	2.030	93.1	2.056
		18	90.9	2.339	94.8	2.463	92.1	2.107	93.8	2.122
		20	90.5	2.444	94.0	2.582	93.4	2.162	93.1	2.171
100	50	16	83.0	2.217	95.3	3.298	88.2	2.294	94.7	2.901
		18	84.3	2.298	95.6	3.455	87.9	2.370	94.2	2.971
		20	82.0	2.349	95.9	3.640	88.1	2.422	94.2	3.030
80	40	16	77.6	2.031	94.8	3.631	83.2	2.241	93.8	3.181
		18	77.1	2.079	95.4	3.811	83.9	2.302	94.1	3.258
		20	76.2	2.076	96.1	4.000	85.1	2.354	93.7	3.319
50	25	16	66.1	1.508	89.9	3.952	73.0	1.832	93.3	3.812
		18	61.2	1.492	91.8	4.216	72.3	1.871	94.6	3.922
		20	61.3	1.470	94.5	4.493	73.9	1.901	95.0	4.005

Similar outcomes were observed when we simulated simultaneous study entry. The difference between the two resampling methods was less pronounced in this case, though.

It follows that martingale-based analysis methods, which rely on the assumption of independent rather than random censoring, are indeed preferable in event-driven trials with staggered entry.

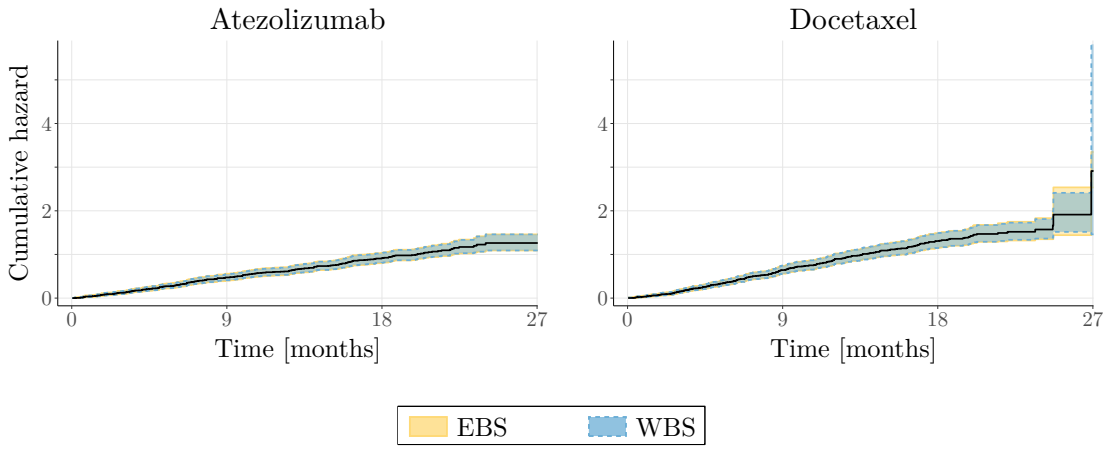
3.3. Analysis of the OAK trial

In addition to the simulation studies, we further examined the implications of Theorem 3.1 considering real study data. The randomized, open-label OAK trial was conducted among 1,225 patients suffering from advanced-stage or metastatic, previously treated non-small cell lung cancer with the aim to compare the efficacy and safety of the immunotherapy agent atezolizumab to that of docetaxel, the standard of care at that time. Participants were enrolled in 194 oncology centres over a period from March to

November 2014 (Rittmeyer et al., 2017). We concentrated on the primary efficacy population that comprised the first 850 enrolled subjects, 425 of whom were assigned to receive docetaxel and atezolizumab, respectively. The primary endpoint was overall survival. According to the statistical analysis plan, the OAK trial was designed event-driven, with the follow-up period scheduled to end when about 595 deaths had occurred (CDER, 2016). Gandara et al. (2018) have made the data for the primary efficacy population publicly available as part of their supplementary material.

We proceeded as described in the previous subsection to derive CIs for the cumulative hazard function of death by means of the EBS and the WBS, respectively. Figure 3.4 depicts the pointwise Nelson-Aalen estimates together with the corresponding 95% CIs. Our (naive) analysis showed a slight advantage of atezolizumab over docetaxel (namely, an estimated cumulative hazard of 1.26 vs. 2.91 after 27 months). It is further obvious that the CIs obtained with the two resampling approaches hardly differ, which we attribute to the large sample size of 850 patients.

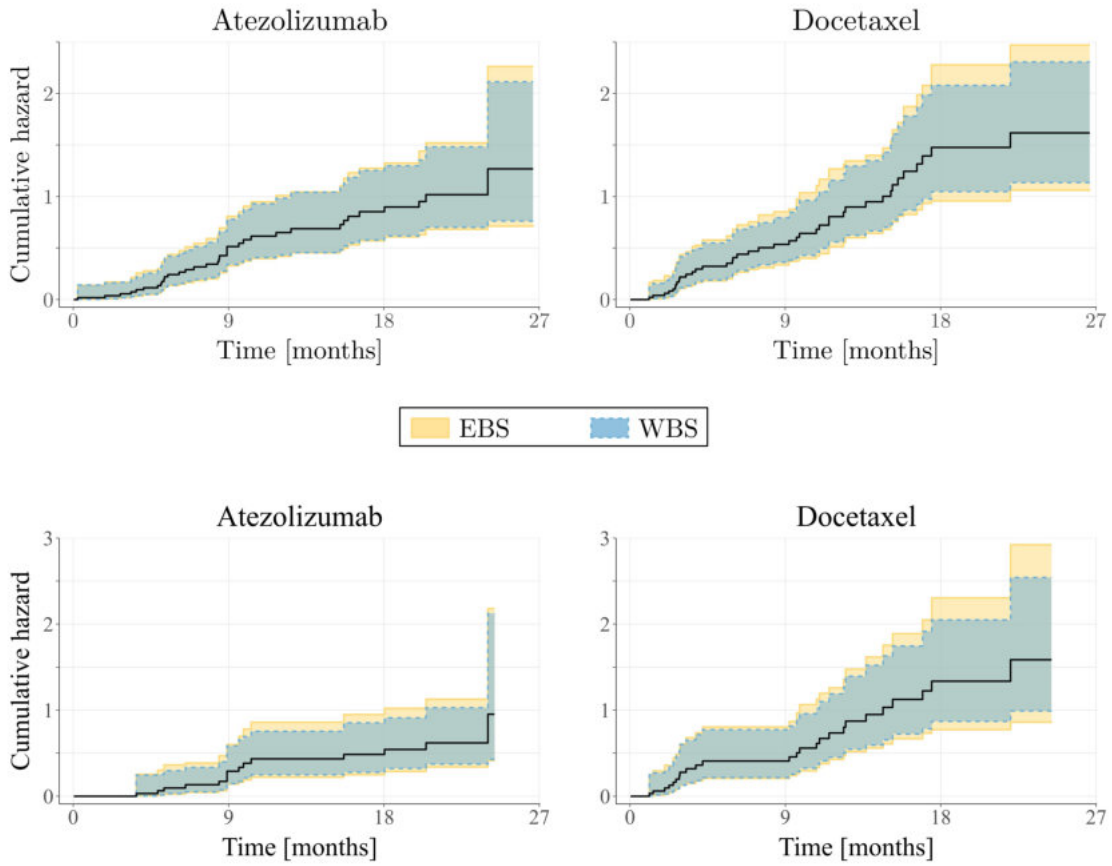
Figure 3.4: 95% CIs for the cumulative hazard of death derived using the EBS and the WBS, respectively.



In order to demonstrate the impact of applying inappropriate methods to smaller samples, we repeated the analysis for subsets of the original data that were generated by randomly drawing observations (without replacement) until a pre-specified number m of observed events was achieved. See Figure 3.5 for an illustration of the CIs in subsets with $m = 75$ and $m = 40$ observed events, respectively. As the figure shows, the differences between the intervals are more distinct here, with the CIs resulting from the classical bootstrap being wider, in particular at later time points.

Such small-sample subsets are relevant in practice in the context of interim analyses, where data are evaluated prior to the closure of the study. Similar to our artificial example, only a fraction of the ultimate number of events has been observed by then.

Figure 3.5: 95% CIs for the cumulative hazard of death derived using the EBS and the WBS in random subsets including 75 (first row) and 40 (second row) observed events, respectively.



We conclude that it is important to carefully deliberate on appropriate analysis techniques when evaluating event-driven trials with staggered entry, in particular with data that only involve few observed events.

4. Resampling-based inference for the average treatment effect in competing-risks data

In the subsequent chapter, we shift our focus to a competing risks setting as described in Subsection 2.1.3. Our goal is to compare the treatment groups in terms of the time to the event of interest, while controlling for confounding bias.

It has been mentioned that the HR – probably the most common effect estimator for TTE data – is subject to non-collapsibility, selection bias, and time invariability (see Subsection 2.2.2). These issues are particularly challenging when causal conclusions are to be drawn. We therefore adhere to the definition of the ATE given in Equation (2.10): The treatment effect is measured by the risk difference between the two exposure groups, considering the event of interest. An estimate can for instance be obtained by means of direct standardization (see Butt et al., 2021) or PS matching (see Chauhan et al., 2022).

In clinical trials, researchers typically not only wish to estimate the extent of the treatment effect, but also assess the (un)certainty of the obtained effect estimate. This can be accomplished by constructing pointwise CIs and time-simultaneous CBs. Due to the complex distribution of stochastic processes associated with causal estimators of the ATE, the derivation of exact confidence regions is quite involved, though. The usual approach to tackle this problem is to approximate the asymptotic distribution of such processes by means of resampling, and in practice, statisticians basically always resort to the classical bootstrap (cf. Neumann and Billionnet, 2016; Lesko and Lau, 2017; Ryalen et al., 2020). It has been indicated in Subsection 2.3.1 and, in particular, in Chapter 3 that the EBS is flawed, however, if the underlying data are not i.i.d.

The chapter at hand thus investigates the performance of alternative resampling approaches – namely a technique building on the IF (see Subsection 2.3.2) as well as the martingale-based WBS (see Subsection 2.3.3) – when applied to approximate the limiting distributions of the stochastic processes that relate to the g-formula estimator and the PS-matched estimator of Equation (2.10), respectively.

We first consider the ATE estimator that results from direct standardization. A martingale representation characterizing the asymptotic distribution of the corresponding process is derived, which we use to prove the validity of the three resampling methods in the given context. For a comparison of the techniques in practical applications, we conduct a simulation study and analyse real study data on the long-term disease progression among patients with early-stage Hodgkin's disease.

Afterwards, the same simulation set-up and study data are further adopted to examine the performance of (suitable variants of) the resampling approaches given PS-matched data.

The proofs in Subsection 4.1.1 have been released in the *Scandinavian Journal of Statistics* before (Rühl and Friedrich, 2024a). In Subsections 4.1.2 and 4.1.3, we further report results – including figures and tables – of the simulation study as well as the real data analysis published with *Statistics and Computing* (Rühl and Friedrich, 2024b).

4.1. Inference using the g-formula

The basis for estimating the ATE in competing risks settings is an i.i.d. data sample of the form $\big((T_i \wedge C_i, D_i, \mathbf{Z}_i)\big)_{i\in\{1,\dots,n\}}$, considering the event indicator $D_i \in \{0, 1, \dots, K\}$ (with $D_i = 1$ w.l.o.g. identifying the cause of interest) and the vector $\mathbf{Z}_i = \big(Z_{Ai}, \mathbf{Z}_{Li}^T\big)^T$, which combines subject i's treatment indicator $Z_{Ai} \in \{0, 1\}$ with their bounded covariate vector $\mathbf{Z}_{Li} \in \mathbb{R}^p$, $i \in \{1, \dots, n\}$ (see Subsection 2.1.3 and Section 2.2). Hereafter, assume the absence of ties and the conditional independence $T_i \perp\!\!\!\perp C_i \mid \mathbf{Z}_i$. Another prerequisite is that the covariates in $\mathbf{Z}_L$ are sufficient to fulfil the identifiability conditions stated in Subsection 2.2.1.

By means of the g-formula, one obtains the following estimator of the ATE:

$$\widehat{ATE}_{ds}(t) = \frac{1}{n}\sum_{i=1}^{n}\Big(\hat{F}_1(t \mid Z_A = 1, \mathbf{Z}_L = \mathbf{Z}_{Li}) - \hat{F}_1(t \mid Z_A = 0, \mathbf{Z}_L = \mathbf{Z}_{Li})\Big).$$

We may derive $\hat{F}_1(t \mid a, \mathbf{l})$ by fitting cause-specific Cox models with covariates Z_A and $\mathbf{Z}_L$ for each event type, which yields the estimated values of

$$\hat{A}_k\big(t \mid Z_A = a, \mathbf{Z}_L = \mathbf{l}\big) = \hat{A}_{0k}(t)\exp\Big(\hat{\boldsymbol{\beta}}_k^T\big(a, \mathbf{l}^T\big)^T\Big),$$

with

$$\hat{A}_{0k}(t) = \int_0^t \frac{\mathrm{d}N_k(u)}{\sum_{i=1}^n Y_i(u)\exp\big(\hat{\boldsymbol{\beta}}_k^T\mathbf{Z}_i\big)},$$

for $k \in \{1, \dots, K\}$, and plugging $\hat{A}_k(t \mid a, \mathbf{l})$ into the subsequent formula proposed by Benichou and Gail (1990) (see Ozenne, Scheike, et al., 2020):

$$\hat{F}_1\big(t \mid Z_A = a, \mathbf{Z}_L = \mathbf{l}\big) = \int_0^t \exp\Big(-\sum_{k=1}^{K}\hat{A}_k(u \mid a, \mathbf{l})\Big)\,\mathrm{d}\hat{A}_1(u \mid a, \mathbf{l}).$$

Here, $\hat{\boldsymbol{\beta}}_k = \big(\hat{\beta}_{kA}, \hat{\boldsymbol{\beta}}_{kL}^T\big)^T \in \mathbb{R}^{p+1}$ denotes the vector of the regression coefficients estimated in the Cox model for cause k. Note that it is reasonable to take the (log-)HRs into account despite their shortcomings in the context of causal inference, since the in-

terpretation of the ATE does not directly depend on $\hat{\boldsymbol{\beta}}_k$. Besides, it should be mentioned that the described approach of fitting k cause-specific Cox models implicitly imposes the proportional hazards assumption on each event type.

4.1.1. Asymptotic distribution of the process for the average treatment effect & resampling-based approximations

We study the asymptotic distribution of the stochastic process $\big(U_n(t)\big)_{t\in[0,\tau]}$ determined by

$$U_n(t) = \sqrt{n}\Big(\widehat{ATE}_{ds}(t) - ATE(t)\Big)$$

as $n \to \infty$. Before doing so, recall the definitions of the expressions $\mathbf{S}^{(r)}(\boldsymbol{\beta}, t)$, $\mathbf{s}^{(r)}(\boldsymbol{\beta}, t)$ (for $r \in \{0, 1, 2\}$), $\mathbf{E}(\boldsymbol{\beta}, t)$, $\mathbf{e}(\boldsymbol{\beta}, t)$, and $\boldsymbol{\Sigma}_k$ in Subsection 2.1.2, which have been adapted to the competing risks setting here by using the quantities obtained in the k^{th} cause-specific Cox model, respectively ($k \in \{1, \ldots, K\}$). In addition, take account of

$$\mathbf{h}_k\big(t \mid Z_A = a, \mathbf{Z}_L = \mathbf{l}\big) = \int_0^t \Big(\big(a, \mathbf{l}^T\big)^T - \mathbf{e}(\boldsymbol{\beta}_{0k}, u)\Big)\, \mathrm{d}A_k(u \mid a, \mathbf{l}),$$

$$\boldsymbol{\varphi}_1\big(t \mid Z_A = a, \mathbf{Z}_L = \mathbf{l}\big) = \int_0^t S(u- \mid a, \mathbf{l})\, \mathrm{d}\mathbf{h}_1(u \mid a, \mathbf{l}),$$

$$\boldsymbol{\psi}_{1k}\big(t \mid Z_A = a, \mathbf{Z}_L = \mathbf{l}\big) = \int_0^t \big(F_1(t \mid a, \mathbf{l}) - F_1(u \mid a, \mathbf{l})\big)\, \mathrm{d}\mathbf{h}_k(u \mid a, \mathbf{l}).$$

These functions stem from the appendix of Cheng, Fine, and Wei (1998), with the vector $\boldsymbol{\beta}_{0k} = \big(\beta_{0kA}, \boldsymbol{\beta}_{0kL}^T\big)^T \in \mathbb{R}^{p+1}$ denoting the true regression coefficients in the proportional hazards model for cause k, and $S(t \mid a, \mathbf{l}) = \exp\big(-\sum_{k=1}^K A_k(t \mid a, \mathbf{l})\big)$.

For $k \in \{1, \ldots, K\}$, $i \in \{1, \ldots, n\}$, we introduce

$$H_{k1i}(u, t) = \frac{\widetilde{H}_{k1}(u, t)}{\sqrt{n}\, S^{(0)}(\boldsymbol{\beta}_{0k}, u)}$$

and

$$H_{k2i}(u, t) = \frac{1}{\sqrt{n}}\Big(\widetilde{\mathbf{H}}_{k2}(t)\Big)^T \boldsymbol{\Sigma}_k^{-1}\left(\mathbf{Z}_i - \mathbf{E}(\boldsymbol{\beta}_{0k}, u)\right),$$

with

$$\widetilde{H}_{11}(u, t) = \frac{1}{n}\sum_{i=1}^n \Bigg(\Big(S\big(u- \mid Z_A{=}1, \mathbf{Z}_{Li}\big) - F_1\big(t \mid Z_A{=}1, \mathbf{Z}_{Li}\big) + F_1\big(u \mid Z_A{=}1, \mathbf{Z}_{Li}\big)\Big) \cdot \exp\Big(\boldsymbol{\beta}_{01}^T\big(1, \mathbf{Z}_{Li}^T\big)^T\Big) - \Big(S\big(u- \mid Z_A{=}0, \mathbf{Z}_{Li}\big) - F_1\big(t \mid Z_A{=}0, \mathbf{Z}_{Li}\big) + F_1\big(u \mid Z_A{=}0, \mathbf{Z}_{Li}\big)\Big) \cdot \exp\Big(\boldsymbol{\beta}_{01}^T\big(0, \mathbf{Z}_{Li}^T\big)^T\Big)\Bigg)$$

as well as

$$\widetilde{H}_{k1}(u,t) = \frac{1}{n}\sum_{i=1}^{n}\Big(\Big(F_1\big(t \mid Z_A{=}0, \mathbf{Z}_{Li}\big) - F_1\big(u \mid Z_A{=}0, \mathbf{Z}_{Li}\big)\Big)\exp\Big(\boldsymbol{\beta}_{0k}^T\big(0, \mathbf{Z}_{Li}^T\big)^T\Big) - \Big(F_1\big(t \mid Z_A{=}1, \mathbf{Z}_{Li}\big) - F_1\big(u \mid Z_A{=}1, \mathbf{Z}_{Li}\big)\Big)\exp\Big(\boldsymbol{\beta}_{0k}^T\big(1, \mathbf{Z}_{Li}^T\big)^T\Big)\Big) \quad (k \in \{2,\ldots,K\}),$$

and

$$\widetilde{\mathbf{H}}_{12}(t) = \frac{1}{n}\sum_{i=1}^{n}\Big(\Big(\boldsymbol{\varphi}_1\big(t \mid Z_A{=}1, \mathbf{Z}_{Li}\big) - \boldsymbol{\psi}_{11}\big(t \mid Z_A{=}1, \mathbf{Z}_{Li}\big)\Big) - \Big(\boldsymbol{\varphi}_1\big(t \mid Z_A{=}0, \mathbf{Z}_{Li}\big) - \boldsymbol{\psi}_{11}\big(t \mid Z_A{=}0, \mathbf{Z}_{Li}\big)\Big)\Big)$$

along with

$$\widetilde{\mathbf{H}}_{k2}(t) = \frac{1}{n}\sum_{i=1}^{n}\Big(\boldsymbol{\psi}_{1k}\big(t \mid Z_A{=}0, \mathbf{Z}_{Li}\big) - \boldsymbol{\psi}_{1k}\big(t \mid Z_A{=}1, \mathbf{Z}_{Li}\big)\Big) \ (k \in \{2,\ldots,K\}).$$

All the definitions above allow us to proceed similarly as Cheng, Fine, and Wei (1998) in order to represent the limiting distribution of the process $\big(U_n(t)\big)$ based on the martingales $M_{ki}(t) = N_{ki}(t) - \int_0^t Y_i(u)\,\mathrm{d}A_k\big(u \mid Z_{Ai}, \mathbf{Z}_{Li}\big)$ ($k \in \{1,\ldots,K\}$, $i \in \{1,\ldots,n\}$). Asymptotics refer to the setting where $n \to \infty$ in the following.

Lemma 4.1 (Martingale representation of $U_n(t)$)**:**
It holds that

$$U_n(t) = \widetilde{U}_n(t) + o_P(1)$$

for $t \in [0,\tau]$, with

$$\widetilde{U}_n(t) = \sum_{k=1}^{K}\sum_{i=1}^{n}\Big(\int_0^t H_{k1i}(u,t)\,\mathrm{d}M_{ki}(u) + \int_0^\tau H_{k2i}(u,t)\,\mathrm{d}M_{ki}(u)\Big).$$

Proof:

We can describe the process $\big(U_n(t)\big)$ evaluated at time t by

$$U_n(t) = \sqrt{n}\,\Big(\frac{1}{n}\sum_{i=1}^{n}\Big(\hat{F}_1\big(t \mid Z_A{=}1, \mathbf{Z}_{Li}\big) - \hat{F}_1\big(t \mid Z_A{=}0, \mathbf{Z}_{Li}\big)\Big) - \mathbb{E}_{\mathbf{Z}_L}\Big(F_1\big(t \mid Z_A{=}1, \mathbf{Z}_L\big) - F_1\big(t \mid Z_A{=}0, \mathbf{Z}_L\big)\Big)\Big).$$

The strong law of large numbers suggests that

$$
\begin{aligned}
U_n(t) = \frac{\sqrt{n}}{n} \sum_{i=1}^{n} \Bigg(\Bigg(& \int_0^t \hat{S}\big(u- \mid Z_A = 1, \mathbf{Z}_{Li}\big)\, \mathrm{d}\hat{A}_1\big(u \mid Z_A = 1, \mathbf{Z}_{Li}\big) \\
& \quad - \int_0^t S\big(u- \mid Z_A = 1, \mathbf{Z}_{Li}\big)\, \mathrm{d}A_1\big(u \mid Z_A = 1, \mathbf{Z}_{Li}\big) \Bigg) \\
& - \Bigg(\int_0^t \hat{S}\big(u- \mid Z_A = 0, \mathbf{Z}_{Li}\big)\, \mathrm{d}\hat{A}_1\big(u \mid Z_A = 0, \mathbf{Z}_{Li}\big) \\
& \quad - \int_0^t S\big(u- \mid Z_A = 0, \mathbf{Z}_{Li}\big)\, \mathrm{d}A_1\big(u \mid Z_A = 0, \mathbf{Z}_{Li}\big) \Bigg) \Bigg) \\
& + o_P(1) \\
= \frac{\sqrt{n}}{n} \sum_{i=1}^{n} \Bigg(\Bigg(& \int_0^t \Big(\hat{S}\big(u- \mid Z_A = 1, \mathbf{Z}_{Li}\big) - S\big(u- \mid Z_A = 1, \mathbf{Z}_{Li}\big) \Big) \\
& \qquad \cdot \mathrm{d}\hat{A}_1\big(u \mid Z_A = 1, \mathbf{Z}_{Li}\big) \\
& \quad + \int_0^t S\big(u- \mid Z_A = 1, \mathbf{Z}_{Li}\big)\, \mathrm{d}\Big(\hat{A}_1\big(u \mid Z_A = 1, \mathbf{Z}_{Li}\big) \\
& \qquad - A_1\big(u \mid Z_A = 1, \mathbf{Z}_{Li}\big) \Big) \Bigg) \\
& - \Bigg(\int_0^t \Big(\hat{S}\big(u- \mid Z_A = 0, \mathbf{Z}_{Li}\big) - S\big(u- \mid Z_A = 0, \mathbf{Z}_{Li}\big) \Big) \\
& \qquad \cdot \mathrm{d}\hat{A}_1\big(u \mid Z_A = 0, \mathbf{Z}_{Li}\big) \\
& \quad + \int_0^t S\big(u- \mid Z_A = 0, \mathbf{Z}_{Li}\big)\, \mathrm{d}\Big(\hat{A}_1\big(u \mid Z_A = 0, \mathbf{Z}_{Li}\big) \\
& \qquad - A_1\big(u \mid Z_A = 0, \mathbf{Z}_{Li}\big) \Big) \Bigg) \Bigg) \Bigg) \\
& + o_P(1).
\end{aligned}
$$

Now consider the martingale representation of the process $\sqrt{n}\big(\hat{A}_k(t \mid a, \mathbf{l}) - A_k(t \mid a, \mathbf{l})\big)$ via

$$
\begin{aligned}
\widetilde{W}_k(t \mid a, \mathbf{l}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Bigg(& \int_0^t \frac{\exp\big(\boldsymbol{\beta}_{0k}^T \big(a, \mathbf{l}^T\big)^T\big)}{S^{(0)}(\boldsymbol{\beta}_{0k}, u)}\, \mathrm{d}M_{ki}(u) \\
& + \big(\mathbf{h}_k(t \mid a, \mathbf{l})\big)^T \boldsymbol{\Sigma}_k^{-1} \int_0^\tau \big(\mathbf{Z}_i - \mathbf{E}(\boldsymbol{\beta}_{0k}, u)\big)\, \mathrm{d}M_{ki}(u) \Bigg),
\end{aligned}
$$

as proposed by Lin, Fleming, and Wei (1994) (see Subsection 2.1.2). Together with the (uniform) consistency of the estimators $\hat{\boldsymbol{\beta}}_1$ and $\hat{A}_{01}$ (see Subsection 2.1.2), a first-order Taylor approximation of the function $f\colon x \mapsto \exp(-x)$ around $x = \sum_{k=1}^{K} A_k(t \mid a, \mathbf{l})$

yields

$$\hat{S}(t- \mid a, \mathbf{l}) - S(t- \mid a, \mathbf{l}) = -\frac{1}{\sqrt{n}}\, S(t- \mid a, \mathbf{l}) \sum_{k=1}^{K} \widetilde{W}_k(t \mid a, \mathbf{l}) + o_P(1),$$

and thus, noting that $S(t- \mid a, \mathbf{l})\, \mathrm{d}\hat{A}_1(t \mid a, \mathbf{l}) = \mathrm{d}F_1(t \mid a, \mathbf{l}) + o_P(1)$, we have

$$\begin{aligned} U_n(t) = \frac{1}{n} \sum_{i=1}^{n} \Bigg(\Bigg(& \int_0^t S\big(u- \mid Z_A{=}1, \mathbf{Z}_{Li}\big)\, \mathrm{d}\widetilde{W}_1\big(u \mid Z_A{=}1, \mathbf{Z}_{Li}\big) \\ & - \sum_{k=1}^{K} \int_0^t \widetilde{W}_k\big(u \mid Z_A{=}1, \mathbf{Z}_{Li}\big)\, \mathrm{d}F_1\big(u \mid Z_A{=}1, \mathbf{Z}_{Li}\big) \Bigg) \\ & - \Bigg(\int_0^t S\big(u- \mid Z_A{=}0, \mathbf{Z}_{Li}\big)\, \mathrm{d}\widetilde{W}_1\big(u \mid Z_A{=}0, \mathbf{Z}_{Li}\big) \\ & \quad - \sum_{k=1}^{K} \int_0^t \widetilde{W}_k\big(u \mid Z_A{=}0, \mathbf{Z}_{Li}\big)\, \mathrm{d}F_1\big(u \mid Z_A{=}0, \mathbf{Z}_{Li}\big) \Bigg) \Bigg) \\ & + o_P(1). \end{aligned}$$

Furthermore, integration by parts implies

$$\int_0^t \widetilde{W}_k(u \mid a, \mathbf{l})\, \mathrm{d}F_1(u \mid a, \mathbf{l}) = \widetilde{W}_k(t \mid a, \mathbf{l})\, F_1(t \mid a, \mathbf{l}) - \int_0^t F_1(u \mid a, \mathbf{l})\, \mathrm{d}\widetilde{W}_k(u \mid a, \mathbf{l}),$$

so that

$$\begin{aligned} U_n(t) = \frac{1}{n} \sum_{i=1}^{n} \Bigg(\Bigg(& \int_0^t S\big(u- \mid Z_A{=}1, \mathbf{Z}_{Li}\big)\, \mathrm{d}\widetilde{W}_1\big(u \mid Z_A{=}1, \mathbf{Z}_{Li}\big) \\ & - \sum_{k=1}^{K} \int_0^t \Big(F_1\big(t \mid Z_A{=}1, \mathbf{Z}_{Li}\big) - F_1\big(u \mid Z_A{=}1, \mathbf{Z}_{Li}\big) \Big) \\ & \qquad \cdot \mathrm{d}\widetilde{W}_k\big(u \mid Z_A{=}1, \mathbf{Z}_{Li}\big) \Bigg) \\ & - \Bigg(\int_0^t S\big(u- \mid Z_A{=}0, \mathbf{Z}_{Li}\big)\, \mathrm{d}\widetilde{W}_1\big(u \mid Z_A{=}0, \mathbf{Z}_{Li}\big) \\ & \quad - \sum_{k=1}^{K} \int_0^t \Big(F_1\big(t \mid Z_A{=}0, \mathbf{Z}_{Li}\big) - F_1\big(u \mid Z_A{=}0, \mathbf{Z}_{Li}\big) \Big) \\ & \qquad \cdot \mathrm{d}\widetilde{W}_k\big(u \mid Z_A{=}0, \mathbf{Z}_{Li}\big) \Bigg) \Bigg) \\ & + o_P(1). \end{aligned}$$

The desired representation is eventually obtained by inserting the definition of $\widetilde{W}_k$ and reordering the terms. □

Let the functions $\tilde{h}_{k1}$ and $\tilde{\mathbf{h}}_{k2}$ be defined as the large-sample limits of $\widetilde{H}_{k1}$ and $\widetilde{\mathbf{H}}_{k2}$, respectively, e.g.

$$\tilde{\mathbf{h}}_{12}(u,t) = \mathbb{E}_{\mathbf{Z}_L}\Big(\Big(\boldsymbol{\varphi}_1\big(t \mid Z_A{=}1, \mathbf{Z}_L\big) - \boldsymbol{\psi}_{11}\big(t \mid Z_A{=}1, \mathbf{Z}_L\big)\Big) - \Big(\boldsymbol{\varphi}_1\big(t \mid Z_A{=}0, \mathbf{Z}_L\big) - \boldsymbol{\psi}_{11}\big(t \mid Z_A{=}0, \mathbf{Z}_L\big)\Big)\Big).$$

Lemma 4.1 provides the basis to prove the subsequent core statement:

Theorem 4.1 (Asymptotic distribution of $\big(U_n(t)\big)$)**:**
The process $\big(U_n(t)\big)$ converges weakly to a zero-mean Gaussian process with covariance function $\xi(t_1,t_2) = \sum_{k=1}^K \xi^{(k)}(t_1,t_2)$,

$$\xi^{(k)}(t_1,t_2) = \int_0^{t_1\wedge t_2} \tilde{h}_{k1}(u,t_1)\,\tilde{h}_{k1}(u,t_2)\,\frac{\mathrm{d}A_{0k}(u)}{s^{(0)}(\boldsymbol{\beta}_{0k},u)} + \big(\tilde{\mathbf{h}}_{k2}(t_1)\big)^T \boldsymbol{\Sigma}_k^{-1}\,\tilde{\mathbf{h}}_{k2}(t_2)\,,$$

on the Skorokhod space $D[0,\tau]$.

Proof:

We treat the covariate vectors $\mathbf{Z}_i$, $i \in \{1,\ldots,n\}$, as fixed from now on. According to Lemma 4.1, it is sufficient to study the limiting distribution of the process $\big(\widetilde{U}_n(t)\big)$.

Note that the counting processes N_{ki} and N_{li} associated with subject i cannot jump at the same time for distinct causes $k \neq l$, which means that the martingales $M_{ki}(t)$ and $M_{li}(t)$ are orthogonal. In addition, w.r.t. the predictable covariation process given below, we find that

$$\begin{aligned}
&\Big\langle \sum_{i=1}^n \int_0^{\cdot} \frac{1}{\sqrt{n}\,S^{(0)}(\boldsymbol{\beta}_{0k},u)}\,\mathrm{d}M_{ki}(u),\ \sum_{i=1}^n \int_0^{\cdot} \frac{1}{\sqrt{n}}\,\big(\mathbf{Z}_i - \mathbf{E}(\boldsymbol{\beta}_{0k},u)\big)\,\mathrm{d}M_{ki}(u)\Big\rangle(t) \\
&\quad= \frac{1}{n}\sum_{i=1}^n \int_0^t \frac{1}{S^{(0)}(\boldsymbol{\beta}_{0k},u)}\,\big(\mathbf{Z}_i - \mathbf{E}(\boldsymbol{\beta}_{0k},u)\big)\,Y_i(u)\exp\big(\boldsymbol{\beta}_{0k}^T\mathbf{Z}_i\big)\,\mathrm{d}A_{0k}(u) \\
&\quad= \int_0^t \frac{1}{S^{(0)}(\boldsymbol{\beta}_{0k},u)}\,\Big(\mathbf{S}^{(1)}(\boldsymbol{\beta}_{0k},u) - \mathbf{E}(\boldsymbol{\beta}_{0k},u)\,S^{(0)}(\boldsymbol{\beta}_{0k},u)\Big)\,\mathrm{d}A_{0k}(u) \\
&\quad= \int_0^t \Big(\mathbf{E}(\boldsymbol{\beta}_{0k},u) - \mathbf{E}(\boldsymbol{\beta}_{0k},u)\Big)\,\mathrm{d}A_{0k}(u) = 0,
\end{aligned} \tag{4.1}$$

so that the terms $\sum_{i=1}^n \int_0^t H_{k1i}(u,t)\,\mathrm{d}M_{ki}(u)$ and $\sum_{i=1}^n \int_0^t H_{k2i}(u,t)\,\mathrm{d}M_{ki}(u)$ referring to cause $k \in \{1,\ldots,K\}$ are also orthogonal by the definitions of H_{k1i} and H_{k2i}.

Now recall the asymptotic normality of the score function at $\boldsymbol{\beta}_{0k}$ (see Subsection 2.1.2). Together with the fact that the functions $\boldsymbol{\varphi}_1$ and $\boldsymbol{\psi}_{1k}$ are deterministic, it follows that the second summand $\sum_{k=1}^K \sum_{i=1}^n \int_0^\tau H_{k2i}(u,t)\,\mathrm{d}M_{ki}(u)$ in $\widetilde{U}_n$ – as a sum of asymptotically normal distributed terms with mean zero – is likewise asymptotically normal with mean zero.

What remains to be examined is the limiting distribution of the first summand. The processes $\widetilde{H}_{k1}(u,t)$, $k \in \{1,\dots,K\}$, are deterministic and continuous in $u \leq t$, with

$$\begin{aligned}\left|\widetilde{H}_{k1}(u,t)\right| &\leq \big(\exp(\beta_{0kA})+1\big) \max_{i\in\{1,\dots,n\}} \exp\big(\boldsymbol{\beta}_{0kL}^T \mathbf{Z}_{Li}\big),\\ \left|\Big(\widetilde{H}_{k1}(u,t)\Big)^2\right| &\leq \big(\exp(2\beta_{0kA})+2\exp(\beta_{0kA})+1\big) \max_{i,j\,\in\,\{1,\dots,n\}} \exp\big(\boldsymbol{\beta}_{0kL}^T(\mathbf{Z}_{Li}+\mathbf{Z}_{Lj})\big).\end{aligned}$$

Besides, provided that $P\big(Y_i(t)=1\big)>0 \;\; \forall i \in \{1,\dots,n\}, \;\; t\in[0,\tau]$, the function $S^{(0)}(\boldsymbol{\beta}_{0k},t)$ converges uniformly to $s^{(0)}(\boldsymbol{\beta}_{0k},t)$ in probability on $[0,\tau]$, with the limit $s^{(0)}(\boldsymbol{\beta}_{0k},t)$ being bounded away from zero (see Subsection 2.1.2). The conditions of Theorem 2.1 are hence fulfilled w.r.t. the process $\big(H_{k1i}(u,t)\big)_{u\in[0,t]}$ for $k\in\{1,\dots,K\}$ and $t\in[0,\tau]$: It holds that

$$\sum_{i=1}^n \int_0^t \big(H_{k1i}(u,t)\big)^2 Y_i(u) \exp\big(\boldsymbol{\beta}_{0k}^T \mathbf{Z}_i\big)\,\mathrm{d}A_{0k}(u) \xrightarrow{P} \int_0^t \frac{\big(\tilde{h}_{k1}(u,t)\big)^2}{s^{(0)}(\boldsymbol{\beta}_{0k},u)}\,\mathrm{d}A_{0k}(u)$$

based on the definitions of H_{k1i} and $S^{(0)}$, and $|H_{k1i}(u,t)| \xrightarrow{P} 0 \;\; \forall i\in\{1,\dots,n\}$ as $n\to\infty$. One may deduce that the stochastic integral $\sum_{i=1}^n \int H_{k1i}(u,\cdot)\,\mathrm{d}M_{ki}(u)$ converges weakly to a Gaussian process with mean zero on $D[0,\tau]$, and by the previous considerations, so does $\big(\widetilde{U}_n\big)$.

As a final step, we derive the covariance function $\tilde{\xi}$ of $\big(\widetilde{U}_n\big)$. Because of the equivalence

$$\sum_{i=1}^n \int_0^{\cdot} H_{k1i}(u,t_1)\, H_{k2i}(u,t_2)\, Y_i(u) \exp\big(\boldsymbol{\beta}_{0k}^T \mathbf{Z}_i\big) = 0$$

for $k\in\{1,\dots,K\}$, $t_1,t_2\in[0,\tau]$ (cf. Equation (4.1)), it is

$$\begin{aligned}\tilde{\xi}(t_1,t_2) = \Big\langle &\sum_{k=1}^K \sum_{i=1}^n \left(\int_0^{\cdot} \big(H_{k1i}(u,t_1)\,\mathbb{1}\{u\leq t_1\} + H_{k2i}(u,t_1)\big)\,\mathrm{d}M_{ki}(u)\right),\\ &\sum_{k=1}^K \sum_{i=1}^n \left(\int_0^{\cdot} \big(H_{k1i}(u,t_2)\,\mathbb{1}\{u\leq t_2\} + H_{k2i}(u,t_2)\big)\,\mathrm{d}M_{ki}(u)\right)\Big\rangle(\tau)\\ = \sum_{k=1}^K \Bigg(&\sum_{i=1}^n \int_0^{t_1\wedge t_2} \frac{\widetilde{H}_{k1}(u,t_1)\,\widetilde{H}_{k1}(u,t_2)}{n\left(S^{(0)}(\boldsymbol{\beta}_{0k},u)\right)^2} Y_i(u)\exp\big(\boldsymbol{\beta}_{0k}^T\mathbf{Z}_i\big)\,\mathrm{d}A_{0k}(u)\\ &+\sum_{i=1}^n \int_0^\tau \frac{1}{n}\Big(\widetilde{\mathbf{H}}_{k2}(t_1)\Big)^T \boldsymbol{\Sigma}_k^{-1}\big(\mathbf{Z}_i-\mathbf{E}(\boldsymbol{\beta}_{0k},u)\big)\\ &\qquad\cdot\Big(\widetilde{\mathbf{H}}_{k2}(t_2)\Big)^T \boldsymbol{\Sigma}_k^{-1}\big(\mathbf{Z}_i-\mathbf{E}(\boldsymbol{\beta}_{0k},u)\big)\\ &\qquad\qquad\cdot Y_i(u)\exp\big(\boldsymbol{\beta}_{0k}^T\mathbf{Z}_i\big)\,\mathrm{d}A_{0k}(u)\Bigg)\end{aligned}$$

$$= \sum_{k=1}^{K} \Bigg(\int_0^{t_1 \wedge t_2} \frac{\widetilde{H}_{k1}(u,t_1)\,\widetilde{H}_{k1}(u,t_2)}{S^{(0)}(\boldsymbol{\beta}_{0k},u)}\,\mathrm{d}A_{0k}(u) + \Big(\widetilde{\mathbf{H}}_{k2}(t_1)\Big)^T \boldsymbol{\Sigma}_k^{-1} \Bigg(\int_0^{\tau} \frac{1}{n} \sum_{i=1}^{n} \big(\mathbf{Z}_i - \mathbf{E}(\boldsymbol{\beta}_{0k},u)\big)\big(\mathbf{Z}_i - \mathbf{E}(\boldsymbol{\beta}_{0k},u)\big)^T \cdot Y_i(u) \exp\Big(\boldsymbol{\beta}_{0k}^T \mathbf{Z}_i\Big)\,\mathrm{d}A_{0k}(u) \Bigg) \cdot \Big(\boldsymbol{\Sigma}_k^{-1}\Big)^T \widetilde{\mathbf{H}}_{k2}(t_2) \Bigg).$$

Noting that

$$\begin{aligned} &\frac{1}{n} \sum_{i=1}^{n} \big(\mathbf{Z}_i - \mathbf{E}(\boldsymbol{\beta}_{0k},u)\big)\big(\mathbf{Z}_i - \mathbf{E}(\boldsymbol{\beta}_{0k},u)\big)^T Y_i(u) \exp\Big(\boldsymbol{\beta}_{0k}^T \mathbf{Z}_i\Big) \\ &= \mathbf{S}^{(2)}(\boldsymbol{\beta}_{0k},u) - \mathbf{S}^{(1)}(\boldsymbol{\beta}_{0k},u)\big(\mathbf{E}(\boldsymbol{\beta}_{0k},u)\big)^T - \mathbf{E}(\boldsymbol{\beta}_{0k},u)\Big(\mathbf{S}^{(1)}(\boldsymbol{\beta}_{0k},u)\Big)^T \\ &\qquad + \big(\mathbf{E}(\boldsymbol{\beta}_{0k},u)\big)^{\otimes 2} S^{(0)}(\boldsymbol{\beta}_{0k},u) \\ &= \left(\frac{\mathbf{S}^{(2)}(\boldsymbol{\beta}_{0k},u)}{S^{(0)}(\boldsymbol{\beta}_{0k},u)} - \big(\mathbf{E}(\boldsymbol{\beta}_{0k},u)\big)^{\otimes 2} \right) S^{(0)}(\boldsymbol{\beta}_{0k},u)\,, \end{aligned}$$

we have

$$\tilde{\xi}(t_1,t_2) = \sum_{k=1}^{K} \Bigg(\int_0^{t_1 \wedge t_2} \frac{\widetilde{H}_{k1}(u,t_1)\,\widetilde{H}_{k1}(u,t_2)}{S^{(0)}(\boldsymbol{\beta}_{0k},u)}\,\mathrm{d}A_{0k}(u) + \Big(\widetilde{\mathbf{H}}_{k2}(t_1)\Big)^T \boldsymbol{\Sigma}_k^{-1} \Bigg(\int_0^{\tau} \left(\frac{\mathbf{S}^{(2)}(\boldsymbol{\beta}_{0k},u)}{S^{(0)}(\boldsymbol{\beta}_{0k},u)} - \big(\mathbf{E}(\boldsymbol{\beta}_{0k},u)\big)^{\otimes 2} \right) \cdot S^{(0)}(\boldsymbol{\beta}_{0k},u)\,\mathrm{d}A_{0k}(u) \Bigg) \Big(\boldsymbol{\Sigma}_k^{-1}\Big)^T \widetilde{\mathbf{H}}_{k2}(t_2) \Bigg),$$

and by the strong law of large number as well as the continuous mapping theorem, the uniform convergence of $\mathbf{S}^{(r)}(\boldsymbol{\beta}_{0k},t)$, $r \in \{0,1,2\}$, together with Equation (2.6), yields that $\tilde{\xi}(t_1,t_2) \xrightarrow{P} \xi(t_1,t_2)$. □

Thus, the asymptotic distribution of the process $\big(U_n(t)\big)$ has been established. We proceed from Theorem 4.1 hereafter to show that this distribution can be approximated by different resampling approaches.

In the following, let $\widehat{\mathbf{h}}_k$, $\widehat{\boldsymbol{\varphi}}_1$, $\widehat{\boldsymbol{\psi}}_{1k}$, $\widehat{\widetilde{H}}_{k1}$, and $\widehat{\widetilde{\mathbf{H}}}_{k2}$ refer to the plug-in estimators of $\mathbf{h}_k$, $\boldsymbol{\varphi}_1$, $\boldsymbol{\psi}_{1k}$, $\widetilde{H}_{k1}$, and $\widetilde{\mathbf{H}}_{k2}$, which are obtained by replacing $\boldsymbol{\beta}_{0k}$, A_{0k}, A_k, $\mathbf{s}^{(r)}$, S, F_1 with $\hat{\boldsymbol{\beta}}_k$, $\hat{A}_{0k}$, $\hat{A}_k$, $\mathbf{S}^{(r)}$, $\hat{S}$, $\hat{F}_1$, respectively ($k \in \{1,\ldots,K\}$, $r \in \{0,1,2\}$). Furthermore, we introduce

$$\widehat{\boldsymbol{\Sigma}}_k = \frac{1}{n} \int_0^{\tau} \left(\frac{\mathbf{S}^{(2)}\big(\hat{\boldsymbol{\beta}}_k,u\big)}{S^{(0)}\big(\hat{\boldsymbol{\beta}}_k,u\big)} - \Big(\mathbf{E}\big(\hat{\boldsymbol{\beta}}_k,u\big)\Big)^{\otimes 2} \right) \mathrm{d}N_k(u).$$

Efron's nonparametric bootstrap (EBS)

Our first focal point is the classical nonparametric bootstrap.

Theorem 4.2 (Approximation of the distribution of $\big(U_n(t)\big)$ by the EBS)**:**
The process

$$\widehat{U}_n^{EBS}(t) = \sqrt{n}\,\Big(\widehat{ATE}^{*}_{ds}(t) - \widehat{ATE}_{ds}(t)\Big),$$

with $\widehat{ATE}^{*}_{ds}(t)$ *being the ATE estimated in the bootstrap sample, converges weakly to the same limiting process as* $\big(U_n(t)\big)$ *for almost all data samples* $\big((T_i \wedge C_i, D_i, \mathbf{Z}_i)\big)_{i\in\{1,\ldots,n\}}$ *if* $\inf_{t\in[0,\tau]} Y(t) \xrightarrow{P} \infty$.

As the EBS is the most widely applied resampling approach for making inferences about the ATE, it is important to prove its validity in the setting at hand.

Proof:
Akritas (1986) pointed out that the general martingale arguments are valid conditional on ω for almost all outcomes ω in the sample space Ω. Let $\big((T_i \wedge C_i, D_i, \mathbf{Z}_i)\big)_{i\in\{1,\ldots,n\}}$ be the data sample corresponding to such an $\omega \in \Omega$, and suppose that the bootstrap sample $\Big(\big((T\wedge C)^*_i, D^*_i, \mathbf{Z}^*_i\big)\Big)_{i\in\{1,\ldots,n\}}$ is obtained according to the approach described in Subsection 2.3.1. Following the argumentation of Akritas (1986), the target functions in the bootstrap sample correspond to the estimators in the original sample:

$$A^*_{0k}(t) = \hat{A}_{0k}(t), \qquad A^*_k(t \mid a, \mathbf{l}) = \hat{A}_k(t \mid a, \mathbf{l}),$$

and

$$\mathbf{s}^{(r)^*}(\boldsymbol{\beta}, u) = \mathbf{S}^{(r)}(\boldsymbol{\beta}, u), \qquad \mathbf{e}^*(\boldsymbol{\beta}, t) = \mathbf{E}(\boldsymbol{\beta}, t),$$

for $k \in \{1, \ldots, K\}$, $r \in \{0, 1, 2\}$ and $t \in [0, \tau^*]$, with $\tau^* = \max_{i\in\{1,\ldots,n\}} (T\wedge C)^*_i$. The validity of the latter two equivalences is easy to see if we characterize the resampled data via multinomial weights assigned to the original observations. Consider e.g.

$$S^{(0)^*}(\boldsymbol{\beta}, t) = \frac{1}{n}\sum_{i=1}^{n} w_i\, Y_i(t)\exp\Big(\boldsymbol{\beta}^T\mathbf{Z}_i\Big),$$

$(w_1, \ldots, w_n)^T \sim \mathcal{M}\Big(n, \big(1/n, \ldots, 1/n\big)^T\Big)$, which implies $\mathbb{E}\Big(S^{(0)^*}(\boldsymbol{\beta}, t)\Big) = S^{(0)}(\boldsymbol{\beta}, t)$, given a fixed data sample $\big((T_i \wedge C_i, D_i, \mathbf{Z}_i)\big)_{i\in\{1,\ldots,n\}}$. It follows that

$$\boldsymbol{\Sigma}^*_k = \widehat{\boldsymbol{\Sigma}}_k$$

as well as

$$\mathbf{h}^*_k(t \mid a, \mathbf{l}) = \widehat{\mathbf{h}}_k(t \mid a, \mathbf{l}), \quad \boldsymbol{\varphi}^*_1(t \mid a, \mathbf{l}) = \widehat{\boldsymbol{\varphi}}_1(t \mid a, \mathbf{l}), \quad \boldsymbol{\psi}^*_{1k}(t \mid a, \mathbf{l}) = \widehat{\boldsymbol{\psi}}_{1k}(t \mid a, \mathbf{l}).$$

According to the reasoning of Prentice and Kalbfleisch (2003), we derive the asymptotic equivalences $\hat{\boldsymbol{\beta}}_k^* = \hat{\boldsymbol{\beta}}_k + o_P(1)$ and $\hat{A}_{0k}^*(t) = \hat{A}_{0k}(t) + o_P(1)$ on $[0, \tau^*]$ as $n \to \infty$. Moreover, one can show that

$$\begin{aligned}
&\sqrt{n}\Big(\hat{A}_k^*(t \mid a, \mathbf{l}) - \hat{A}_k(t \mid a, \mathbf{l})\Big) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Bigg(\int_0^t \frac{\exp\Big(\hat{\boldsymbol{\beta}}_k^T \left(a, \mathbf{l}^T\right)^T\Big)}{S^{(0)^*}\big(\hat{\boldsymbol{\beta}}_k, u\big)} \, M_{ki}^*(\mathrm{d}u) \\
&\qquad\qquad + \Big(\widehat{\mathbf{h}}_k(t \mid a, \mathbf{l})\Big)^T \widehat{\boldsymbol{\Sigma}}_k^{-1} \int_0^{\tau^*} \Big(\mathbf{Z}_i - \mathbf{E}^*\big(\hat{\boldsymbol{\beta}}_k, u\big)\Big) \, M_{ki}^*(\mathrm{d}u) \Bigg) \\
&\quad + o_P(1)
\end{aligned}$$

by transferring the ideas of Lin, Fleming, and Wei (1994) to the situation at hand. A time-discrete setting is considered here, and our focus is on the (discrete-time) martingales $M_{ki}^*(t) = w_i \left(N_{ki}(t) - \int_0^t Y_i(u) \, \mathrm{d}\hat{A}_k\big(u \mid Z_{Ai}, \mathbf{Z}_{Li}\big)\right)$, $k \in \{1, \ldots, K\}$, $i \in \{1, \ldots, n\}$.

Let the functions H_{k1i}^* and H_{k2i}^* be defined analogously to H_{k1i} and H_{k2i}, but on the basis of the bootstrap sample $\Big(\big((T \wedge C)_i^*, D_i^*, \mathbf{Z}_i^*\big)\Big)_{i \in \{1, \ldots, n\}}$. The findings above allow us to proceed as in the proof of Lemma 4.1 in order to demonstrate that

$$\widehat{U}_n^{EBS}(t) = \sum_{k=1}^{K} \sum_{i=1}^{n} \Bigg(\int_0^t H_{k1i}^*(u, t) \, M_{ki}^*(\mathrm{d}u) + \int_0^{\tau^*} H_{k2i}^*(u, t) \, M_{ki}^*(\mathrm{d}u) \Bigg) + o_P(1)$$

on $[0, \tau^*]$.

We will continue similarly as it has already been done in the proof of Theorem 4.1. First, note the considerations of Gill (1980, Theorem 2.3.1) on discrete-time martingales. It follows that

$$\begin{aligned}
&\Big\langle \sum_{i=1}^{n} \int_0^{\cdot} \frac{1}{\sqrt{n} \, S^{(0)^*}\big(\hat{\boldsymbol{\beta}}_k, u\big)} \, M_{ki}^*(\mathrm{d}u), \; \sum_{i=1}^{n} \int_0^{\cdot} \frac{1}{\sqrt{n}} \Big(\mathbf{Z}_i - \mathbf{E}^*\big(\hat{\boldsymbol{\beta}}_k, u\big)\Big) \, M_{ki}^*(\mathrm{d}u) \Big\rangle(t) \\
&= \frac{1}{n} \sum_{i=1}^{n} \int_0^t \frac{1}{S^{(0)^*}\big(\hat{\boldsymbol{\beta}}_k, u\big)} \Big(\mathbf{Z}_i - \mathbf{E}^*\big(\hat{\boldsymbol{\beta}}_k, u\big)\Big) \, w_i \, Y_i(u) \Big(1 - \hat{A}_k\big(\Delta u \mid Z_{Ai}, \mathbf{Z}_{Li}\big)\Big) \\
&\qquad\qquad \cdot \hat{A}_k\big(\mathrm{d}u \mid Z_{Ai}, \mathbf{Z}_{Li}\big) \\
&\quad = \int_0^t \Big(\mathbf{E}^*\big(\hat{\boldsymbol{\beta}}_k, u\big) - \mathbf{E}^*\big(\hat{\boldsymbol{\beta}}_k, u\big)\Big) \, \hat{A}_{0k}(\mathrm{d}u) \\
&\qquad - \frac{1}{n} \sum_{i=1}^{n} \int_0^t \frac{1}{S^{(0)^*}\big(\hat{\boldsymbol{\beta}}_k, u\big)} \Big(\mathbf{Z}_i - \mathbf{E}^*\big(\hat{\boldsymbol{\beta}}_k, u\big)\Big) \, w_i \, Y_i(u) \, \exp\Big(2\hat{\boldsymbol{\beta}}_k^T \mathbf{Z}_i\Big) \\
&\qquad\qquad \cdot \hat{A}_{0k}(\Delta u) \, \hat{A}_{0k}(\mathrm{d}u),
\end{aligned}$$

and this expression is equal to

$$\int_0^t \frac{1}{S^{(0)}\big(\hat{\boldsymbol{\beta}}_k, u\big)} \Big(\mathbf{E}\big(\hat{\boldsymbol{\beta}}_k, u\big) S_2^{(0)}\big(\hat{\boldsymbol{\beta}}_k, u\big) - \mathbf{S}_2^{(1)}\big(\hat{\boldsymbol{\beta}}_k, u\big) \Big) \, \hat{A}_{0k}(\Delta u) \, \hat{A}_{0k}(\mathrm{d}u) + o_P(1) \quad (4.2)$$

for $\mathbf{S}_2^{(r)}(\boldsymbol{\beta}, u) = \frac{1}{n} \sum_{i=1}^n Y_i(t) \exp\big(2\boldsymbol{\beta}^T \mathbf{Z}_i\big) \, \mathbf{Z}_i^{\otimes r}$, $r \in \{0, 1, 2\}$.

One may further exploit the arguments given in the proof of Theorem 4.1 as well as in Theorem 2.4.1 by Gill (1980) to infer that the limiting distribution of the stochastic integral $\sum_{i=1}^n \int H_{k1i}^*(u, \cdot) \, M_{ki}^*(\mathrm{d}u)$ is a zero-mean Gaussian process with covariance function asymptotically equivalent to

$$\begin{aligned} \sum_{k=1}^K \int_0^{t_1 \wedge t_2} \Bigg(& \frac{\widehat{\widehat{H}}_{k1}(u, t_1) \, \widehat{\widehat{H}}_{k1}(u, t_2)}{S^{(0)}\big(\hat{\boldsymbol{\beta}}_k, u\big)} \, \hat{A}_{0k}(\mathrm{d}u) \\ & - \frac{\widehat{\widehat{H}}_{k1}(u, t_1) \, \widehat{\widehat{H}}_{k1}(u, t_2)}{\Big(S^{(0)}\big(\hat{\boldsymbol{\beta}}_k, u\big)\Big)^2} \, S_2^{(0)}\big(\hat{\boldsymbol{\beta}}_k, u\big) \, \hat{A}_{0k}(\Delta u) \, \hat{A}_{0k}(\mathrm{d}u) \Bigg). \end{aligned} \quad (4.3)$$

Eventually, Prentice and Kalbfleisch (2003) showed the normality of

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \int_0^{\tau^*} \Big(\mathbf{Z}_i - \mathbf{E}^*\big(\hat{\boldsymbol{\beta}}_k, u\big) \Big) \, M_{ki}^*(\mathrm{d}u)$$

as $n \to \infty$. The mean of this expression tends to the null vector and the covariance matrix can be approximated by $\widehat{\boldsymbol{\Sigma}}_k - \widehat{\boldsymbol{\Sigma}}_{k,2}$, where $\widehat{\boldsymbol{\Sigma}}_{k,2}$ is defined by

$$\begin{aligned} \int_0^{\tau^*} \Bigg(& \mathbf{S}_2^{(2)}\big(\hat{\boldsymbol{\beta}}_k, u\big) - \mathbf{S}_2^{(1)}\big(\hat{\boldsymbol{\beta}}_k, u\big) \Big(\mathbf{E}\big(\hat{\boldsymbol{\beta}}_k, u\big)\Big)^T - \mathbf{E}\big(\hat{\boldsymbol{\beta}}_k, u\big) \Big(\mathbf{S}_2^{(1)}\big(\hat{\boldsymbol{\beta}}_k, u\big)\Big)^T \\ & + \mathbf{E}\big(\hat{\boldsymbol{\beta}}_k, u\big) \Big(\mathbf{E}\big(\hat{\boldsymbol{\beta}}_k, u\big)\Big)^T S_2^{(0)}\big(\hat{\boldsymbol{\beta}}_k, u\big) \Bigg) \, \hat{A}_{0k}(\Delta u) \, \hat{A}_{0k}(\mathrm{d}u). \end{aligned} \quad (4.4)$$

Now, bearing in mind that there are no ties in the original sample, we deduce the inequality

$$\begin{aligned} & \mathbf{S}_2^{(r)}\big(\hat{\boldsymbol{\beta}}_k, t\big) \, \hat{A}_{0k}(\Delta t) \, \hat{A}_{0k}(\mathrm{d}t) \\ = \; & \frac{1}{n} \sum_{i=1}^n Y_i(t) \exp\big(2\hat{\boldsymbol{\beta}}_k^T \mathbf{Z}_i\big) \, \mathbf{Z}_i^{\otimes r} \frac{\Delta N_k(t) \, \mathrm{d}N_k(t)}{\Big(\sum_{i=1}^n Y_i(t) \exp\big(\hat{\boldsymbol{\beta}}_k^T \mathbf{Z}_i\big)\Big)^2} \\ \leq \; & \frac{\max_{i \in \{1, \ldots, n\} \colon Y_i(t) = 1} \Big\{ \exp\big(2\hat{\boldsymbol{\beta}}_k^T \mathbf{Z}_i\big) \, \mathbf{Z}_i^{\otimes r} \Big\}}{\big(Y(t)\big)^2 \min_{i \in \{1, \ldots, n\} \colon Y_i(t) = 1} \Big\{ \exp\big(2\hat{\boldsymbol{\beta}}_k^T \mathbf{Z}_i\big) \Big\}} \end{aligned}$$

for $k \in \{1, \ldots, K\}$, $r \in \{0, 1, 2\}$, $t \in [0, \tau^*]$. It follows that $\mathbf{S}_2^{(r)}\big(\hat{\boldsymbol{\beta}}_k, t\big)\, \hat{A}_{0k}(\Delta t)\, \hat{A}_{0k}(\mathrm{d}t)$ vanishes as $n \to \infty$, for $r \in \{0, 1, 2\}$, due to the boundedness of the covariates, and thus, the expression in Equation (4.2), the subtrahend in Equation (4.3), as well as the matrix in Equation (4.4) do likewise.

We finally conclude that $\widehat{U}_n^{EBS}(t)$ converges weakly to a zero-mean Gaussian process with covariance function ξ on $D[0, \tau^*]$, using the arguments from the proof of Theorem 4.1. □

Influence function approach (IF)

Scheike and Zhang (2008) proposed another resampling approach that relies on the theory revolving around IFs.

Theorem 4.3 (Approximation of the distribution of $\big(U_n(t)\big)$ by the IF)**:**
For i.i.d. multipliers $G_i \sim \mathcal{N}(0, 1)$, $i \in \{1, \ldots, n\}$, the process

$$\widehat{U}_n^{IF}(t) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \widehat{IF}_{ATE}(t;\, T_i \wedge C_i, D_i, \mathbf{Z}_i)\, G_i$$

converges weakly to the same limiting process as $\big(U_n(t)\big)$ on $D[0, \tau]$, conditional on the data $\big((T_i \wedge C_i, D_i, \mathbf{Z}_i)\big)_{i \in \{1,\ldots,n\}}$.
The definition of $\widehat{IF}_{ATE}$ is delineated in the subsequent proof.

Proof:
Let $F_{\mathbf{O}}$ denote the CDF of the random vector $\mathbf{O} = \big(T \wedge C, D, \mathbf{Z}^T\big)^T$ that gathers the available data. One finds that $\sqrt{n}\Big(\widehat{ATE}_{ds} - ATE\Big) \xrightarrow{\mathscr{D}} \partial ATE_{F_{\mathbf{O}}}\big(W(F_{\mathbf{O}})\big)$ on $D[0, \tau]$ by Theorem 2.2 and Donsker's theorem. (We label the Brownian motion by $\big(W(t)\big)$.)

On the other hand, due to the linearity of the Hadamard derivative, the process $\widehat{U}_n^{IF}(\cdot)$ can be expressed as

$$\begin{aligned}
&\frac{1}{\sqrt{n}} \sum_{i=1}^{n} IF_{ATE}(\cdot; \mathbf{O}_i)\big|_{F_{\mathbf{O}} = \hat{F}_{\mathbf{O}}}\, G_i \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \partial ATE_{F_{\mathbf{O}}}\big(\hat{F}_{\mathbf{O}_i} - F_{\mathbf{O}}\big)\big|_{F_{\mathbf{O}} = \hat{F}_{\mathbf{O}}} G_i \\
&= \partial ATE_{F_{\mathbf{O}}}\bigg(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \big(\hat{F}_{\mathbf{O}_i} - F_{\mathbf{O}}\big)\, G_i\bigg)\bigg|_{F_{\mathbf{O}} = \hat{F}_{\mathbf{O}}},
\end{aligned}$$

where $\hat{F}_{\mathbf{O}_i} = \mathbb{1}_{[T_i \wedge C_i, \infty] \times \{D_i, D_i+1, \ldots, K\} \times \{Z_{Ai}, 1\} \times [Z_{Li_1}, \infty] \times \cdots \times [Z_{Li_p}, \infty]}$ and $\hat{F}_{\mathbf{O}} = \frac{1}{n} \sum_{i=1}^{n} \hat{F}_{\mathbf{O}_i}$. It is easy to see that for fixed data, $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \big(\hat{F}_{\mathbf{O}_i} - F_{\mathbf{O}}\big)\, G_i$ converges to a Gaussian process with mean zero and a covariance function characterized by

$$
\begin{aligned}
&\operatorname{Cov}\Bigg(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \big(\hat{F}_{\mathbf{O}_i}(t_1, d_1, a_1, \mathbf{l}_1) - F_{\mathbf{O}}(t_1, d_1, a_1, \mathbf{l}_1) \big) G_i, \\
&\qquad \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \big(\hat{F}_{\mathbf{O}_i}(t_2, d_2, a_2, \mathbf{l}_2) - F_{\mathbf{O}}(t_2, d_2, a_2, \mathbf{l}_2) \big) G_i \Bigg) \\
&\overset{G_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)}{=} \Bigg(\frac{1}{n} \sum_{i=1}^{n} \big(\hat{F}_{\mathbf{O}_i}(t_1, d_1, a_1, \mathbf{l}_1) - F_{\mathbf{O}}(t_1, d_1, a_1, \mathbf{l}_1) \big) \\
&\qquad \cdot \big(\hat{F}_{\mathbf{O}_i}(t_2, d_2, a_2, \mathbf{l}_2) - F_{\mathbf{O}}(t_2, d_2, a_2, \mathbf{l}_2) \big) \Bigg) \\
&\xrightarrow{\text{a.s.}} F_{\mathbf{O}}\left(t_1 \wedge t_2, d_1 \wedge d_2, a_1 \wedge a_2, l_{1_1} \wedge l_{2_1}, \ldots, l_{1_p} \wedge l_{2_p}\right) - F_{\mathbf{O}}(t_1, d_1, a_1, \mathbf{l}_1)\, F_{\mathbf{O}}(t_2, d_2, a_2, \mathbf{l}_2)
\end{aligned}
$$

on $D[0, \tau]$. This process coincides exactly with $W(F_{\mathbf{O}})$.

According to the definition of the Hadamard derivative, and by the uniform consistency of $\hat{F}_{\mathbf{O}}$, we conclude that $\big(U_n(t)\big)$ and $\Big(\widehat{U}_n^{IF}(t)\Big)$ have the same limiting distribution.

In order to determine the IF of the ATE, we proceed as suggested by Kennedy (2022), i.e. we treat the data as if they were discrete and compute Gateaux derivatives, exploiting standard differentiation rules.

The first step is to rewrite the (discretized) ATE as

$$
\sum_{\mathbf{l}} \Big(F_1\big(\cdot \mid Z_A = 1, \mathbf{Z}_L = \mathbf{l}\big) - F_1\big(\cdot \mid Z_A = 0, \mathbf{Z}_L = \mathbf{l}\big) \Big)\, P(\mathbf{Z}_L = \mathbf{l})
$$

and to apply the product rule, yielding

$$
\begin{aligned}
IF_{ATE}(\cdot;\, \mathbf{o}) &= \sum_{\tilde{\mathbf{l}}} \Big(\big(IF_{F_1}(\cdot, a' = 1, \mathbf{l}' = \tilde{\mathbf{l}};\, \mathbf{o}) - IF_{F_1}(\cdot, a' = 0, \mathbf{l}' = \tilde{\mathbf{l}};\, \mathbf{o}) \big) P(\mathbf{Z}_L = \tilde{\mathbf{l}}) \\
&\qquad + \big(F_1\big(\cdot \mid Z_A = 1, \mathbf{Z}_L = \tilde{\mathbf{l}}\big) - F_1\big(\cdot \mid Z_A = 0, \mathbf{Z}_L = \tilde{\mathbf{l}}\big) \big)\, IF_{P(\mathbf{Z}_L = \tilde{\mathbf{l}})}(\mathbf{l}) \Big) \\
&= \mathbb{E}_{\mathbf{Z}_L}\big(IF_{F_1}(\cdot, a' = 1, \mathbf{l}' = \mathbf{Z}_L;\, \mathbf{o}) - IF_{F_1}(\cdot, a' = 0, \mathbf{l}' = \mathbf{Z}_L;\, \mathbf{o}) \big) \\
&\quad + \sum_{\tilde{\mathbf{l}}} \Big(F_1\big(\cdot \mid Z_A = 1, \mathbf{Z}_L = \tilde{\mathbf{l}}\big) - F_1\big(\cdot \mid Z_A = 0, \mathbf{Z}_L = \tilde{\mathbf{l}}\big) \Big) \\
&\qquad \cdot \Big(\mathbb{1}\{\tilde{\mathbf{l}} = \mathbf{l}\} - P(\mathbf{Z}_L = \tilde{\mathbf{l}}) \Big) \\
&= \mathbb{E}_{\mathbf{Z}_L}\big(IF_{F_1}(\cdot, a' = 1, \mathbf{l}' = \mathbf{Z}_L;\, \mathbf{o}) - IF_{F_1}(\cdot, a' = 0, \mathbf{l}' = \mathbf{Z}_L;\, \mathbf{o}) \big) \\
&\quad + \big(F_1\big(\cdot \mid Z_A = 1, \mathbf{Z}_L = \mathbf{l}\big) - F_1\big(\cdot \mid Z_A = 0, \mathbf{Z}_L = \mathbf{l}\big) \big) \\
&\quad - \mathbb{E}_{\mathbf{Z}_L}\big(F_1\big(\cdot \mid Z_A = 1, \mathbf{Z}_L\big) - F_1\big(\cdot \mid Z_A = 0, \mathbf{Z}_L\big) \big),
\end{aligned}
$$

where $\mathbf{o} = (t, d, a, \mathbf{l}^T)^T$. Note that $IF_{P(\mathbf{Z}_L = \tilde{\mathbf{l}})}(\mathbf{l})$ is obtained by computing the Gateaux derivative

$$
\frac{\partial}{\partial \epsilon} \big((1 - \epsilon) P(\mathbf{Z}_L = \tilde{\mathbf{l}}) + \epsilon\, \mathbb{1}\{\tilde{\mathbf{l}} = \mathbf{l}\} \big) \Big|_{\epsilon = 0} = \; \mathbb{1}\{\tilde{\mathbf{l}} = \mathbf{l}\} - P(\mathbf{Z}_L = \tilde{\mathbf{l}})\,.
$$

Next, consider the function $F_1(\cdot \mid a, \mathbf{l}) = \int_0 \exp\Big(-\sum_{k=1}^K A_k(u \mid a, \mathbf{l})\Big)\, \alpha_1(u \mid a, \mathbf{l})\, \mathrm{d}s$. The product and chain rules suggest that

$$\begin{aligned} IF_{F_1}(\cdot, a', \mathbf{l}'; \mathbf{o}) = & \int_0^{\cdot} - \exp\Big(-\sum_{k=1}^K A_k(u \mid a', \mathbf{l}')\Big) \Big(\sum_{k=1}^K IF_{A_k}(u, a'=a', \mathbf{l}'=\mathbf{l}'; \mathbf{o})\Big) \\ & \qquad \cdot \mathrm{d}A_1(u \mid a', \mathbf{l}') \\ & + \int_0^{\cdot} \exp\Big(-\sum_{k=1}^K A_k(u \mid a', \mathbf{l}')\Big)\, \mathrm{d}IF_{A_1}(u, a'=a', \mathbf{l}'=\mathbf{l}'; \mathbf{o}). \end{aligned}$$

Similarly, for $A_k(\cdot \mid a, \mathbf{l}) = \exp\Big(\boldsymbol{\beta}_{0k}^T (a, \mathbf{l}^T)^T\Big) A_{0k}(\cdot)$, $k \in \{1, \ldots, K\}$, one finds

$$\begin{aligned} IF_{A_k}(\cdot, a', \mathbf{l}'; \mathbf{o}) = & \exp\Big(\boldsymbol{\beta}_{0k}^T (a', \mathbf{l}'^T)^T\Big) (a', \mathbf{l}'^T)\, IF_{\boldsymbol{\beta}_{0k}}(\mathbf{o})\, A_{0k}(\cdot) \\ & + \exp\Big(\boldsymbol{\beta}_{0k}^T (a', \mathbf{l}'^T)^T\Big)\, IF_{A_{0k}}(\cdot; \mathbf{o}). \end{aligned}$$

Let now $P(t, d, a, \mathbf{l})$ be short for $P(T \wedge C = t, D = d, Z_A = a, \mathbf{Z}_L = \mathbf{l})$. The fourth (discretized) function to be studied,

$$A_{0k}(\cdot) = \int_0 \frac{\sum_{t,d,a,\mathbf{l}} \mathbb{1}\{t \leq u, d = k\}\, P(t, d, a, \mathbf{l})}{\sum_{t,d,a,\mathbf{l}} \mathbb{1}\{t \geq u\} \exp\Big(\boldsymbol{\beta}_{0k}^T (a, \mathbf{l}^T)^T\Big) P(t, d, a, \mathbf{l})}\, \mathrm{d}u,$$

has the IF

$$\begin{aligned} IF_{A_{0k}}(\cdot; \mathbf{o}) = & \int_0^{\cdot} \frac{\sum_{\tilde{t},\tilde{d},\tilde{a},\tilde{\mathbf{l}}} \mathbb{1}\{\tilde{t} \leq u, \tilde{d} = k\}\, IF_{P(\tilde{t},\tilde{d},\tilde{a},\tilde{\mathbf{l}})}(\mathbf{o})}{\sum_{\tilde{t},\tilde{d},\tilde{a},\tilde{\mathbf{l}}} \mathbb{1}\{\tilde{t} \geq u\} \exp\Big(\boldsymbol{\beta}_{0k}^T (\tilde{a}, \tilde{\mathbf{l}}^T)^T\Big) P(\tilde{t}, \tilde{d}, \tilde{a}, \tilde{\mathbf{l}})}\, \mathrm{d}u \\ & - \int_0^{\cdot} \Bigg(\frac{\sum_{\tilde{t},\tilde{d},\tilde{a},\tilde{\mathbf{l}}} \mathbb{1}\{\tilde{t} \leq u, \tilde{d} = k\}\, P(\tilde{t}, \tilde{d}, \tilde{a}, \tilde{\mathbf{l}})}{\Big(\sum_{\tilde{t},\tilde{d},\tilde{a},\tilde{\mathbf{l}}} \mathbb{1}\{\tilde{t} \geq u\} \exp\Big(\boldsymbol{\beta}_{0k}^T (\tilde{a}, \tilde{\mathbf{l}}^T)^T\Big) P(\tilde{t}, \tilde{d}, \tilde{a}, \tilde{\mathbf{l}})\Big)^2} \\ & \qquad \cdot \sum_{\tilde{t},\tilde{d},\tilde{a},\tilde{\mathbf{l}}} \mathbb{1}\{\tilde{t} \geq u\} \exp\Big(\boldsymbol{\beta}_{0k}^T (\tilde{a}, \tilde{\mathbf{l}}^T)^T\Big) \Big(\tilde{a}, \tilde{\mathbf{l}}^T\Big)\, IF_{\boldsymbol{\beta}_{0k}}(\mathbf{o})\, P(\tilde{t}, \tilde{d}, \tilde{a}, \tilde{\mathbf{l}}) \Bigg) \mathrm{d}u \\ & - \int_0^{\cdot} \Bigg(\frac{\sum_{\tilde{t},\tilde{d},\tilde{a},\tilde{\mathbf{l}}} \mathbb{1}\{\tilde{t} \leq u, \tilde{d} = k\}\, P(\tilde{t}, \tilde{d}, \tilde{a}, \tilde{\mathbf{l}})}{\Big(\sum_{\tilde{t},\tilde{d},\tilde{a},\tilde{\mathbf{l}}} \mathbb{1}\{\tilde{t} \geq u\} \exp\Big(\boldsymbol{\beta}_{0k}^T (\tilde{a}, \tilde{\mathbf{l}}^T)^T\Big) P(\tilde{t}, \tilde{d}, \tilde{a}, \tilde{\mathbf{l}})\Big)^2} \\ & \qquad \cdot \sum_{\tilde{t},\tilde{d},\tilde{a},\tilde{\mathbf{l}}} \mathbb{1}\{\tilde{t} \geq u\} \exp\Big(\boldsymbol{\beta}_{0k}^T (\tilde{a}, \tilde{\mathbf{l}}^T)^T\Big)\, IF_{P(\tilde{t},\tilde{d},\tilde{a},\tilde{\mathbf{l}})}(\mathbf{o}) \Bigg) \mathrm{d}u \end{aligned}$$

on account of the quotient and product rules. Using that

$$IF_{P(\tilde{t},\tilde{d},\tilde{a},\tilde{\mathbf{l}})}(\mathbf{o}) = \mathbb{1}\{\tilde{t} = t, \tilde{d} = d, \tilde{a} = a, \tilde{\mathbf{l}} = \mathbf{l}\} - P(\tilde{t}, \tilde{d}, \tilde{a}, \tilde{\mathbf{l}})$$

(see the Gateaux derivative above) and

$$\frac{\sum_{t,d,a,\mathbf{l}} \mathbb{1}\{t \leq u, d=k\}\, P(t,d,a,\mathbf{l})}{\sum_{t,d,a,\mathbf{l}} \mathbb{1}\{t \geq u\} \exp\big(\boldsymbol{\beta}_{0k}^T(a,\mathbf{l}^T)^T\big) P(t,d,a,\mathbf{l})} = \alpha_{0k}(u),$$

we eventually obtain

$$\begin{aligned}
IF_{A_{0k}}(\cdot;\, \mathbf{o}) = &\int_0^{\cdot} \frac{\mathbb{1}\{t \leq u, d=k\}}{\sum_{\tilde{t},\tilde{d},\tilde{a},\tilde{\mathbf{l}}} \mathbb{1}\{\tilde{t} \geq u\} \exp\big(\boldsymbol{\beta}_{0k}^T(\tilde{a},\tilde{\mathbf{l}}^T)^T\big) P(\tilde{t},\tilde{d},\tilde{a},\tilde{\mathbf{l}})}\, \mathrm{d}u \\
&- A_{0k}(\cdot) \\
&- \int_0^{\cdot} \frac{\sum_{\tilde{t},\tilde{d},\tilde{a},\tilde{\mathbf{l}}} \mathbb{1}\{\tilde{t} \geq u\} \exp\big(\boldsymbol{\beta}_{0k}^T(\tilde{a},\tilde{\mathbf{l}}^T)^T\big) \big(\tilde{a},\tilde{\mathbf{l}}^T\big) P(\tilde{t},\tilde{d},\tilde{a},\tilde{\mathbf{l}})}{\sum_{\tilde{t},\tilde{d},\tilde{a},\tilde{\mathbf{l}}} \mathbb{1}\{\tilde{t} \geq u\} \exp\big(\boldsymbol{\beta}_{0k}^T(\tilde{a},\tilde{\mathbf{l}}^T)^T\big) P(\tilde{t},\tilde{d},\tilde{a},\tilde{\mathbf{l}})}\, \mathrm{d}A_{0k}(u)\, IF_{\boldsymbol{\beta}_{0k}}(\mathbf{o}) \\
&- \int_0^{\cdot} \frac{\mathbb{1}\{t \geq u\} \exp\big(\boldsymbol{\beta}_{0k}^T(a,\mathbf{l}^T)^T\big)}{\sum_{\tilde{t},\tilde{d},\tilde{a},\tilde{\mathbf{l}}} \mathbb{1}\{\tilde{t} \geq u\} \exp\big(\boldsymbol{\beta}_{0k}^T(\tilde{a},\tilde{\mathbf{l}}^T)^T\big) P(\tilde{t},\tilde{d},\tilde{a},\tilde{\mathbf{l}})}\, \mathrm{d}A_{0k}(u) \\
&+ A_{0k}(\cdot) \\
= &\frac{\mathbb{1}\{t \leq \cdot, d=k\}}{s^{(0)}(\boldsymbol{\beta}_{0k}, t)} - \int_0^{\cdot} \big(\mathbf{e}(\boldsymbol{\beta}_{0k}, u)\big)^T \mathrm{d}A_{0k}(u)\, IF_{\boldsymbol{\beta}_{0k}}(\mathbf{o}) \\
&- \exp\big(\boldsymbol{\beta}_{0k}^T(a,\mathbf{l}^T)^T\big) \int_0^{\cdot \wedge t} \frac{\mathrm{d}A_{0k}(u)}{s^{(0)}(\boldsymbol{\beta}_{0k}, u)}.
\end{aligned}$$

The IF

$$\begin{aligned}
IF_{\boldsymbol{\beta}_{0k}}(\mathbf{o}) = \boldsymbol{\Sigma}_k^{-1} \Bigg(&\mathbb{1}\{d=k\}\big((a,\mathbf{l}^T)^T - \mathbf{e}(\boldsymbol{\beta}_{0k}, t)\big) \\
&- \exp\big(\boldsymbol{\beta}_{0k}^T(a,\mathbf{l}^T)^T\big) \int \mathbb{1}\{\tilde{t} \leq t, \tilde{d}=k\} \frac{\big((a,\mathbf{l}^T)^T - \mathbf{e}(\boldsymbol{\beta}_{0k}, \tilde{t})\big)}{s^{(0)}(\boldsymbol{\beta}_{0k}, \tilde{t})}\, \mathrm{d}F_{\mathbf{O}}(\tilde{t},\tilde{d},\tilde{a},\tilde{\mathbf{l}}) \Bigg)
\end{aligned}$$

can lastly be derived based on the score function and the information matrix of the Cox partial likelihood (Gerds and Schumacher, 2001).

Empirical counterparts $\widehat{IF}_{ATE}$, $\widehat{IF}_{F_1}$, $\widehat{IF}_{A_k}$, $\widehat{IF}_{A_{0k}}$, $\widehat{IF}_{\boldsymbol{\beta}_{0k}}$ of the IFs derived above arise simply by the plug-in principle. See also Ozenne, Sørensen, et al. (2017) for comparison. □

Wild bootstrap (WBS)

Let us now turn to the third resampling approach, namely the WBS with its variants.

One may notice the parallels between the subsequent theorem and Theorem 1 proposed by Dobler, Beyersmann, and Pauly (2017).

Theorem 4.4 (Approximation of the distribution of $\big(U_n(t)\big)$ by the WBS)**:**
Consider the random multipliers G_i, $i \in \{1,\dots,n\}$, that fulfil the following conditions:

(i) $\sqrt{n}\,\max_{i\in\{1,\dots,n\}} \mathbb{E}\big(G_i \mid \mathscr{F}_\tau\big) \xrightarrow{P} 0$,
(ii) $\max_{i\in\{1,\dots,n\}} \mathrm{Var}\big(G_i \mid \mathscr{F}_\tau\big) \xrightarrow{P} 1$,
(iii) $\frac{1}{\sqrt{n}}\,\max_{i\in\{1,\dots,n\}} \mathbb{E}\big(G_i^4 \mid \mathscr{F}_\tau\big) \xrightarrow{P} 0$,
(iv) $P\big(G_1 \le g_1,\dots,G_n \le g_n \mid \mathscr{F}_\tau\big) = \prod_{i=1}^n P\big(G_i \le g_i \mid \mathscr{F}_\tau\big)$,
(v) $\sum_{i=1}^n \mathbb{E}\left(\frac{\big(G_i-\mathbb{E}(G_i|\mathscr{F}_\tau)\big)^2}{\sum_{j=1}^n \big(\mathrm{Var}(G_j|\mathscr{F}_\tau)\big)} \cdot \mathbb{1}\left\{\frac{\big(G_i-\mathbb{E}(G_i|\mathscr{F}_\tau)\big)^2}{\sum_{j=1}^n \big(\mathrm{Var}(G_j|\mathscr{F}_\tau)\big)} > \epsilon\right\} \,\middle|\, \mathscr{F}_\tau\right) \xrightarrow{P} 0 \;\forall \epsilon > 0$.

It holds that the plug-in estimator

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

converges weakly to the same process as $\big(U_n(t)\big)$ on $D[0,\tau]$ conditional on the data $\big((T_i \wedge C_i, D_i, \mathbf{Z}_i)\big)_{i\in\{1,\dots,n\}}$ in probability.

Remark 4.1 (see also Example 1 by Dobler, Beyersmann, and Pauly, 2017)*:*
The following choices of multipliers fulfil the conditions in Theorem 4.4:

- independent standard normal multipliers $G_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, $i \in \{1,\dots,n\}$ (cf. Lin, Wei, and Ying, 1993),
- independent, centred unit Poisson multipliers $G_i \overset{\text{i.i.d.}}{\sim} \mathcal{P}oi(1) - 1$, $i \in \{1,\dots,n\}$ (cf. Beyersmann, Di Termini, and Pauly, 2013),
- conditionally independent, centred binomial multipliers that correspond to the weird bootstrap described by Andersen, Borgan, et al. (1993, Subsection IV.1.4), i.e. $G_i \sim \mathcal{B}in\big(Y(T_i \wedge C_i),\, 1/Y(T_i \wedge C_i)\big) - 1$, with $G_{i_1} \perp\!\!\!\perp G_{i_2} \mid \mathscr{F}_\tau$ for $i_1 \neq i_2$.

In order to demonstrate the convergence claimed in Theorem 4.4, we rely on several lemmata, which are proven in Section A.2 of Appendix A based on ideas of Beyersmann, Di Termini, and Pauly (2013), Dobler and Pauly (2014), as well as Dobler, Beyersmann, and Pauly (2017).

Let $i \in \{1,\dots,n\}$, $k \in \{1,\dots,K\}$, and $0 \le t_1 \le \dots \le t_l \le \tau$ for $l \in \mathbb{N}$. The triangular arrays $\mathbf{X}_{n,i}^{(k)} = \big(X_{n,i}^{(k)}(t_1),\dots,X_{n,i}^{(k)}(t_l)\big)^T$ serve as a starting point for our considerations, with

$$\begin{aligned} X_{n,i}^{(k)}(t) = &\int_0^t \widehat{\widetilde{H}}_{k1}(u,t)\, \frac{\mathrm{d}N_{ki}(u)}{\sqrt{n}\, S^{(0)}\big(\hat{\boldsymbol{\beta}}_k, u\big)} \\ &+ \int_0^\tau \frac{1}{\sqrt{n}} \Big(\widehat{\widetilde{\mathbf{H}}}_{k2}(t)\Big)^T \widehat{\boldsymbol{\Sigma}}_k^{-1} \Big(\mathbf{Z}_i - \mathbf{E}\big(\hat{\boldsymbol{\beta}}_k, u\big)\Big)\, \mathrm{d}N_{ki}(u). \end{aligned}$$

It should be noted that $\widehat{U}_n^{WBS}(t)$ can be expressed as $\sum_{k=1}^{K}\sum_{i=1}^{n} X_{n,i}^{(k)}(t)\,G_i$ according to the definition above.

Lemma 4.2:
For each $k \in \{1,\ldots,K\}$, the triangular arrays $\mathbf{X}_{n,i}^{(k)}$ satisfy the following conditions:

(i) $\max_{i\in\{1,\ldots,n\}} \left\|\mathbf{X}_{n,i}^{(k)}\right\| \xrightarrow{P} 0$ *(where $\|\cdot\|$ denotes the Euclidean norm),*
(ii) $\sum_{i=1}^{n} \mathbf{X}_{n,i}^{(k)} \left(\mathbf{X}_{n,i}^{(k)}\right)^T \xrightarrow{P} \left(\xi^{(k)}(t_r,t_s)\right)_{r,s\,\in\,\{1,\ldots,l\}}$.

Lemma 4.3:
For $0 \le t_r \le t_s \le \tau$ and $k \in \{1,\ldots,K\}$, it holds that

$$\max_{i\in\{1,\ldots,n\}} \left| X_{n,i}^{(k)}(t_s) - X_{n,i}^{(k)}(t_r) \right| \in O_P\left(n^{-1/2}\right),$$

where $O_P(a_n)$ indicates asymptotic boundedness in probability by the sequence a_n.

The bound in Lemma 4.3 is in fact independent of the considered time points!

Lemma 4.4:
Consider $0 \le t_q \le t_r \le t_s \le \tau$, $k \in \{1,\ldots,K\}$, and the function $L_n^{(k)}(t)$ defined in Section A.2 of Appendix A. Provided that the conditions in Theorem 4.4 are fulfilled, the expectation

$$\mathbb{E}\left(\left(\sum_{i=1}^{n} X_{n,i}^{(k)}(t_r)\,G_i - \sum_{i=1}^{n} X_{n,i}^{(k)}\left(t_q\right) G_i\right)^2 \left(\sum_{i=1}^{n} X_{n,i}^{(k)}(t_s)\,G_i - \sum_{i=1}^{n} X_{n,i}^{(k)}(t_r)\,G_i\right)^2 \,\middle|\, \mathscr{F}_\tau\right)$$

has the upper bound $\left(L_n^{(k)}(t_s) - L_n^{(k)}\left(t_q\right)\right)^{3/2} \cdot O_P(1)$.

Proof of Theorem 4.4:

Lemma 4.2 and the conditions w.r.t. the multipliers G_i in Theorem 4.4 ensure that the assumptions stated in Lemma 1 of Dobler, Beyersmann, and Pauly (2017, Supplementary Material) are fulfilled for each $k \in \{1,\ldots,K\}$. Hence, the finite-dimensional distributions of $\widehat{U}_n^{WBS;\,(k)}(t) = \sum_{i=1}^{n}\left(\widehat{H}_{k1i}(T_i \wedge C_i, t)\,N_{ki}(t)\,G_i + \widehat{H}_{k2i}(T_i \wedge C_i, t)\,N_{ki}(\tau)\,G_i\right)$ converge weakly to Gaussian processes with mean zero and covariance functions $\xi^{(k)}$ conditional on the history $\mathscr{F}_\tau$ in probability.

One may further apply Theorem 2.1 to demonstrate that

$$\frac{1}{n}\sum_{i=1}^{n} \int_0^\tau \left(\mathbf{Z}_i - \mathbf{E}\left(\hat{\boldsymbol{\beta}}_k, u\right)\right)^{\otimes 2} \mathrm{d}N_{ki}(u) \xrightarrow{P} \boldsymbol{\Sigma}_k,$$

which suggests – again by Theorem 2.1 – that the functions $L_n^{(k)}$ introduced in Lemma 4.4 converge uniformly to the non-decreasing and continuous mappings

$$
\begin{aligned}
l^{(k)}(t) = {} & \left(\exp(2\beta_{0kA}) + 2\exp(\beta_{0kA}) + 1\right) \mathbb{E}_{\mathbf{Z}_L}\Big(\exp\Big(2\boldsymbol{\beta}_{0kL}^T \mathbf{Z}_L\Big)\Big) \int_0^t \frac{\mathrm{d}A_{0k}(u)}{s^{(0)}(\boldsymbol{\beta}_{0k}, u)} \\
& + \mathbb{E}_{\mathbf{Z}_L}\Big(\Big(\big(F_1\big(t \mid Z_A = 1, \mathbf{Z}_L\big)\big)^2 \exp(2\beta_{0kA}) + \big(F_1\big(t \mid Z_A = 0, \mathbf{Z}_L\big)\big)^2\Big) \\
& \qquad \cdot \exp\Big(2\boldsymbol{\beta}_{0kL}^T \mathbf{Z}_L\Big)\Big) \int_0^\tau \frac{\mathrm{d}A_{0k}(u)}{s^{(0)}(\boldsymbol{\beta}_{0k}, u)} \\
& + \int_0^t \mathbb{E}_{\mathbf{Z}_L}\Big(\Big(\big(1, \mathbf{Z}_L^T\big)^T - \mathbf{e}\left(\boldsymbol{\beta}_{0k}, u\right)\Big)^T \boldsymbol{\Sigma}_k^{-1} \Big(\big(1, \mathbf{Z}_L^T\big)^T - \mathbf{e}\Big(\boldsymbol{\beta}_{0k}, u\Big)\Big) \\
& \qquad \cdot \exp\Big(2\boldsymbol{\beta}_{0k}\big(1, \mathbf{Z}_L^T\big)^T\Big)\Big) \frac{\mathrm{d}A_{0k}(u)}{s^{(0)}(\boldsymbol{\beta}_{0k}, u)} \\
& + \int_0^t \mathbb{E}_{\mathbf{Z}_L}\Big(\Big(\big(0, \mathbf{Z}_L^T\big)^T - \mathbf{e}\left(\boldsymbol{\beta}_{0k}, u\right)\Big)^T \boldsymbol{\Sigma}_k^{-1} \Big(\big(0, \mathbf{Z}_L^T\big)^T - \mathbf{e}\Big(\boldsymbol{\beta}_{0k}, u\Big)\Big) \\
& \qquad \cdot \exp\Big(2\boldsymbol{\beta}_{0k}\big(0, \mathbf{Z}_L^T\big)^T\Big)\Big) \frac{\mathrm{d}A_{0k}(u)}{s^{(0)}(\boldsymbol{\beta}_{0k}, u)} \\
& + \mathbb{E}_{\mathbf{Z}_L}\Big(\big(F_1\big(t \mid Z_A = 1, \mathbf{Z}_L\big)\big)^2 \\
& \qquad \cdot \int_0^\tau \Big(\big(1, \mathbf{Z}_L^T\big)^T - \mathbf{e}\left(\boldsymbol{\beta}_{0k}, u\right)\Big)^T \boldsymbol{\Sigma}_k^{-1} \Big(\big(1, \mathbf{Z}_L^T\big)^T - \mathbf{e}\Big(\boldsymbol{\beta}_{0k}, u\Big)\Big) \\
& \qquad\qquad \cdot \exp\Big(2\boldsymbol{\beta}_{0k}\big(1, \mathbf{Z}_L^T\big)^T\Big)\Big) \frac{\mathrm{d}A_{0k}(u)}{s^{(0)}(\boldsymbol{\beta}_{0k}, u)} \\
& + \mathbb{E}_{\mathbf{Z}_L}\Big(\big(F_1\big(t \mid Z_A = 0, \mathbf{Z}_L\big)\big)^2 \\
& \qquad \cdot \int_0^\tau \Big(\big(0, \mathbf{Z}_L^T\big)^T - \mathbf{e}\left(\boldsymbol{\beta}_{0k}, u\right)\Big)^T \boldsymbol{\Sigma}_k^{-1} \Big(\big(0, \mathbf{Z}_L^T\big)^T - \mathbf{e}\Big(\boldsymbol{\beta}_{0k}, u\Big)\Big) \\
& \qquad\qquad \cdot \exp\Big(2\boldsymbol{\beta}_{0k}\big(0, \mathbf{Z}_L^T\big)^T\Big)\Big) \frac{\mathrm{d}A_{0k}(u)}{s^{(0)}(\boldsymbol{\beta}_{0k}, u)}
\end{aligned}
$$

on $[0, \tau]$ (cf. the proof of Lemma 4.2). As a result, we can show the conditional tightness of the processes $\big(\widehat{U}_n^{WBS;\,(k)}(t)\big)$ (cf. Dobler and Pauly, 2014, proof of Theorem 3.1): The subsequence principle for convergence in probability implies that for every subsequence of $\mathbb{N}$, there is another subsequence n such that for almost every fixed ω in the sample space, one finds values $n_0 \in \mathbb{N}$, $\gamma > 0$, and a sequence of non-decreasing, continuous functions $l_n^{(k)}$ (converging uniformly to $l^{(k)}$) so that

$$
\begin{aligned}
\mathbb{E}\left(\left(\sum_{i=1}^n X_{n,i}^{(k)}(t_r) G_i - \sum_{i=1}^n X_{n,i}^{(k)}\big(t_q\big) G_i\right)^2 \left(\sum_{i=1}^n X_{n,i}^{(k)}(t_s) G_i - \sum_{i=1}^n X_{n,i}^{(k)}(t_r) G_i\right)^2 \,\middle|\, \mathscr{F}_\tau\right) \\
\leq \gamma \Big(l_n^{(k)}(t_s) - l_n^{(k)}\big(t_q\big)\Big)^{3/2}
\end{aligned}
$$

if $n \geq n_0$ (cf. Beyersmann, Di Termini, and Pauly, 2013). Note that the choices of n_0 and γ are independent of the time points t_q, t_r, t_s. As indicated by Dobler and Pauly (2014), the conditional tightness follows by extending Theorem 13.5 in Billingsley (1999) pointwise along subsequences. This establishes the conditional convergence in distribution of $\big(\widehat{U}_n^{WBS;\,(k)}(t)\big)$ in probability for each $k \in \{1, \ldots, K\}$.

Eventually, considering that the processes $\big(\widehat{U}_n^{WBS;\,(k)}(t)\big)$ and $\big(\widehat{U}_n^{WBS;\,(k')}(t)\big)$ are independent conditional on $\mathscr{F}_\tau$ for $k \neq k'$ (because $\mathrm{d}N_{ki}(t)\,\mathrm{d}N_{k'i}(t) = 0$), we conclude that $\big(\widehat{U}_n^{WBS}(t)\big)$ converges weakly to a zero-mean Gaussian process with covariance function ξ on $[0, \tau]$ given the data in probability. □

4.1.2. Simulation study comparing the resampling approaches

The performance of the discussed resampling approaches was compared by means of intensive empirical investigations along the lines of the simulation study presented by Ozenne, Scheike, et al. (2020).

We generated competing risks data involving $K = 2$ event types as follows: Based on $p = 12$ independent covariates $Z_{L_1}, \ldots, Z_{L_{12}}$, with $Z_{L_1}, \ldots, Z_{L_6} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$ and $Z_{L_7}, \ldots, Z_{L_{12}} \overset{\text{i.i.d.}}{\sim} \mathcal{B}in(1, 0.5)$, the treatment variable Z_A, the event and censoring times T and C, as well as the event indicator D were modelled as

$$
\begin{aligned}
Z_A &\sim \mathcal{B}in\,(1, p_A)\,, \\
T &\sim \mathcal{W}b\Big(2, \sqrt{\tfrac{2t}{\alpha_1(t) + \alpha_2(t)}}\Big), \\
C &\sim \mathcal{W}b\Big(2, \sqrt{\tfrac{2t}{\alpha_C(t)}}\Big), \\
D &= \mathbb{1}\{T \leq C\}\, D_T,
\end{aligned}
$$

with

$$
\begin{aligned}
p_A &= \Big(1 + \exp\Big(-\big(\gamma_0 + \log(2)\big(Z_{L_1} - Z_{L_2} + Z_{L_6} + Z_{L_7} - Z_{L_8} - Z_{L_{12}}\big)\big)\Big)\Big)^{-1}, \\
\alpha_1(t) &= 0.02 \exp\Big(\beta_{01A} + \log(2)\big(Z_{L_1} + Z_{L_3} + Z_{L_6} + Z_{L_7} + Z_{L_9} + Z_{L_{12}}\big)\Big)\, t, \\
\alpha_2(t) &= 0.02 \exp\Big(\log(2)\big(-Z_{L_1} + Z_{L_5} + Z_{L_6} - Z_{L_7} + Z_{L_{11}} + Z_{L_{12}}\big)\Big)\, t, \\
\alpha_C(t) &= \frac{2}{\delta} \exp\Big(\log(2)\big(-Z_{L_1} + Z_{L_4} - Z_{L_6} - Z_{L_7} + Z_{L_{10}} - Z_{L_{12}}\big)\Big)\, t, \\
D_T - 1 &\sim \mathcal{B}in\Big(1, \alpha_2(T)\,/\big(\alpha_1(T) + \alpha_2(T)\big)\Big).
\end{aligned}
$$

Accordingly, the data conform to a multistate model with cause-specific Weibull hazards α_1 and α_2 (cf. Beyersmann, Latouche, et al., 2009). The DAG in Figure 4.1 depicts the relations between the individual variables; see also Table 4.1 for an overview of the covariate effects.

Figure 4.1: Causal relations between the covariates, treatment and the event times.

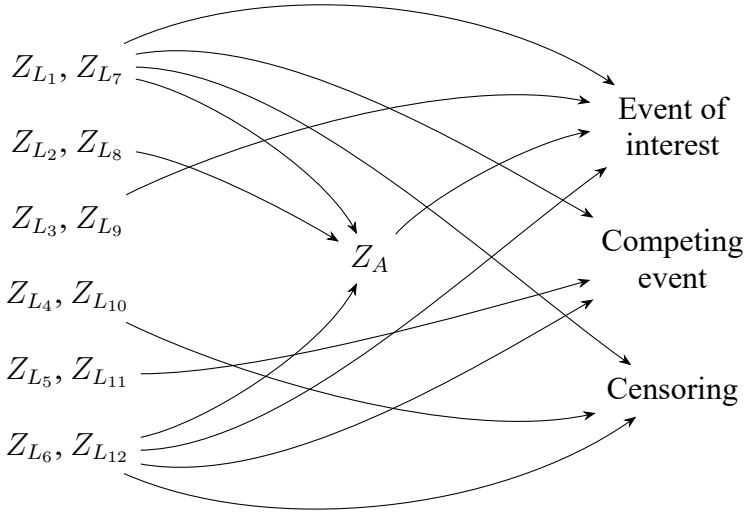


Table 4.1: Effects of the covariates on the treatment probability, the event and the censoring times.

Covariate	Odds Ratio w.r.t. treatment probability[a]	Hazard ratio w.r.t. event of interest[b]	Hazard ratio w.r.t. competing event[c]	Hazard ratio w.r.t. censoring[d]
Z_A	–	$\exp(\beta_{01A})$, $\beta_{01A} \in \{-2, 0, 2\}$	$\exp(\beta_{02A}) = 1.0$	$\exp(\beta_{0CA}) = 1.0$
Z_{L_1} Z_{L_7}	$\frac{\exp(\gamma_{L_1})}{\exp(\gamma_{L_7})} = 2.0$	$\frac{\exp(\beta_{01L_1})}{\exp(\beta_{01L_7})} = 2.0$	$\frac{\exp(\beta_{02L_1})}{\exp(\beta_{02L_7})} = 0.5$	$\frac{\exp(\beta_{0CL_1})}{\exp(\beta_{0CL_7})} = 0.5$
Z_{L_2} Z_{L_8}	$\frac{\exp(\gamma_{L_2})}{\exp(\gamma_{L_8})} = 0.5$	$\frac{\exp(\beta_{01L_2})}{\exp(\beta_{01L_8})} = 1.0$	$\frac{\exp(\beta_{02L_2})}{\exp(\beta_{02L_8})} = 1.0$	$\frac{\exp(\beta_{0CL_2})}{\exp(\beta_{0CL_8})} = 1.0$
Z_{L_3} Z_{L_9}	$\frac{\exp(\gamma_{L_3})}{\exp(\gamma_{L_9})} = 1.0$	$\frac{\exp(\beta_{01L_3})}{\exp(\beta_{01L_9})} = 2.0$	$\frac{\exp(\beta_{02L_3})}{\exp(\beta_{02L_9})} = 1.0$	$\frac{\exp(\beta_{0CL_3})}{\exp(\beta_{0CL_9})} = 1.0$
Z_{L_4} $Z_{L_{10}}$	$\frac{\exp(\gamma_{L_4})}{\exp(\gamma_{L_{10}})} = 1.0$	$\frac{\exp(\beta_{01L_4})}{\exp(\beta_{01L_{10}})} = 1.0$	$\frac{\exp(\beta_{02L_4})}{\exp(\beta_{02L_{10}})} = 1.0$	$\frac{\exp(\beta_{0CL_4})}{\exp(\beta_{0CL_{10}})} = 2.0$
Z_{L_5} $Z_{L_{11}}$	$\frac{\exp(\gamma_{L_5})}{\exp(\gamma_{L_{11}})} = 1.0$	$\frac{\exp(\beta_{01L_5})}{\exp(\beta_{01L_{11}})} = 1.0$	$\frac{\exp(\beta_{02L_5})}{\exp(\beta_{02L_{11}})} = 2.0$	$\frac{\exp(\beta_{0CL_5})}{\exp(\beta_{0CL_{11}})} = 1.0$
Z_{L_6} $Z_{L_{12}}$	$\frac{\exp(\gamma_{L_6})}{\exp(\gamma_{L_{12}})} = 2.0$	$\frac{\exp(\beta_{01L_6})}{\exp(\beta_{01L_{12}})} = 2.0$	$\frac{\exp(\beta_{02L_6})}{\exp(\beta_{02L_{12}})} = 2.0$	$\frac{\exp(\beta_{0CL_6})}{\exp(\beta_{0CL_{12}})} = 0.5$

[a] $P(Z_A = 1) = \big(1 + \exp\big(\gamma_0 + \boldsymbol{\gamma}_L^T \mathbf{Z}_L\big)\big)^{-1}$, with $\boldsymbol{\gamma}_L = (\gamma_{L_1}, \ldots, \gamma_{L_{12}})^T$.

[b] $\alpha_1(t) = 0.02 \exp\big(\beta_{01A} Z_A + \boldsymbol{\beta}_{01L}^T \mathbf{Z}_L\big)\big) t$, with $\boldsymbol{\beta}_{01L} = (\beta_{01L_1}, \ldots, \beta_{01L_{12}})^T$.

[c] $\alpha_2(t) = 0.02 \exp\big(\beta_{02A} Z_A + \boldsymbol{\beta}_{02L}^T \mathbf{Z}_L\big)\big) t$, with $\boldsymbol{\beta}_{02L} = (\beta_{02L_1}, \ldots, \beta_{02L_{12}})^T$.

[d] $\alpha_C(t) = \frac{2}{\delta} \exp\big(\beta_{0CA} Z_A + \boldsymbol{\beta}_{0CL}^T \mathbf{Z}_L\big)\big) t$, with $\boldsymbol{\beta}_{0CL} = (\beta_{01L_1}, \ldots, \beta_{0CL_{12}})^T$.

Other than that, we used the parameters σ^2, γ_0, and δ to control the dispersion of the normally distributed covariates, the overall treatment probability, and the intensity of censoring. In order to investigate the performance of the resampling approaches when censoring is non-random, we further simulated a two-state survival setting with event-driven censoring and staggered study entry (see Chapter 3). Each of the considered scenarios was implemented with sample sizes $n \in \{50, 75, 100, 200, 300\}$ as well as treatment effects according to the coefficient $\beta_{01A} \in \{-2, 0, 2\}$. Table 4.2 summarizes the characteristics of the data in the distinct settings.

Table 4.2: Simulation scenarios considered.

Scenario	% censored[a]			% type 1 events[a]			% treated	$\text{Var}(Z_{L_1})$
	$\beta_{01A} =$			$\beta_{01A} =$				
	-2	0	2	-2	0	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

[a] Determined at $t = 9$ except for the scenario with type II censoring, where the percentages are determined at $t = 10,\ 5,\ 2.5$ for $\beta_{01A} = -2,\ 0,\ 2$, respectively.

After the data had been generated, we derived 95% CIs and CBs for the ATE by means of the resampling approaches explored in Subsection 4.1.1. Our interest w.r.t. the pointwise CIs and the time-simultaneous CBs was in the time points $t \in \{1, 3, 5, 7, 9\}$ and the interval $[0, 9]$, respectively. (In case of the type II censored scenario, we examined $t \in \{2, 4, 6, 8, 10\}$, $\{1, 2, 3, 4, 5\}$, $\{0.5, 1, 1.5, 2, 2.5\}$ and the intervals $[2, 10]$, $[1, 5]$, $[0.5, 2.5]$ for the respective values $-2,\ 0,\ 2$ of β_{01A}.) Each resampling technique was realized using $B = 1,000$ replications.

The CI limits regarding the EBS were simply specified as the empirical 0.025 and 0.975 quantiles of the bootstrap estimates $\left(\widehat{ATE}^*_{ds;\,b}(t)\right)_{b\in\{1,\ldots,B\}}$. The CB w.r.t. the EBS, on the other hand, emerged as

$$\left[\widehat{ATE}_{ds}(t) - q_{EBS}(0.95)\sqrt{\widehat{\text{Var}}^{EBS}(t)},\ \widehat{ATE}_{ds}(t) + q_{EBS}(0.95)\sqrt{\widehat{\text{Var}}^{EBS}(t)}\right],$$

where $q_{EBS}(0.95)$ denotes the 0.95 quantile of

$$\left(\sup_{t\in[0,9]}\left|\frac{\widehat{ATE}^*_{ds;\,b}(t) - \frac{1}{B}\sum_{\tilde{b}=1}^{B}\widehat{ATE}^*_{ds;\,\tilde{b}}(t)}{\sqrt{\widehat{\text{Var}}^{EBS}(t)}}\right|\right)_{b\in\{1,\ldots B\}}$$

and $\widehat{\text{Var}}^{EBS}(t)$ refers to the sample variance of $\left(\widehat{ATE}^{*}_{ds;\,b}(t)\right)_{b\in\{1,\ldots,B\}}$ (cf. Theorem 4.2). Note that we used the 0.95 quantile and the absolute value for q_{EBS} in order to enhance stability.

Concerning the IF approach, the initial step was to determine the empirical estimator $\widehat{\text{Var}}^{IF}(t) = \frac{1}{n^2}\sum_{i=1}^{n}\left(\widehat{IF}_{ATE}(t;\, T_i \wedge C_i, D_i, \mathbf{Z}_i)\right)^2$ and the 0.95 quantile $q_{IF}(0.95)$ of

$$\left(\sup_{t\in[0,9]}\left|\frac{1}{n}\sum_{i=1}^{n}\frac{\widehat{IF}_{ATE}(t;\, T_i\wedge C_i, D_i, \mathbf{Z}_i)}{\sqrt{\widehat{\text{Var}}^{IF}(t)}}\, G^{IF}_{i_b}\right|\right)_{b\in\{1,\ldots B\}},$$

with i.i.d. standard normally distributed multipliers $G^{IF}_{i_b}$, $i \in \{1,\ldots,n\}$, $b \in \{1,\ldots,B\}$. Theorem 4.3 consequently yields the CI

$$\left[\widehat{ATE}_{ds}(t) - q_{\mathcal{N}(0,1)}(0.975)\sqrt{\widehat{\text{Var}}^{IF}(t)},\ \widehat{ATE}_{ds}(t) + q_{\mathcal{N}(0,1)}(0.975)\sqrt{\widehat{\text{Var}}^{IF}(t)}\right]$$

and the CB

$$\left[\widehat{ATE}_{ds}(t) - q_{IF}(0.95)\sqrt{\widehat{\text{Var}}^{IF}(t)},\ \widehat{ATE}_{ds}(t) + q_{IF}(0.95)\sqrt{\widehat{\text{Var}}^{IF}(t)}\right].$$

We lastly obtained

$$\left[\widehat{ATE}_{ds}(t) - q_{WBS}(t;\, 0.95),\ \widehat{ATE}_{ds}(t) + q_{WBS}(t;\, 0.95)\right]$$

and, similarly,

$$\left[\widehat{ATE}_{ds}(t) - q_{WBS}(0.95)\sqrt{\widehat{\text{Var}}^{WBS}(t)},\ \widehat{ATE}_{ds}(t) + q_{WBS}(0.95)\sqrt{\widehat{\text{Var}}^{WBS}(t)}\right]$$

as CI and CB for the WBS according to Theorem 4.4. Here, $q_{WBS}(t;\, 0.95)$ denotes the 0.95 quantile of $\left(\left|\widehat{U}^{WBS;\,(b)}_n(t)/\sqrt{n}\right|\right)_{b\in\{1,\ldots,B\}}$, whereas $q_{WBS}(0.95)$ is the 0.95 quantile of

$$\left(\sup_{t\in[0,9]}\left|\frac{\widehat{U}^{WBS;\,(b)}_n(t)}{\sqrt{n\,\widehat{\text{Var}}^{WBS}(t)}}\right|\right)_{b\in\{1,\ldots B\}},$$

with plug-in estimators

$$\widehat{U}^{WBS;\,(b)}_n(t) = \sum_{k=1}^{K}\sum_{i=1}^{n}\left(\widehat{H}_{k1i}(T_i\wedge C_i, t)\, N_{ki}(t)\, G^{WBS}_{i_b} + \widehat{H}_{k2i}(T_i\wedge C_i, t)\, N_{ki}(\tau)\, G^{WBS}_{i_b}\right),$$

$b \in \{1,\ldots,B\}$, and sample variance $\widehat{\text{Var}}^{WBS}(t)$ of $\left(\widehat{U}^{WBS;\,(b)}_n(t)\right)_{b\in\{1,\ldots,B\}}$. We examined different choices of multipliers $G^{WBS}_{i_b}$ pursuant to Remark 4.1: While standard

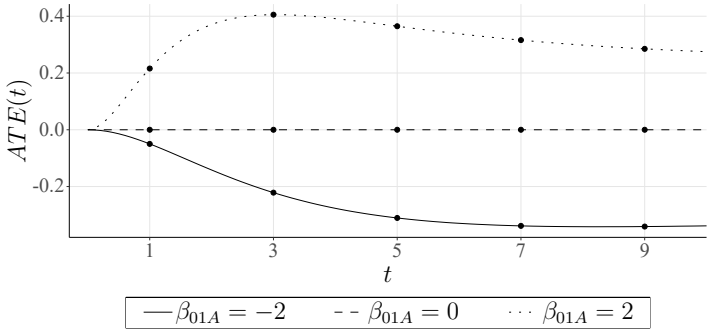
normally distributed random variables conform to the original wild bootstrap, centred Poisson variables and binomial multipliers as specified in the remark are supposed to improve small-sample performance (Beyersmann, Di Termini, and Pauly, 2013; Dobler, Beyersmann, and Pauly, 2017). The subversions of the WBS according to these multipliers will hereafter be referred to as the Lin, Beyersmann, and weird bootstrap approaches, respectively.

The capability of the individual resampling techniques to imitate the distribution of the process $\big(U_n(t)\big)$ was finally compared considering the coverage probabilities of the corresponding confidence regions. Each simulation scenario was implemented 5,000 times so that the MCSE w.r.t. the coverage was restricted below 0.75%.

Since we investigated rather small samples, some Monte Carlo iterations entailed too few observed events for the cause-specific Cox models to converge, and as a consequence, we could not determine any CIs and CBs. The respective coverage probabilities were therefore based on less than 5,000 iterations. For similar reasons, parts of the confidence regions relating to the EBS were obtained using less than 1,000 bootstrap samples. Table B.8 in Appendix B quantifies the frequency of these issues. It becomes apparent that the simulation outcomes for the settings with sample size $n = 50$ and parameter value $\beta_{01A} = 2$ should be treated carefully.

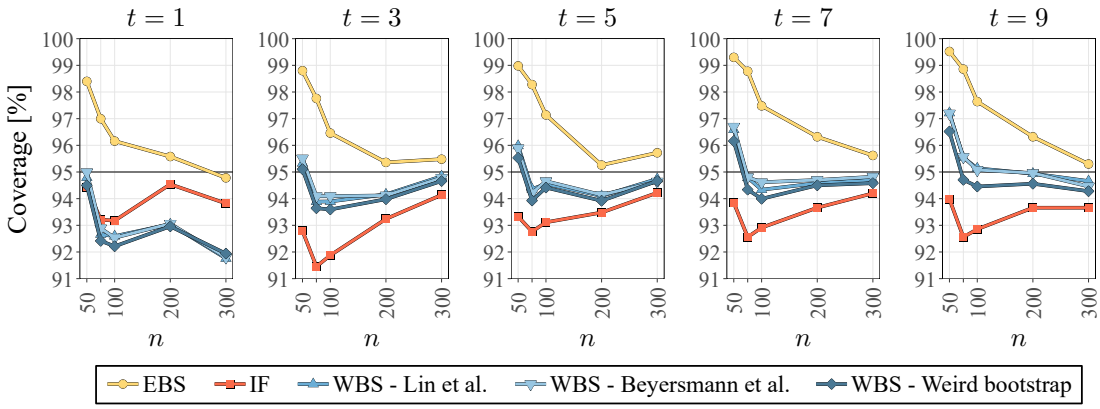
In order to assess the validity of the confidence regions, we had to determine the true ATE, but its analytical calculation is difficult when multiple covariates are involved. We hence approximated its value numerically. For this purpose, 1,000 data sets were generated as described earlier, with sample size $n = 100,000$. The treatment groups were however allocated at random (by setting the parameter p_A to 0.5) and censoring was suppressed. We then estimated the difference $F_1\big(t \mid Z_A{=}1\big) - F_1\big(t \mid Z_A{=}0\big)$ in each of the 1,000 data sets and determined the median. Figure 4.2 shows the resulting approximation of the ATE over the time interval $[0, 9]$.

Figure 4.2: Approximated ATE (except for the scenarios with $\sigma^2 \neq 1$ and those with type II censoring).



Looking at the generated CIs, the coverage probabilities obtained by the distinct resampling approaches generally complied with the scheme depicted in Figure 4.3: The EBS and the IF approach produced the most conservative and the most liberal CIs, respectively, whereas the WBS yielded coverages in between that were, on average, the most correct. Taking into account all scenarios, sample sizes, and time points, the mean absolute deviation from the target level of 95% was 2.42% for the WBS as compared to 2.49% and 2.61% for the IF approach and the EBS (see Section B.2 in Appendix B for the coverage probabilities in the scenarios not displayed here). This ranking applied to most settings with $\beta_{01A} \in \{0, 2\}$, although the performance of the various methods was not consistent w.r.t. early time points t (see Figures B.32, B.37, B.39, B.44, and B.45).

Figure 4.3: Coverage of the g-formula CIs in the scenario with light censoring and $\beta_{01A} = 2$.



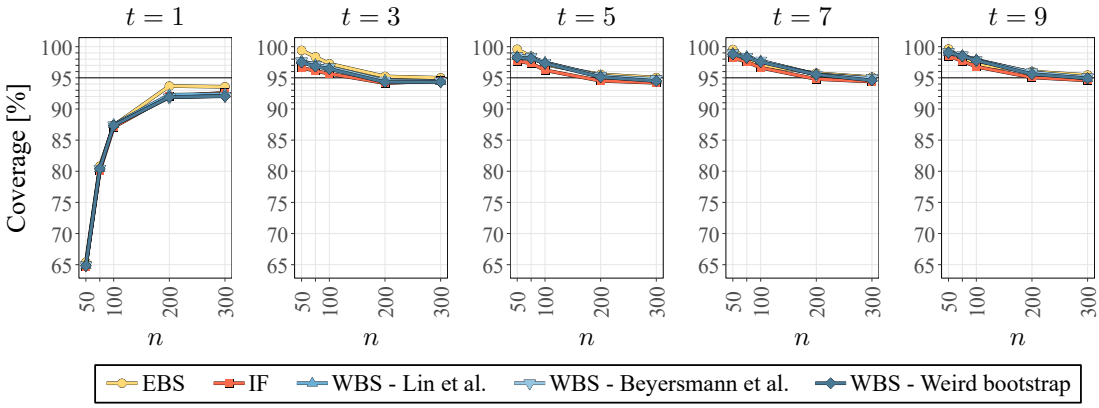
It should be mentioned that the accuracy of the WBS CIs peaked in particular at later analysis times (see Figures 4.3, B.32, B.39, B.42, B.45, B.51). The reason for this upturn is that the WBS relies on the product of the counting processes $N_{ki}(t)$ with the multipliers $G_{i_b}^{WBS}$, $k \in \{1, \ldots, K\}$, $i \in \{1, \ldots, n\}$, $b \in \{1, \ldots, B\}$. At late times, more instances of N_{ki} jump, and thus, a larger number of multipliers is factored in, which leads to a better approximation of the target distribution.

What is more, the different multipliers considered w.r.t. the WBS did not result in any noteworthy differences except that the CIs obtained by the Lin and Beyersmann approaches were more conservative. The associated coverages consequently exceeded those attributed to the weird bootstrap approach, but we did not find a general rule that characterizes the validity of the distinct WBS variants.

An exception to the ranking of the resampling methods according to Figure 4.3 occurred for instance in the scenario with high variance of the normally distributed covariates. Here, all resampling techniques entailed rather conservative CIs so that the IF approach – as the most liberal method – outperformed the other resampling techniques by a small margin (see Figures B.47 and B.48). Similar findings were made in several

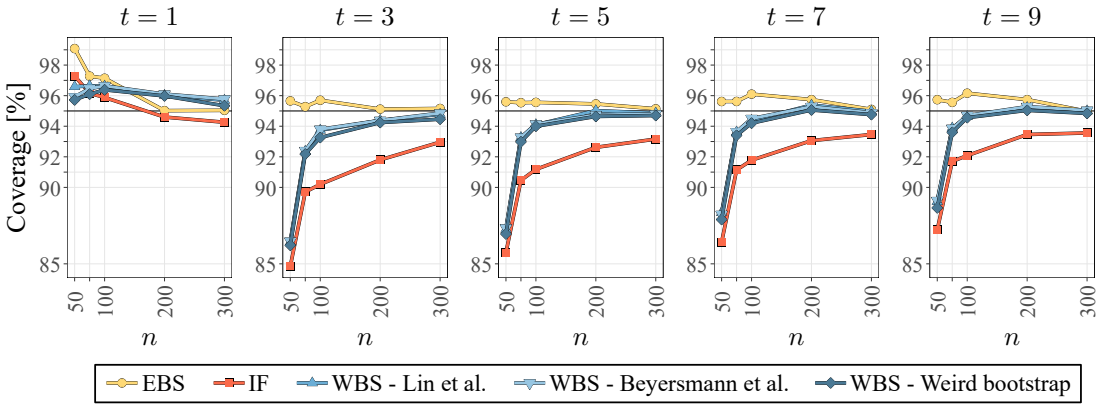
settings with $\beta_{01A} = -2$ (excluding the time point $t = 1$, where coverages generally fell short; see Figure 4.4 as well as Figures B.33, B.35, B.41, and B.43). The common trait of all these cases is that the competing event dominated over the cause of interest (see Table 4.2), and in consequence of the small proportion of observed type I events, the CIs apparently became too wide.

Figure 4.4: Coverage of the g-formula CIs in the scenario with no censoring and $\beta_{01A} = -2$.



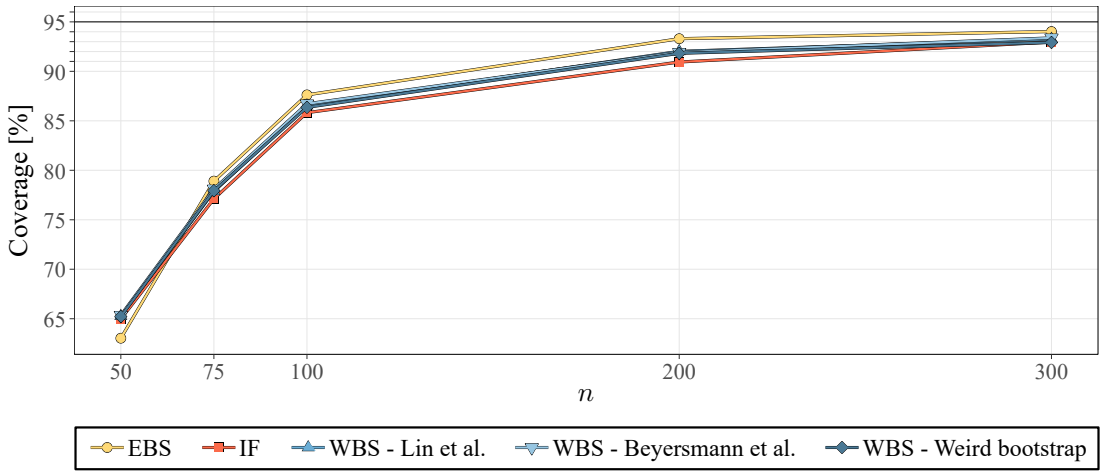
There were, conversely, several settings involving no treatment effect ($\beta_{01A} = 0$) where the IF approach yielded coverage probabilities that did not come particularly close to 95% even if the sample size approached 300 (see Figures 4.5, B.31, B.36, B.39). Ozenne, Scheike, et al. (2020) made similar observations, and they counteracted this distortion by means of a non-robust variance estimator.

Figure 4.5: Coverage of the g-formula CIs in the scenario with high treatment probability and $\beta_{01A} = 0$.



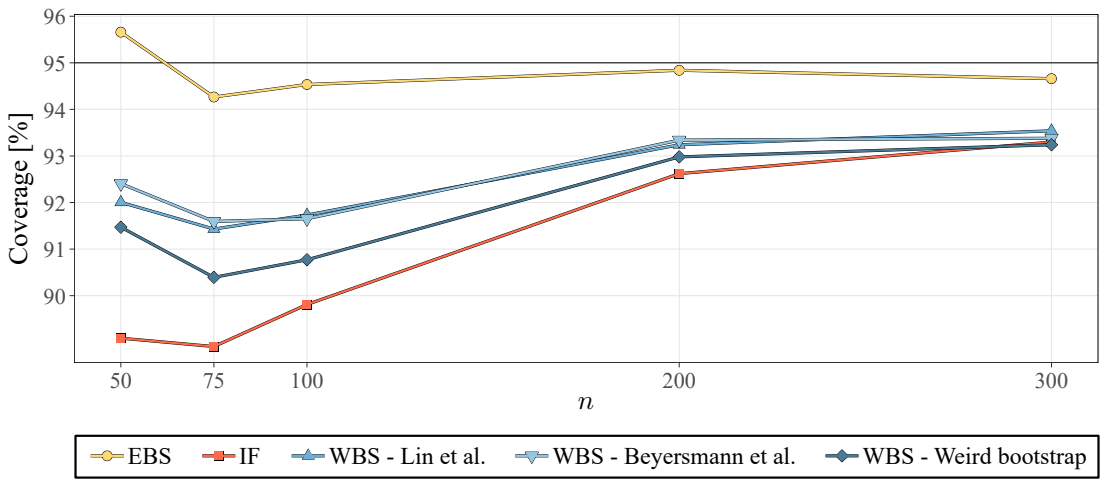
Lastly, it is worth noting that the WBS did not outperform the other resampling methods in the scenario with type II censoring and staggered study entry. Despite non-random censoring, the dependence structure in our examples was apparently not strong enough to reveal any issues with the methods relying on random censoring (cf. Chapter 3).

Figure 4.6: Coverage of the g-formula CBs in the scenario with heavy censoring and $\beta_{01A} = -2$.



We observed comparable patterns to those described above considering the time-simultaneous CBs. The coverage probabilities w.r.t. EBS, WBS, and IF mostly followed a descending order (see Figures B.53, B.56, B.58, B.61, B.64, B.67), yet all resampling techniques achieved only low coverages if $\beta_{01A} = -2$ and n was small (see Figures 4.6, B.52, B.55, B.63, B.66). The latter finding is likely attributable to an inadequate approximation of the distribution of $U_n(t)$ for the time point $t = 1$, taking account of the similarity between the listed figures and the first panel in e.g. Figure 4.4. It follows that the conservative EBS bands accomplished the most accurate coverages overall, with a mean absolute deviation of 4.75% from the nominal level, in comparison to 5.53% and 5.70% for the WBS and IF approaches, respectively. This result is also due to the fact that the EBS performed particularly well in settings with $\beta_{01A} = 2$ (see e.g. Figure 4.7).

Figure 4.7: Coverage of the g-formula CBs in the scenario with low variance of the covariates and $\beta_{01A} = 2$.

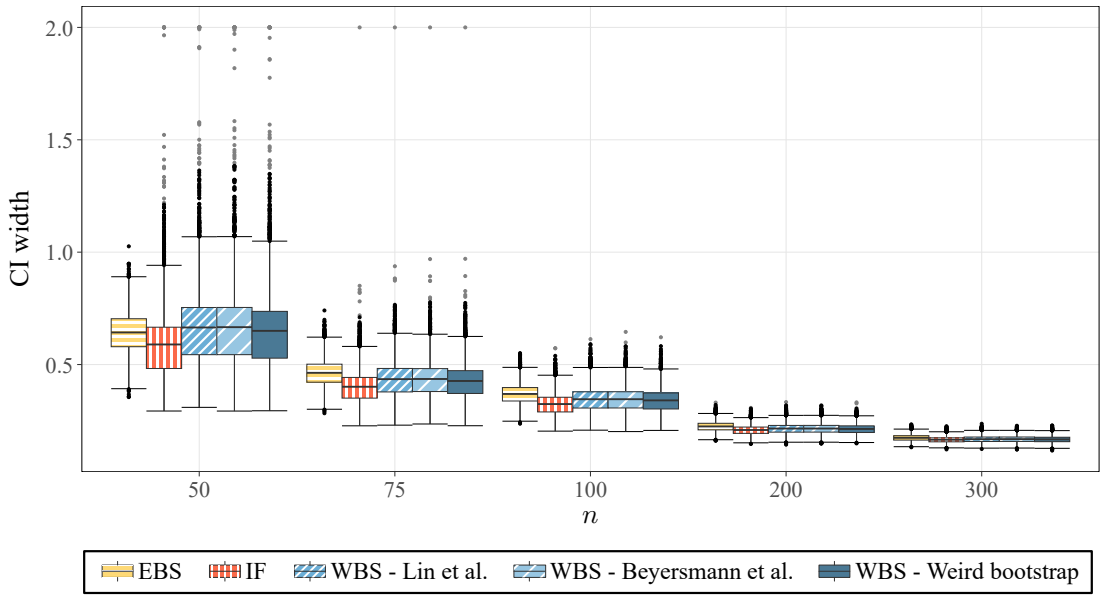


As already noted for the pointwise CIs, the CBs likewise showed no evidence that the different multiplier options for the WBS might have any significant impact. In addition, the CBs derived by the WBS did not stand out particularly in the scenario with type II censoring and staggered entry.

We further examined the size of the CIs and CBs. Figure 4.8 depicts the distribution of the CI widths at time $t = 5$ in the case without censoring and with parameter value $\beta_{01A} = 2$. It is evident that the IF approach led to the narrowest intervals, and this observation applied to nearly all examined scenarios. The range of the EBS CIs was either between or above the widths of the IF and WBS intervals, with the exception of the settings where $\beta_{01A} = -2$. Here, the CI widths varied considerably. With increasing sample sizes, the disparities w.r.t. the distinct resampling approaches became negligible, though.

A similar scheme was also observed in view of the time-simultaneous CBs.

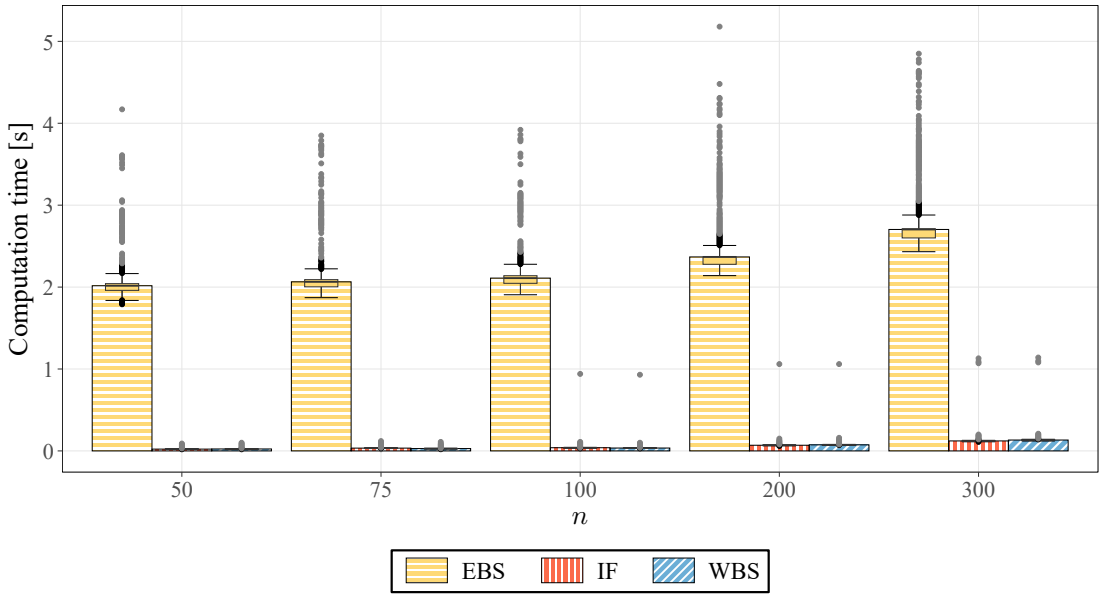
Figure 4.8: Widths of the g-formula CIs at time $t = 5$ in the scenario with no censoring and $\beta_{01A} = 2$. Note the spacing of the x-axis!



As a last aspect, we compared the running times of the methods, and in this context, it should be mentioned that the computation of the bootstrap replicates for the EBS approach was parallelized, whereas `C++` code was interfaced to perform the calculations for the IF and WBS methods. (The EBS and IF approaches were implemented by adapting the function '*ate*' of the `R` package '*riskRegression*' created by Gerds, Ohlendorff, and Ozenne, 2023.) We ran the simulation study on a high-performance computing cluster operating on 2.4 GHz Intel® processors with 128 GB RAM, while using 16 cores for

parallel computations. Figure 4.9 illustrates the execution times corresponding to each resampling method in the scenario without censoring and with parameter value $\beta_{01A} = 2$. (Note that the computation time for the WBS comprises all three subversions, since the additional steps necessary to cover the distinct multipliers contribute only marginally to the overall running time after the general components of the process have been determined.) The computation time for the EBS clearly surpassed that of the IF and WBS approaches by many times, which might limit its relevance in practical applications that involve larger sample sizes. The reason for the large imbalance between the execution times is that the EBS relies on repeated calculation of the ATE, whereas the IF method and the WBS approximate the distribution of the target process based on multipliers (see Section 2.3).

Figure 4.9: Mean computation times for the g-formula confidence regions in the scenario with no censoring and $\beta_{01A} = 2$. Note the spacing of the x-axis!



To sum up, our simulations showed that the EBS, the IF approach, as well as the WBS can be used for valid inference regarding the ATE. The performance of the distinct methods varied, however, depending on the circumstances: When a sufficient amount of events had been observed for the type of interest, the WBS yielded the most accurate results. Otherwise, our results suggested to resort to the IF approach.

4.1.3. Analysis of the Hodgkin's disease study

In order to illustrate the application of the resampling methods to real-world data, we analysed the ATE of radiation combined with chemotherapy vs. radiation therapy alone

in terms of the long-term disease progression among patients suffering from early-stage Hodgkin's disease (cf. Pintilie, 2006). The study data included 865 subjects with stage I or II lymphoma who had been admitted to the Princess Margaret Hospital in Toronto between 1968 and 1986. As a primary endpoint, we considered the time (in years) from diagnosis to first relapse or death, whichever occurred first. Table 4.3 summarizes the covariates recorded for each subject. One may access the data of the Hodgkin's disease study through the '*randomForestSRC*' `R` package created by Ishwaran and Kogalur (2024).

Table 4.3: Summary of the covariates recorded for the Hodgkin's disease study.

Covariate	Therapy	
	Radiation ($n = 616$)	Radiation & chemotherapy ($n = 249$)
Age [mean (SD)]	35.93 (16.37)	33.77 (12.86)
Sex		
female	285 (46.27%)	117 (46.99%)
male	331 (53.73%)	132 (53.01%)
Lymphoma stage		
I	266 (43.18%)	30 (12.05%)
II	350 (56.82%)	219 (87.95%)
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		
no	587 (95.29%)	199 (79.92%)
yes	29 (4.71%)	50 (20.08%)

We assumed for the analysis that the variables in Table 4.3 sufficed to fulfil the identifiability conditions presented in Subsection 2.2.1. Besides, it was necessary to manipulate repeated event times by adding $\mathcal{N}(0, 10^{-6})$-distributed random values so that any ties stemming from rounding were broken. Tests on the scaled Schoenfeld residuals were then conducted for cause-specific Cox models addressing relapse and death, respectively, to check whether these models fitted the data appropriately (cf. Subsection 2.1.2). The outcomes implied no violations of the proportional hazards assumption except for the variable 'age' in the relapse model, and the estimated coefficient in a corresponding model with time-dependent covariate was nearly constant (see Figures B.74 and B.75 in Section B.2 in Appendix B). We thus proceeded by analysing the ATE based on simple Cox models with time-constant covariates.

Figure 4.10: G-formula confidence regions for the average treatment effect on the risk of relapse.

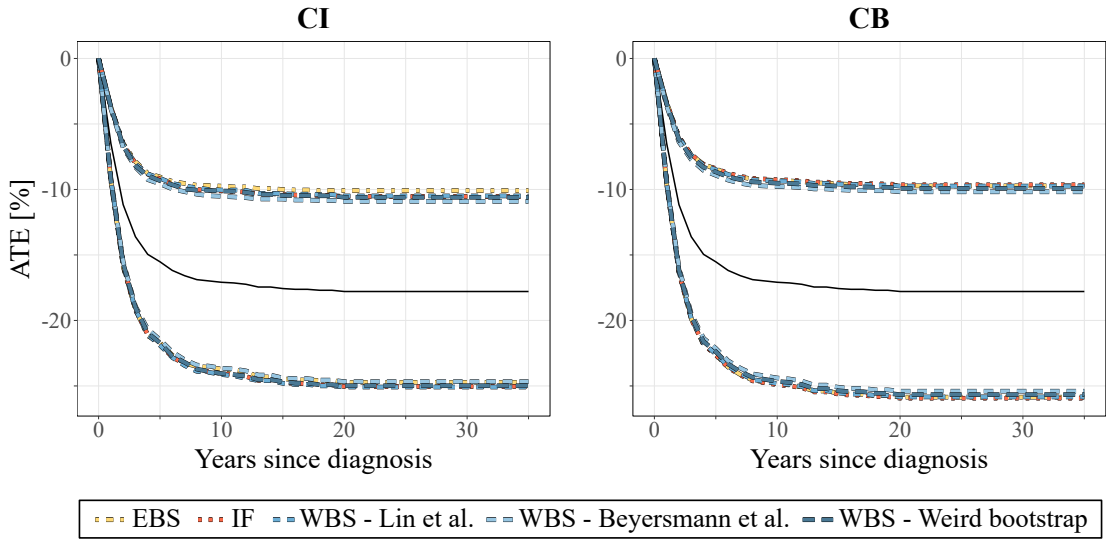
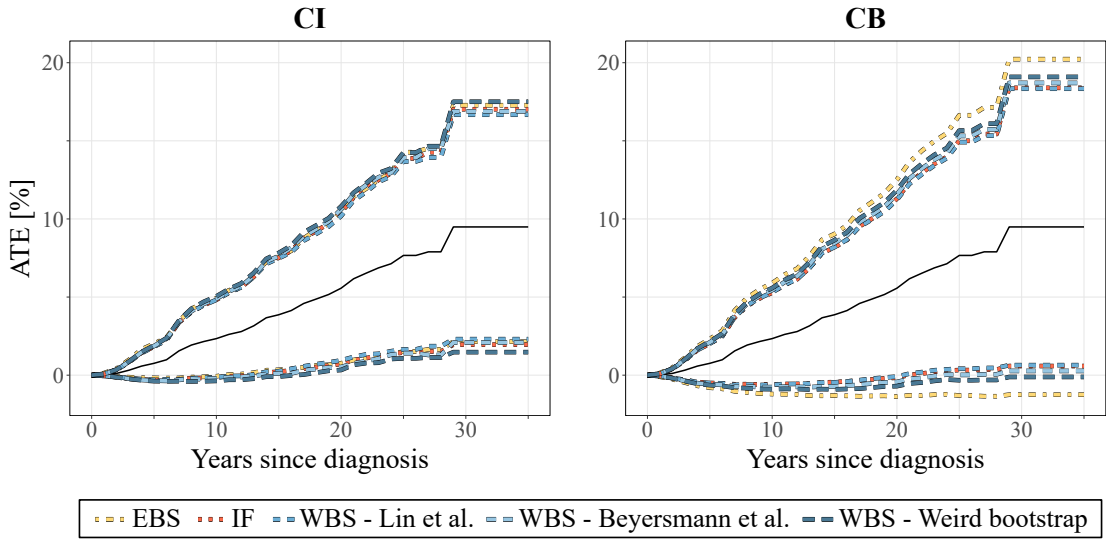


Figure 4.10 depicts the estimated effect of the combination therapy in comparison to treatment with radiation alone, the focus being on the event of first relapse. According to our analysis, the risk of recidivism would be reduced by 17.89% after 30 years had every patient been exposed to both radiation and chemotherapy instead of everyone being treated with radiation only. The risk of death, on the other hand, would rise by 9.49% (see Figure 4.11).

Figure 4.11: G-formula confidence regions for the average treatment effect on the risk of death.



It can further be seen that the ATE shown in Figure 4.10 decreases rather suddenly within the first 5 years after diagnosis, while the risk difference concerning death rises gradu-

ally throughout the entire study period. Our conclusion is that the combination therapy effectively prevents relapse among the studied patient collective. This means, however, that more subjects remain who will die later without experiencing any relapse event.

Figures 4.10 and 4.11 also illustrate the CIs and CBs for the ATE that have been obtained using the EBS, the IF approach and the WBS, respectively. Since there were 291 observed relapses and 135 deaths, the variation between the confidence regions in Figure 4.10 is somewhat smaller than in Figure 4.11. Overall, the differences are minor, though, with the only peculiarity relating to the CB for the event of death that has been determined by the EBS: Its width exceeds that of the remaining CBs, and this finding conforms to the results of the simulation study presented in Subsection 4.1.2, which identified the EBS CBs as the most conservative bands. Other than that, we did not find the order of the confidence widths in Figures 4.10 and 4.11 to be consistent, however.

4.2. Inference using propensity score matching

So far we have concentrated on the estimation of the ATE by means of the g-formula. An alternative approach is to identify $\left(\widehat{ATE}_m(t)\right)$ based on a PS-matched sample of the original population, as outlined in Subsection 2.2.1. Due to the nature of the matched data, one needs to take certain matters into account when employing resampling methods for inference about the ATE.

Consider an i.i.d. sample of n observations that includes the (general) outcome O_i and the vector $\mathbf{Z}_i$, which combines the treatment indicator Z_{Ai} and the covariate vector $\mathbf{Z}_{Li}$, for $i \in \{1, \ldots, n\}$. We will focus on one-to-one PS matching with replacement here. For each individual i, a counterpart j_i is determined, e.g. by nearest-neighbour matching on the PS:

$$j_i = \underset{j \in \{1,\ldots,n\}:\, Z_{Aj} \neq Z_{Ai}}{\arg\min} \left| \widehat{PS}\left(\mathbf{Z}_{Li}\right) - \widehat{PS}\left(\mathbf{Z}_{Lj}\right) \right|.$$

The variable $n_i = \sum_{k=1}^{n} \mathbb{1}\{j_k = i\}$ specifies how many times observation i serves as a match. Then, the matching estimator of $ATE = \mathbb{E}\left(O^{Z_A=1}\right) - \mathbb{E}\left(O^{Z_A=0}\right)$ is given by

$$\widehat{ATE}_m = \frac{1}{n} \sum_{i=1}^{n} \left(1 + n_i\right) \left(\mathbb{1}\{Z_{Ai}=1\} - \mathbb{1}\{Z_{Ai}=0\}\right) O_i.$$

Abadie and Imbens (2006) investigated the asymptotic properties of a broader class of matching estimators that incorporates the one above as a special case. They discovered that $\widehat{ATE}_m$ is biased, or more specifically, that the difference $\widehat{ATE}_m - ATE - B_n$ is asymptotically normally distributed with mean zero for $n \to \infty$. The bias term B_n is

given by

$$\frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{1}\{Z_{Ai}=1\}\cdot\left(\mathbb{E}\left(O \mid Z_A=0,\widehat{PS}(\mathbf{Z}_{Li})\right)+n_i\,\mathbb{E}\left(O \mid Z_A=1,\widehat{PS}(\mathbf{Z}_{Li})\right)\right)\right.$$
$$\left.-\mathbb{1}\{Z_{Ai}=0\}\cdot\left(\mathbb{E}\left(O \mid Z_A=1,\widehat{PS}(\mathbf{Z}_{Li})\right)+n_i\,\mathbb{E}\left(O \mid Z_A=0,\widehat{PS}(\mathbf{Z}_{Li})\right)\right)\right). \tag{4.5}$$

It follows that the corrected estimator $\widehat{ATE}_m^c=\widehat{ATE}_m-\widehat{B}_n$ should be preferred over $\widehat{ATE}_m$, where $\widehat{B}_n$ results from replacing $\mathbb{E}\left(O \mid Z_A=a,\mathbf{Z}_L=\mathbf{l}\right)$ in Equation (4.5) by a consistent estimator $\widehat{\mathbb{E}}\left(O \mid Z_A=a,\mathbf{Z}_L=\mathbf{l}\right)$ (Abadie and Imbens, 2011).

Let us now focus on TTE data $\left((T_i\wedge C_i,D_i,\mathbf{Z}_i)\right)_{i\in\{1,\ldots,n\}}$ of the same set-up as already considered in Section 4.1. Our interest is on the outcome $O=\mathbb{1}\{T\leq t,D=1\}$. One finds that

$$\begin{aligned}\widehat{ATE}_m(t)&=\frac{1}{n}\sum_{i=1}^{n}(1+n_i)\left(\mathbb{1}\{Z_{Ai}=1,T_i\leq t,D_i=1\}\right.\\&\qquad\left.-\mathbb{1}\{Z_{Ai}=0,T_i\leq t,D_i=1\}\right)\\&=2\hat{P}_m(Z_A=1,T\leq t,D=1)-2\hat{P}_m(Z_A=0,T\leq t,D=1)\\&=\hat{P}_m\left(T\leq t,D=1 \mid Z_A=1\right)-\hat{P}_m\left(T\leq t,D=1 \mid Z_A=0\right),\end{aligned}$$

with $\hat{P}_m$ denoting the empirical probability in the matched population, and due to censoring, we determine $\widehat{ATE}_m(t)$ as the difference between the estimated CIFs given $Z_A=1$ and $Z_A=0$, respectively, after matching. The bias estimator $\widehat{B}_n$ further results as

$$\frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{1}\{Z_{Ai}=1\}\cdot\left(\hat{F}_1\left(t \mid Z_A=0,\widehat{PS}(\mathbf{Z}_{Li})\right)+n_i\,\hat{F}_1\left(t \mid Z_A=1,\widehat{PS}(\mathbf{Z}_{Li})\right)\right)\right.$$
$$\left.-\mathbb{1}\{Z_{Ai}=0\}\cdot\left(\hat{F}_1\left(t \mid Z_A=1,\widehat{PS}(\mathbf{Z}_{Li})\right)+n_i\,\hat{F}_1\left(t \mid Z_A=0,\widehat{PS}(\mathbf{Z}_{Li})\right)\right)\right).$$

It has been pointed out by Abadie and Imbens (2008) that the EBS does not provide valid variance estimates in the given setting because it fails to replicate the distribution of n_i, the number of times observation i is used as a match. This is because the matching step is performed anew in each bootstrap sample, and consequently, the matches vary between the samples. As a remedy, Otsu and Rai (2017) proposed to resample from the linear form of $\widehat{ATE}_m^c$ in the context of a weighted bootstrap procedure, so that n_i is regarded as part of the data. Their solution refers to the case of matching on the full vector $\mathbf{Z}_L$, however, and cannot simply be adopted when matching on the PS, as it does not account for the variability of the estimator $\widehat{PS}(\mathbf{Z}_L)$. Adaptations of the suggested bootstrap procedure to PS matching have e.g. been investigated by Bodory et al. (2016) and Adusumilli (2022).

When TTE outcomes are considered, matters are complicated even more, since the usual survival estimators cannot be represented as linear forms. Wang et al. (2024) thus approximated the variance of the estimated causal HR by means of martingale residuals, while using a double-resampling (DR) approach similar to that suggested by Adusumilli (2022) in order to take the uncertainty of the PS estimators into account. We will hereafter adjust their technique to make inferences that pertain to the ATE.

The proof of Lemma 4.1 implies that

$$\begin{aligned}&\sqrt{n}\left(\widehat{ATE}_m(t) - ATE(t)\right)\\ &= \sum_{i=1}^{n} (1+n_i)\left(\mathbb{1}\{Z_{Ai}=1\} - \mathbb{1}\{Z_{Ai}=0\}\right) MR_i(t) + o_P(1),\end{aligned}$$

with martingale residuals

$$\begin{aligned}MR_i(t) = \frac{1}{\sqrt{n}}\sum_{k=1}^{K}\int_0^t &\Big(\mathbb{1}\{k=1\}\cdot S\big(u- \mid Z_{Ai}\big) - F_1\big(t \mid Z_{Ai}\big) + F_1\big(u \mid Z_{Ai}\big)\Big)\\ &\cdot \frac{\exp(\beta_{0k}Z_{Ai})}{S^{(0)}(\beta_{0k},u)}\,\mathrm{d}\big(N_{ki}(u) - Y_i(u)\exp(\beta_{0k}Z_{Ai})\,A_{0k}(u)\big)\\ + \frac{1}{\sqrt{n}}\sum_{k=1}^{K}\int_0^t &\Big(\mathbb{1}\{k=1\}\cdot S\big(v- \mid Z_{Ai}\big) - F_1\big(t \mid Z_{Ai}\big) + F_1\big(v \mid Z_{Ai}\big)\Big)\\ &\cdot \big(Z_{Ai} - e(\beta_{0k},v)\big)\exp(\beta_{0k}Z_{Ai})\;\mathrm{d}A_{0k}(v)\\ &\cdot \int_0^\tau \frac{Z_{Ai} - e(\beta_{0k},u)}{\Sigma_k}\,\mathrm{d}\big(N_{ki}(u) - Y_i(u)\exp(\beta_{0k}Z_{Ai})\,A_{0k}(u)\big)\,,\end{aligned}$$

if for each $k \in \{1,\ldots,K\}$, the structural hazards models $\alpha_k^a(t) = \alpha_{0k}(t)\exp(\beta_{0k}a)$ are valid with corresponding functions $A_{0k}(t)$, $S^{(0)}(\beta_{0k},t)$, $e(\beta_{0k},t)$, $S\big(t \mid a\big)$, $F_1\big(t \mid a\big)$, as well as the Fisher information Σ_k.

The first step of the DR approach is to determine the secondary (nearest-neighbour) matches

$$j_i^{SM}(a) = \begin{cases} i & \text{if } Z_{Ai} = a,\\ \underset{j\in\{1,\ldots,n\}:\, Z_{Aj}=a}{\arg\min} \|\mathbf{Z}_{Li} - \mathbf{Z}_{Lj}\|_2 & \text{if } Z_{Ai} \neq a,\end{cases}$$

and the set S_i of indices $j \in \{1,\ldots,n\}$ with $Z_{Aj} \neq Z_{Ai}$, for which $\widehat{PS}\left(\mathbf{Z}_{Lj}\right)$ falls into the same quintile partition of $\widehat{PS}(\mathbf{Z}_L)$ as $\widehat{PS}(\mathbf{Z}_{Li})$ ($i \in \{1,\ldots,n\}$). After randomly selecting an element j_{S_i} of S_i, the parameter n_i may be imputed by

$$n_i^{imp}(a) = \begin{cases} n_i & \text{if } Z_{Ai} = a,\\ n_{j_{S_i}} & \text{if } Z_{Ai} \neq a.\end{cases}$$

Adusumilli (2022) argues that the secondary matches j_i^{SM} should be based on the full covariate vector $\mathbf{Z}_L$ rather than on $\widehat{PS}(\mathbf{Z}_L)$ to maintain the correlation between $\mathbf{Z}_L$ and the bootstrap residual, conditional on $\widehat{PS}(\mathbf{Z}_L)$. Other than that, one may note that n_i^{imp} is not imputed by matching. The reason why we consider the partition of the PS quintiles instead is to avert correlation between n_i^{imp} and the original parameter n_i.

As a next step, the bootstrap quantities

$$\begin{aligned} Z_{Ai}^{(b)} &\sim \mathcal{B}in\Big(1, \widehat{PS}(\mathbf{Z}_{Li})\Big), \\ \widehat{PS}_i^{(b)} &= \hat{P}\Big(Z_{Ai}^{(b)} = 1 \mid \mathbf{Z}_L = \mathbf{Z}_{Li}\Big), \\ G_{i_b}^{DR} &\sim \mathcal{N}(0,1) \end{aligned}$$

are simulated for each $b \in \{1, \ldots, B\}$, and the estimate

$$\hat{\mu}_{a,i}^{(b)}(t) = \widehat{\mathbb{E}}\Big(MR(t) \mid Z_A = a, \widehat{PS}(\mathbf{Z}_L) = \widehat{PS}_i^{(b)}\Big)$$

is obtained by linear regression. Eventually, the DR technique approximates the distribution of the process U_n through the replicates

$$\Big(\widehat{U}_n^{DR;\,(b)}(t)\Big)_{b\in\{1,\ldots,B\}} = \left(\sum_{i=1}^{n}\Big(r_i^{(b)}(t) - R^{(b)}(t)\Big)\, G_{i_b}^{DR}\right)_{b\in\{1,\ldots,B\}},$$

where

$$\begin{aligned} r_i^{(b)}(t) = \hat{\mu}_{1,i}^{(b)}(t) - \hat{\mu}_{0,i}^{(b)}(t) \\ + \Big(\mathbb{1}\big\{Z_{Ai}^{(b)} = 1\big\} - \mathbb{1}\{Z_{Ai}^{(b)} = 0\}\Big)\Big(1 + n_i^{imp}\big(Z_{Ai}^{(b)}\big)\Big) \\ \cdot \Big(\widehat{MR}_{j_i^{SM}\left(Z_{Ai}^{(b)}\right)}(t) - \hat{\mu}_{Z_{Ai}^{(b)}, j_i^{SM}\left(Z_{Ai}^{(b)}\right)}(t)\Big) \end{aligned}$$

for $i \in \{1, \ldots, n\}$ and

$$\begin{aligned} R^{(b)}(t) = \frac{1}{n}\sum_{i=1}^{n}\Bigg(\hat{\mu}_{1,i}^{(b)}(t) - \hat{\mu}_{0,i}^{(b)}(t) \\ + \widehat{PS}_i^{(b)}\Big(1 + n_i^{imp}(1)\Big)\Big(\widehat{MR}_{j_i^{SM}(1)}(t) - \hat{\mu}_{1,\,j_i^{SM}(1)}(t)\Big) \\ - \Big(1 - \widehat{PS}_i^{(b)}\Big)\Big(1 + n_i^{imp}(0)\Big)\Big(\widehat{MR}_{j_i^{SM}(0)}(t) - \hat{\mu}_{1,\,j_i^{SM}(0)}(t)\Big)\Bigg), \end{aligned}$$

for plug-in estimators $\widehat{MR}_i(t)$ of $MR_i(t)$. Note that the bootstrap residuals specified by Wang et al. (2024) have been adjusted because we consider a linear contrast when examining the ATE instead of the HR.

An alternative way to derive an estimator for the variance of $\widehat{ATE}_m$ has been described by Austin and Cafri (2020). They reason that such an estimator needs to factor in several sources of correlation: Firstly, subjects and their matches have similar covariate values (given that their PSs are close to each other), and therefore, they are likely to have similar outcomes. Secondly, individuals may be used as a match more than once, so that there is correlation between the distinct instances of the same unit within the matched data. The fact that identical subjects can belong to separate matched pairs finally necessitates accounting for the corresponding cross-classification, which is accomplished by treating all subjects in the matched sample as independent, since the same individual will not be part of a matched pair more than once. Austin and Cafri (2020) – who focused on the HR as target outcome – hence proposed the total variance estimator

$$\widehat{\text{Var}}^{\,c} = \frac{n_{mp}}{n_{mp}-1}\,\widehat{\text{Var}}_{mp} + \frac{n_{di}}{n_{di}-1}\,\widehat{\text{Var}}_{di} + \frac{n_{cc}}{n_{cc}-1}\,\widehat{\text{Var}}_{cc}, \tag{4.6}$$

with n_{mp}, n_{di}, and n_{cc} denoting the number of matched pairs (i.e. $n_{mp} = n$), the number of distinct individuals in the matched data ($n_{di} = n$), as well as the size of the matched sample ($n_{cc} = 2n$). The terms $\widehat{\text{Var}}_{mp}$, $\widehat{\text{Var}}_{di}$, and $\widehat{\text{Var}}_{cc}$ furthermore refer to the variance estimators that are obtained by accounting for the clustering of the matched pairs, the distinct individuals, as well as the separate observations in the matched sample, respectively (cf. Lin and Wei, 1989). Note that the factors preceding the variance terms in Equation (4.6) are supposed to improve the performance of the estimator if the corresponding number of clusters is small. By its definition, $\widehat{\text{Var}}^{c}$ can be adapted to the case where inference targets the ATE considered here in a rather straightforward way, although correlations between distinct time points cannot be reflected.

4.2.1. Simulation study comparing the resampling approaches

We investigated the accuracy of the presented methods for PS matching-based inference about the ATE within the scope of a simulation study that followed the set-up described in Subsection 4.1.2. The data were generated the same way as before, but the ATE was estimated by $\widehat{ATE}_m^{\,c}$ instead of $\widehat{ATE}_{ds}$. For this purpose, we approximated $PS(\mathbf{Z}_L)$ by means of a logistic regression according to the model

$$PS(\mathbf{Z}_L) = \left(1 + \exp\Big(-\Big(\gamma_0 + \gamma_1 Z_{L_1} + \cdots + \gamma_{12} Z_{L_{12}}\Big)\Big)\right)^{-1},$$

and performed nearest-neighbour matching with replacement on the resulting estimates $\widehat{PS}(\mathbf{Z}_L)$. Cause-specific Cox models were then fitted on both the original and the matched data, respectively, in order to derive the CIFs that define $\widehat{ATE}_m$ and $\widehat{B}_n$

(see Section 4.2). We used the Peto-Breslow method to handle ties in the matched data (cf. Peto, 1972; Breslow, 1974).

With the aim of illustrating its limitations in the given setting, the EBS was applied to identify confidence regions for the ATE: After drawing (with replacement) from the original data and PS-matching the obtained bootstrap sample, the bias-corrected estimator $\widehat{ATE}^{c^*}_{m;\,b}(t)$ was calculated. We repeated this procedure for $b \in \{1, \ldots, B\}$ (with $B = 1,000$) and computed CIs as well as CBs in the same way as it had been done when conducting the EBS with regard to the g-formula estimator of the ATE (see Subsection 4.1.2).

The DR approach described earlier was additionally implemented with the CI

$$\Big[\widehat{ATE}^{c}_{m}(t) - q_{DR}(t;\,0.95),\ \widehat{ATE}^{c}_{m}(t) + q_{DR}(t;\,0.95)\Big]$$

as well as the CB

$$\Big[\widehat{ATE}^{c}_{m}(t) - q_{DR}(0.95)\sqrt{\widehat{\text{Var}}^{DR}(t)},\ \widehat{ATE}^{c}_{m}(t) + q_{DR}(0.95)\sqrt{\widehat{\text{Var}}^{DR}(t)}\Big].$$

Here, $q_{DR}(t;\,0.95)$ and $q_{DR}(0.95)$ denote the 0.95 quantiles of the sets $\Big(\big|\widehat{U}^{DR;\,(b)}_{n}(t)/\sqrt{n}\big|\Big)_{b\in\{1,\ldots,B\}}$ and

$$\left(\sup_{t\in[0,9]}\left|\frac{\widehat{U}^{DR;\,(b)}_{n}(t)}{\sqrt{n\,\widehat{\text{Var}}^{DR}(t)}}\right|\right)_{b\in\{1,\ldots B\}},$$

respectively. The expression $\widehat{\text{Var}}^{DR}(t)$ furthermore refers to the empirical variance of $\Big(\widehat{U}^{DR;\,(b)}_{n}(t)\Big)_{b\in\{1,\ldots,B\}}$.

Lastly, we employed the clustered variance estimator proposed by Austin and Cafri (2020) for the derivation of the confidence regions. To do so, both the IF and WBS approaches from Section 4.1 were applied to the matched data, but considering treatment as the only covariate and accounting for the respective clusters that lead to the variance estimators $\widehat{\text{Var}}^{IF}_{mp}(t)$, $\widehat{\text{Var}}^{IF}_{di}(t)$, $\widehat{\text{Var}}^{IF}_{cc}(t)$ as well as $\widehat{\text{Var}}^{WBS}_{mp}(t)$, $\widehat{\text{Var}}^{WBS}_{di}(t)$, $\widehat{\text{Var}}^{WBS}_{cc}(t)$. (The estimators for both approaches were determined using standard normally distributed multipliers.) We then calculated $\widehat{\text{Var}}^{cIF}(t)$ and $\widehat{\text{Var}}^{cWBS}(t)$ in line with Formula (4.6). The corresponding CIs were obtained as

$$\Big[\widehat{ATE}^{c}_{m}(t) - q_{\mathcal{N}(0,1)}(0.975)\sqrt{\widehat{\text{Var}}^{cIF}(t)},\ \widehat{ATE}^{c}_{m}(t) + q_{\mathcal{N}(0,1)}(0.975)\sqrt{\widehat{\text{Var}}^{cIF}(t)}\Big]$$

as well as

$$\Big[\widehat{ATE}^{c}_{m}(t) - q_{\mathcal{N}(0,1)}(0.975)\sqrt{\widehat{\text{Var}}^{cWBS}(t)},\ \widehat{ATE}^{c}_{m}(t) + q_{\mathcal{N}(0,1)}(0.975)\sqrt{\widehat{\text{Var}}^{cWBS}(t)}\Big],$$

respectively. For the CBs, we considered the formalizations

$$\left[\widehat{ATE}^{c}_{m}(t) - q_{cIF}(0.95)\sqrt{\widehat{\text{Var}}^{cIF}(t)},\ \widehat{ATE}^{c}_{m}(t) + q_{cIF}(0.95)\sqrt{\widehat{\text{Var}}^{cIF}(t)}\right]$$

and

$$\left[\widehat{ATE}^{c}_{m}(t) - q_{cWBS}(0.95)\sqrt{\widehat{\text{Var}}^{cWBS}(t)},\ \widehat{ATE}^{c}_{m}(t) + q_{cWBS}(0.95)\sqrt{\widehat{\text{Var}}^{cWBS}(t)}\right],$$

with the 0.95 quantile $q_{cIF}(0.95)$ of

$$\left(\sup_{t\in[0,9]}\left(\frac{n_{mp}}{n_{mp}-1}\left|\frac{1}{2n}\sum_{i=1}^{2n}\frac{\widehat{IF}_{ATE;\,mp}(t;\,\mathbf{O}_i^m)}{\sqrt{\widehat{\text{Var}}^{IF}_{mp}(t)}}\,G^{IF}_{i_b}\right|\right.\right.$$
$$+\frac{n_{di}}{n_{di}-1}\left|\frac{1}{2n}\sum_{i=1}^{2n}\frac{\widehat{IF}_{ATE;\,di}(t;\,\mathbf{O}_i^m)}{\sqrt{\widehat{\text{Var}}^{IF}_{di}(t)}}\,G^{IF}_{i_b}\right|$$
$$\left.\left.+\frac{n_{cc}}{n_{cc}-1}\left|\frac{1}{2n}\sum_{i=1}^{2n}\frac{\widehat{IF}_{ATE;\,cc}(t;\,\mathbf{O}_i^m)}{\sqrt{\widehat{\text{Var}}^{IF}_{cc}(t)}}\,G^{IF}_{i_b}\right|\right)\right)_{b\in\{1,\ldots B\}}$$

(for $\widehat{IF}_{ATE;\,c}(t;\,\mathbf{O}_i^m)$ denoting the empirical IF of the ATE in the matched sample w.r.t observation i, allowing for clustering according to $c \in \{mp, di, cc\}$), as well as the 0.95 quantile $q_{cWBS}(0.95)$ of

$$\left(\sup_{t\in[0,9]}\left(\frac{n_{mp}}{n_{mp}-1}\left|\frac{\widehat{U}^{WBS;\,(b)}_{n;\,mp}(t)}{\sqrt{2n\,\widehat{\text{Var}}^{WBS}_{mp}(t)}}\right| + \frac{n_{di}}{n_{di}-1}\left|\frac{\widehat{U}^{WBS;\,(b)}_{n;\,di}(t)}{\sqrt{2n\,\widehat{\text{Var}}^{WBS}_{di}(t)}}\right|\right.\right.$$
$$\left.\left.+\frac{n_{cc}}{n_{cc}-1}\left|\frac{\widehat{U}^{WBS;\,(b)}_{n;\,cc}(t)}{\sqrt{2n\,\widehat{\text{Var}}^{WBS}_{cc}(t)}}\right|\right)\right)_{b\in\{1,\ldots B\}}$$

(with the WBS plug-in estimator $\widehat{U}^{WBS;\,(b)}_{n;\,c}(t)$ determined in the matched sample while accounting for clustering by $c \in \{mp, di, cc\}$). For comparison, we also considered the confidence regions obtained by the standard (unclustered) IF and WBS approaches (see Subsection 4.1.2).

Each of the simulation scenarios listed in Table 4.2 was implemented with 5,000 Monte Carlo iterations. Table B.9 in Appendix B specifies the corresponding numbers of iterations that involved errors. Note that in addition to the issues already encountered with the simulation study described in Subsection 4.1.2, there were also some cases where

the Cox model that relates to the bias term suffered from convergence problems because of extreme PS estimates. For that reason, one should not attach too much importance to the results covering the settings with sample size $n = 50$, in particular for the scenarios with high treatment probability and with high variance of the covariates.

The CIs obtained by the methods that factor in the specific variance structure of the matched data generally led to coverage probabilities that were closer to the nominal level of 95%: Overall, the mean absolute deviation amounted to 6.27%, 6.31%, and 6.71% for the clustered IF approach, the clustered WBS method, and the DR technique, respectively, in comparison to 9.23%, 10.70%, and 15.20% for the WBS, the IF, and the EBS approaches. We did not observe a consistent order w.r.t. the coverages of the distinct methods like it was present for the simulations in Subsection 4.1.2, though.

Figure 4.12: Coverage of the PS-matched CIs in the scenario with light censoring and $\beta_{01A} = -2$.

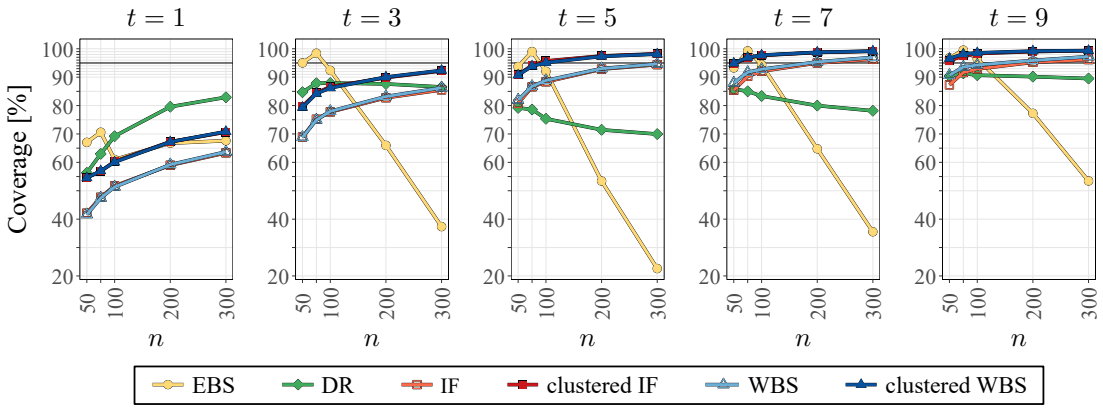
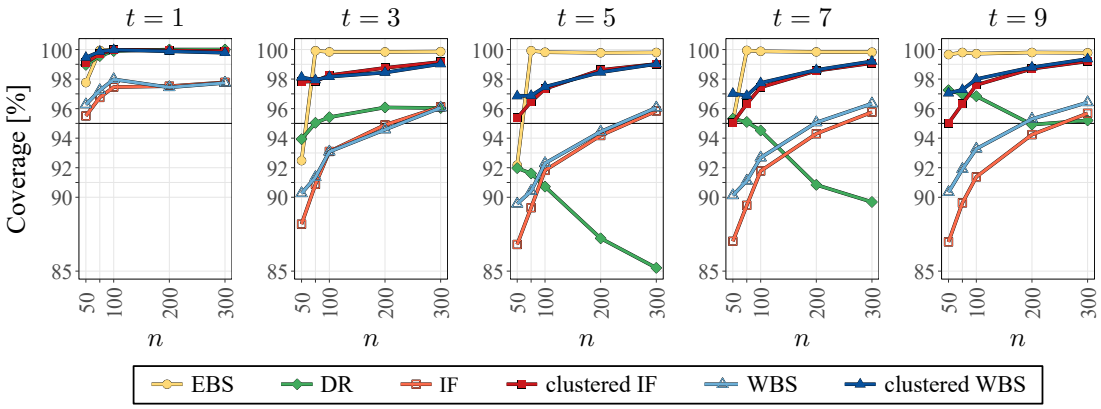


Figure 4.12 depicts the coverage probabilities in the scenario with light censoring and parameter $\beta_{01A} = -2$. It can be seen that for later time points, the clustered variance estimators led to accurate, yet slightly conservative CIs. The unclustered IF and WBS, on the other hand, were too liberal for early time points, but performed better if $t \in \{7, 9\}$. The CIs for the DR method were also found to be liberal, and the associated coverages interestingly dropped between the time points $t = 3$ and $t = 5$, only to increase again thereafter. Eventually, we noted that the EBS CIs were highly inaccurate for large sample sizes. This finding is plausible considering that there is greater variation w.r.t. the subjects that serve as a match if n is high. Since the choice of the matches is not reproduced by the EBS, there are strong deviations in the distribution of the matched bootstrap samples. Other scenarios with a treatment effect according to $\beta_{01A} = -2$ revealed similar outcomes as those shown in Figure 4.12 (see Figures B.76, B.81, B.88, and B.94).

The case with parameter $\beta_{01A} = 0$ yielded findings along the lines of Figure 4.13 (see also Figures B.77, B.79, B.89, B.95). We discovered that the coverages attained by the clustered resampling methods were again somewhat too high, while the standard IF and WBS approaches achieved correct results at later time points, given sufficiently large sample sizes. The DR method did not perform consistently throughout the distinct scenarios, however. The corresponding coverages actually worsened with larger values of n if $t \in \{5, 7\}$. In contrast to Figure 4.12, the EBS further provided very conservative CIs, which highlights that the bias of the associated variance estimator can point into both directions (see also Abadie and Imbens, 2008).

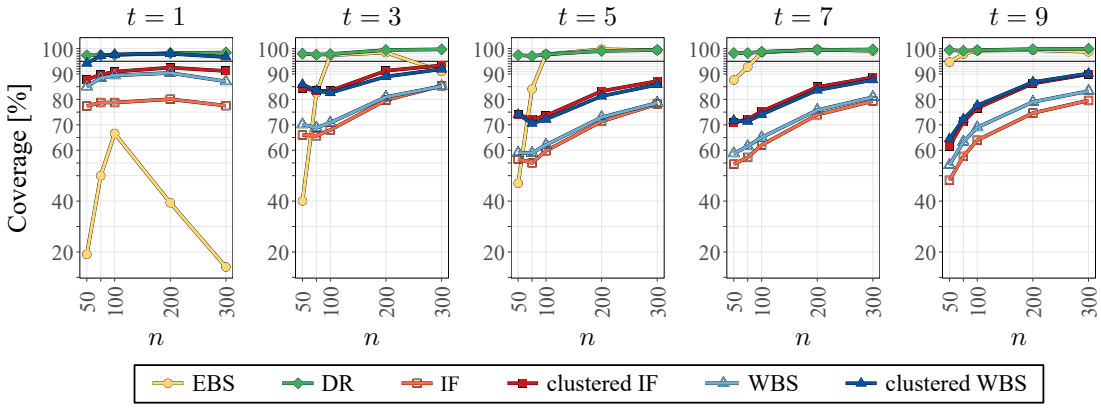
Figure 4.13: Coverage of the PS-matched CIs in the scenario with heavy censoring and $\beta_{01A} = 0$.



The differences between the resampling approaches were less pronounced for treatment effects according to $\beta_{01A} = 2$ (see Figures B.78, B.80, B.82, B.90, B.96). In this setting, the clustered IF and WBS yielded the most accurate coverages, whereas the DR approach was slightly too conservative, and the unclustered IF and WBS methods tended to be too liberal. The failure of the EBS became once again apparent in the form of low coverage probabilities for large sample sizes, especially w.r.t. early analysis time points.

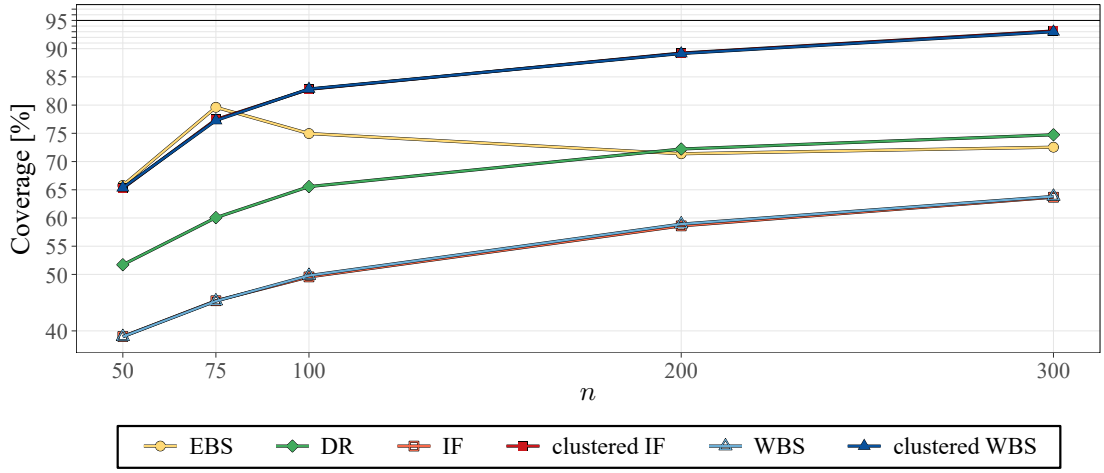
The scenarios with extreme treatment probabilities as well as the one with highly dispersed covariates led to slightly different findings compared to those described so far. At least one treatment group involved a smaller pool of suitable matches here than it had been the case for the other scenarios, so that the resulting pairs were less compatible. As a consequence, the resampling methods tailored to matching-based inference seemed to perform particularly well (see Figures B.83, B.84). The DR technique excelled in the scenario with high treatment probability (Figures 4.14, B.86, B.87). On the other hand, the coverages of the IF and WBS CIs were mostly too low (Figures B.91, B.92, B.93). The EBS further exhibited the same problems as already mentioned.

Figure 4.14: Coverage of the PS-matched CIs in the scenario with high treatment probability and $\beta_{01A} = 2$.



Turning our focus on the time-simultaneous CBs, there were hardly any surprises, except for the fact that the coverage probabilities associated with the EBS were comparably accurate: Their mean absolute deviation of 10.00% from the target level of 95% was only outperformed by the clustered WBS and IF methods, achieving 8.14% and 8.17% each, whereas the DR technique as well as the standard WBS and IF approaches yielded average deviations of 11.80%, 22.80%, and 24.10%, respectively.

Figure 4.15: Coverage of the PS-matched CBs in the scenario with light censoring and $\beta_{01A} = -2$.



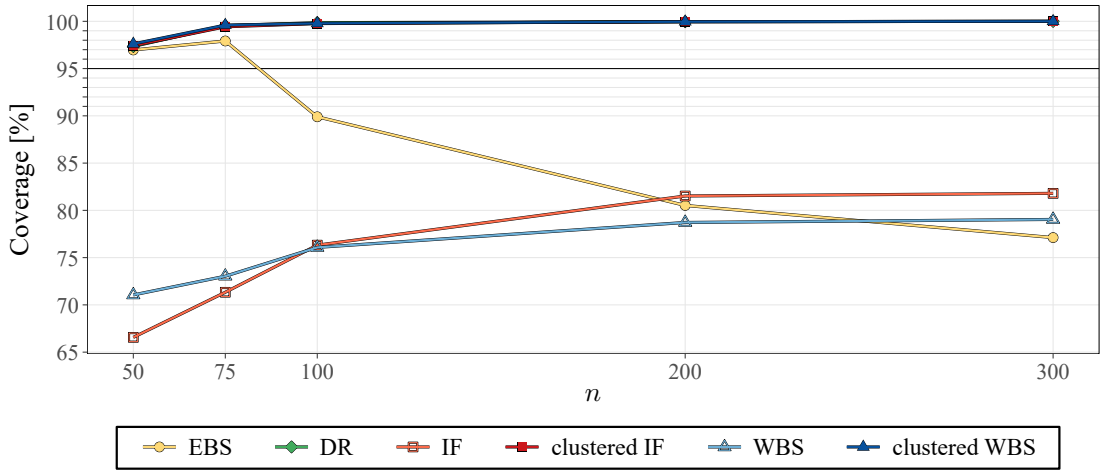
In the settings with $\beta_{01A} = -2$, all methods were rather liberal. The clustered resampling approaches performed best, their coverages ranging above those obtained by the DR method, which in turn exceeded the coverages for the standard IF and WBS. The EBS furthermore provided accurate results for small sample sizes. However, the corre-

sponding CBs became too liberal with increasing values of n (see Figures 4.15, B.97, B.102, B.111, and B.117).

Small sample sizes likewise entailed liberal CBs for the unclustered IF and WBS approaches in the settings without treatment effect, whereas the EBS yielded conservative bands. All three resampling methods improved with growing n. The DR technique moreover performed well for each of the sample sizes considered. What was disappointing is that the clustered versions of the IF and WBS attained too high coverages throughout (see Figures B.98, B.100, B.103, B.112, B.118). It has already been mentioned that the clustered variance estimator cannot reflect the dependence between the increments of the process for the ATE so that this observation is not too striking.

Finally, none of the resampling methods stood out particularly given treatment effects according to $\beta_{01A} = 2$. While the clustered IF and WBS as well as the DR technique produced conservative CBs, the IF and WBS approaches were too liberal. The coverage probabilities for the EBS CBs furthermore proceeded from being too high to assuming fairly low values when n became larger (see Figures 4.16, B.99, B.101, B.104, B.113).

Figure 4.16: Coverage of the PS-matched CBs in the scenario with type II censoring and $\beta_{01A} = 2$.

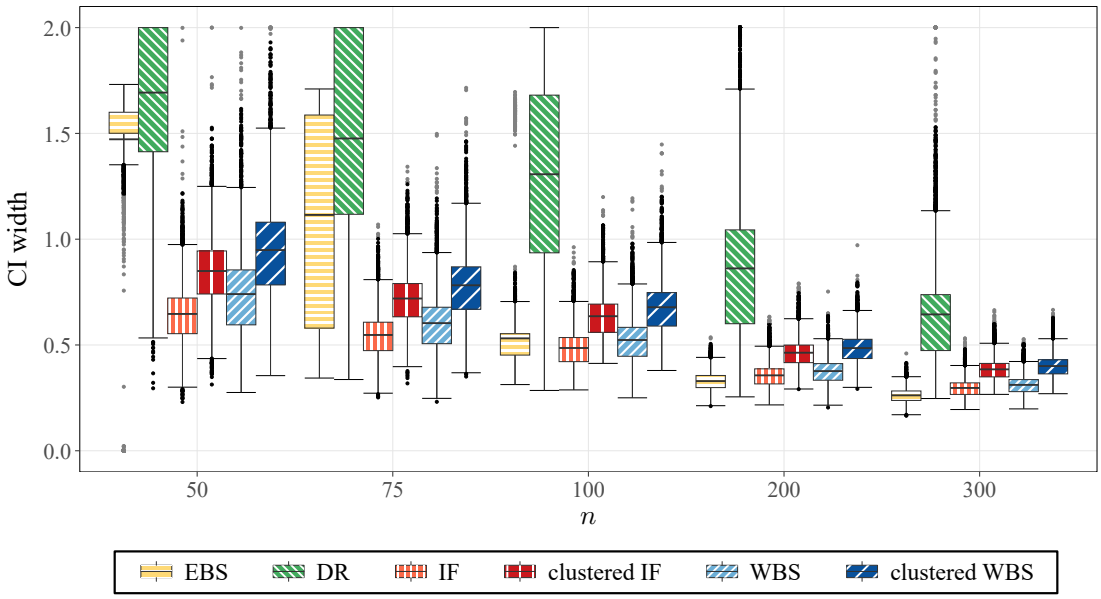


There were again some exceptions to the described findings in the scenarios with extreme treatment probabilities and the one with high variance of the covariates. The unclustered IF and WBS approaches were – as before – too liberal, but the performance of the clustered methods, the DR technique, as well as the EBS varied. For high treatment probabilities, the DR approach provided appropriate coverages (see Figures B.109, B.110), and the EBS generally performed well (see e.g. Figures B.106, B.116). The coverages of the CBs obtained by means of the clustered IF and WBS differed between the settings, though (see e.g. Figures B.105 and B.114).

Figure 4.17 depicts the widths of the CIs derived in the scenario with no censoring and $\beta_{01A} = 2$. As can be seen, the DR technique resulted in the largest intervals. The CIs corresponding to the clustered approaches were somewhat wider than those obtained using their unclustered versions (with the IF CIs having smaller ranges than the WBS CIs). This observation is not surprising considering that the clustered methods account for the additional variance that arises from the matching procedure. The widths of the EBS CIs, relative to the remaining intervals, varied depending on the sample size.

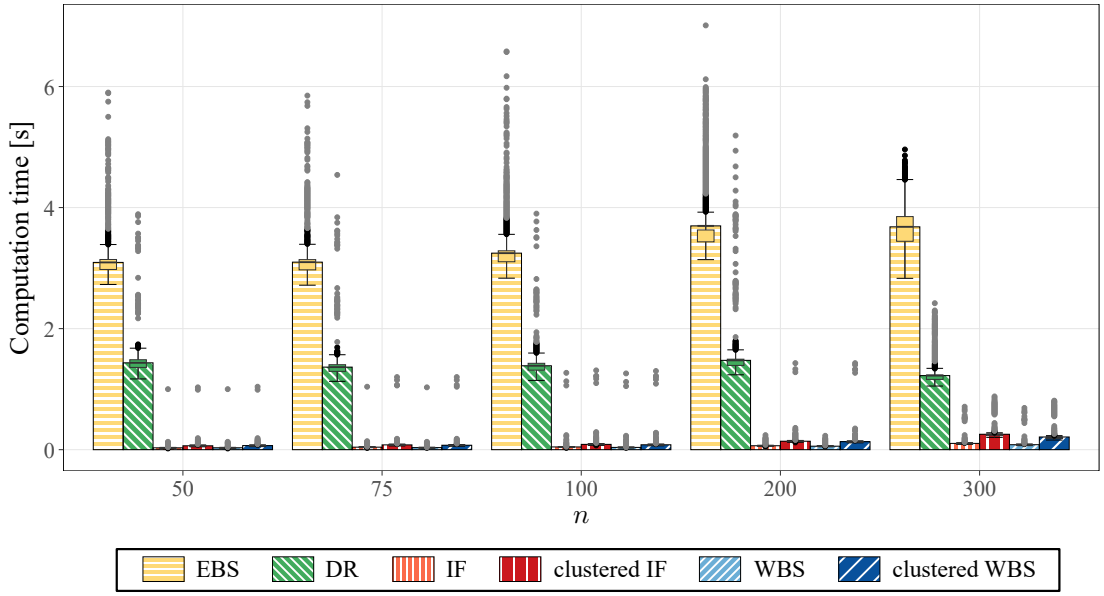
We observed similar patterns for the CI and CB widths in the other settings, although the confidence regions for the DR technique were narrower if $\beta_{01A} = 0$, and especially if $\beta_{01A} = -2$.

Figure 4.17: Widths of the PS-matched CIs at time $t = 5$ in the scenario with no censoring and $\beta_{01A} = 2$. Note the spacing of the x-axis!



Finally, the execution times for the different resampling methods are illustrated in Figure 4.18. With the exception of the EBS, which was conducted via parallel computations, all approaches employed interfaced `C++` code for increased speed. We used the same processors as specified in Subsection 4.1.2 for our simulations. It is evident that the EBS was by far the slowest method, which can be attributed to the repeated calculation of $\widehat{ATE}_m^c$ in distinct samples. The DR technique was about two to three times as fast, but still took significantly more time than the remaining methods due to its complexity. Furthermore, the clustered methods were slightly slower than their unclustered counterparts. This is because they involve performing the same computations three times, namely once w.r.t. each cluster structure.

Figure 4.18: Mean computation times for the PS-matched confidence regions in the scenario with no censoring and $\beta_{01A} = 2$. Note the spacing of the x-axis!

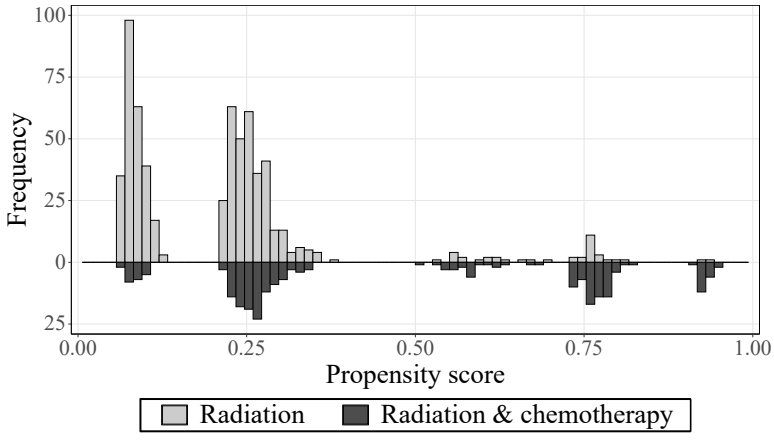


All in all, we found that valid CIs for the PS-matched ATE are provided by the DR technique as well as the clustered IF and WBS, although the latter, in particular, are somewhat conservative and their unclustered versions do not perform crucially worse. The EBS may in contrast entail considerable bias in large samples. To our surprise, the EBS CBs were fairly accurate in the investigated scenarios, however.

4.2.2. Analysis of the Hodgkin's disease study

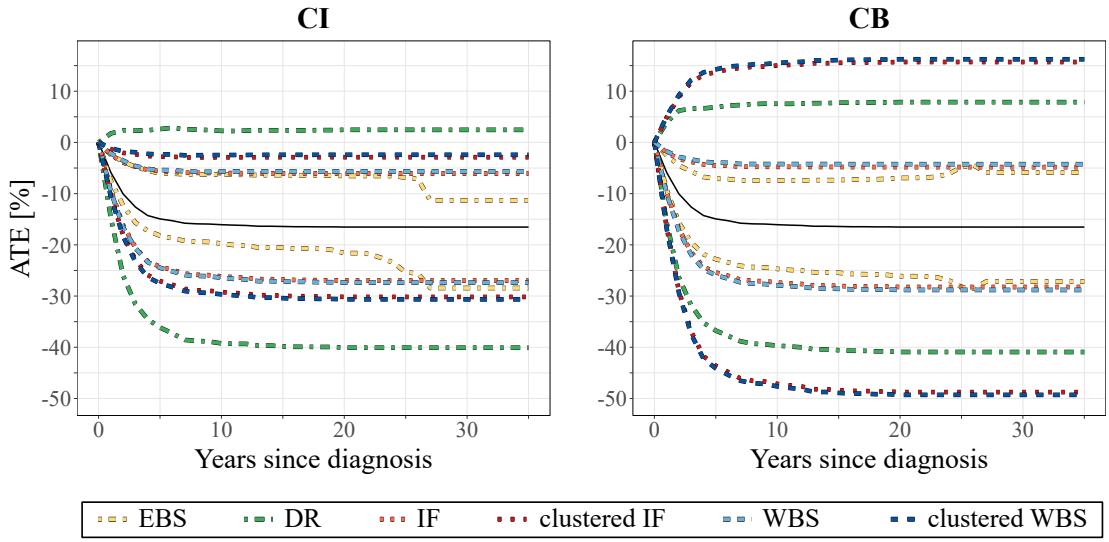
In order to demonstrate the use of the discussed resampling methods, we re-analysed the data already examined in Subsection 4.1.3 by means of PS matching. The PSs were approached by logistic regression on the covariates listed in Table 4.3. Figure 4.19 contrasts the frequencies of the dis-

Figure 4.19: Distribution of the PSs.



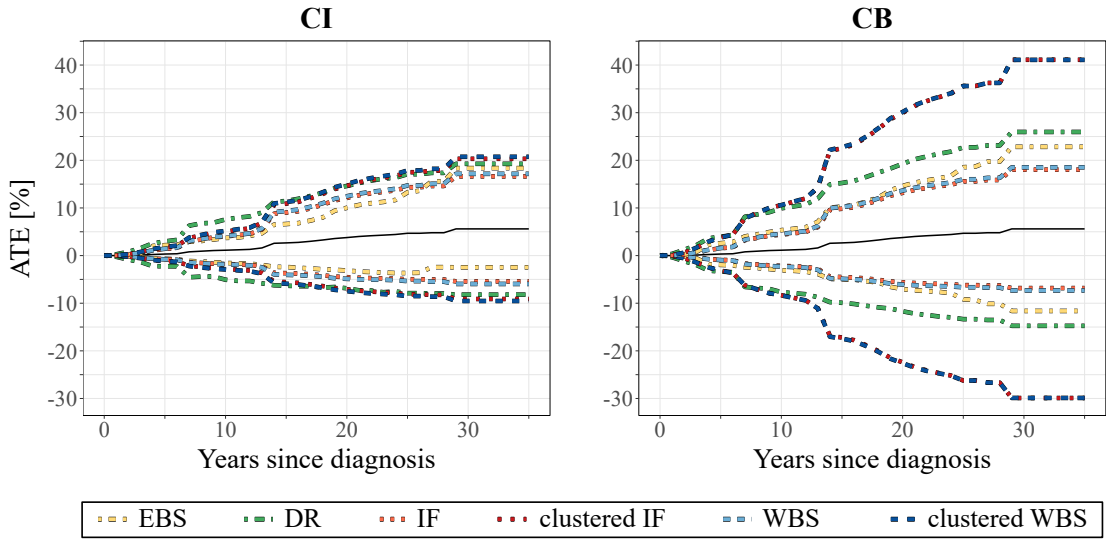
tinct PS estimates in both treatment groups. Even though the supports of both distributions have gaps, they do mostly overlap so that we may assume positivity. Besides, the PSs do not seem to be distributed equally in the groups, implying that there is in fact confounding by the considered covariates.

Figure 4.20: PS-matched confidence regions for the average treatment effect on the risk of relapse.



The estimated ATEs of the combined therapy vs. radiation alone are depicted in Figures 4.20 and 4.21, addressing the events of first relapse and death, respectively. The 30-year estimators of the risk difference reached values of -16.51% as well as 5.59%, and were thus more conservative than the corresponding g-formula estimators (which had been calculated as -17.89% and 9.49%, respectively).

Figure 4.21: PS-matched confidence regions for the average treatment effect on the risk of death.



We found the confidence regions obtained by the matching-based resampling methods to be considerably wider than those computed on the basis of the g-formula (see Figures 4.10 and 4.11). The clustered IF and WBS approaches and the DR technique resulted in the largest confidence regions, with the clustered methods yielding particularly wide ranges for the time-simultaneous CBs. In contrast, the confidence regions obtained by the standard IF and WBS were rather small. The EBS produced CIs and CBs that were comparably narrow, which is in line with our observations in the simulation study described in Subsection 4.2.1, given that the sample size of the Hodgkin's data amounted to 865.

5. Conclusion

In this thesis, we investigated the use of resampling methods for statistical inference on TTE data. Particular attention was paid to the dependence structure of the data in event-driven trials with staggered subject entry as well as to endpoints defined by the causal risk difference.

We summarize and discuss the findings from the previous chapters hereafter and point out potential topics for further research.

5.1. Summary

Our first focal point was the special case of event-driven trials that involve successive entry of the participants. Because the study duration – and thus, administrative censoring – is determined by the event times, the data in such trials are dependent. We established that the condition of independent censoring in the counting process sense is yet fulfilled, provided that the calendar times of the events are concealed. Consequently, it is valid to use standard survival methodology for the analysis of event-driven trials with staggered entry.

We conducted a simulation study to showcase the consequences of conditioning on calendar times when analysing studies of the given type. The additional information on the sequence of the events that is provided through the calendar times led to biased estimators of the associated Cox regression coefficients, and thus, to biased Breslow estimators of the cumulative baseline hazard. The extent of the bias was more pronounced, the more specific the available information on the order of the events was, and the smaller we chose the sample size as well as the number of events to be observed until study closure.

Another simulation study further contrasted the EBS and the WBS in the setting of event-driven trials with staggered entry (cf. Efron, 1979; Lin, Wei, and Ying, 1993). It should be noted that the EBS relies on the assumption of independent data, whereas the WBS is valid under independent censoring. Event-driven trials with staggered entry, however, involve independent, but not random censoring. We used both resampling methods to derive CIs for the cumulative hazard function, and did in fact find the coverage probabilities obtained by the wild bootstrap to be more accurate. The differences between both approaches were more marked in smaller samples and when fewer events had been observed before the end of follow-up.

Lastly, we examined the performance of the EBS and the WBS on the basis of real-world data that covered the survival of patients with non-small cell lung cancer (Rittmeyer et al., 2017). While the application of the resampling methods to the original data did not result in any noteworthy findings, differences between the respective CIs for the EBS and the WBS became apparent in random data subsets involving only few patients.

For the remaining considerations, we shifted our focus to causal effect estimates in general clinical trials. The ATE was defined as the difference between the causal CIFs in the treatment groups, so that both standard survival as well as competing risks settings could be accommodated.

As a first step, we characterized the stochastic process related to the g-formula estimator of the ATE via martingales, which allowed us to prove the validity of three different resampling methods – namely the EBS (Efron, 1979), a technique on the basis of the IF (where resampling is, strictly speaking, only necessary to draw conclusions that apply to multiple time points simultaneously; see Scheike and Zhang, 2008), as well as the WBS (Lin, Wei, and Ying, 1993; Beyersmann, Di Termini, and Pauly, 2013, Andersen, Borgan, et al., 1993, Subsection IV.1.4) – for approximating the asymptotic distribution of the ATE estimator.

The finite-sample performance of the mentioned resampling methods was subsequently examined by simulations. Each approach was applied throughout a variety of scenarios to derive pointwise CIs and time-simultaneous CBs for the ATE, and we found that the confidence regions obtained by the WBS ranged mostly between the more conservative EBS and the liberal IF regions. Overall, the WBS provided accurate coverages for the pointwise CIs unless the amount of the observed events was very small. In that case, the IF performed better, and w.r.t. the time-simultaneous CBs, the EBS achieved the most appropriate coverages (which may be attributed to the poor performance of the other approaches at early analysis time points). As the sample size increased, the confidence regions associated with the distinct methods became more similar. It should finally be mentioned that the computation times for the EBS were considerably higher than those measured for the IF approach and the WBS.

To illustrate the application of the resampling methods, we derived confidence regions for the ATE considering real data on the disease progression among patients suffering from Hodgkin's disease (Pintilie, 2006). The sample size was quite high, though, so that the obtained CIs and CBs differed only marginally.

Our investigations eventually addressed resampling-based inference relying on PS-matched estimators of the ATE. We adapted a DR technique (Wang et al., 2024) and a cluster-based variance estimator (Austin and Cafri, 2020) – which had both been proposed for inference about the causal HR – to the definition of the ATE at hand.

The performance of these methods was evaluated w.r.t. the same simulation set-up as already implemented when exploring the g-formula methods. In order to enable comparisons, we further applied the EBS as well as the standard IF and WBS approaches to the PS-matched data. There was no consistent pattern characterizing the corresponding coverage probabilities, but the DR technique and the clustered variance estimator generally provided decent, yet somewhat conservative confidence regions. The standard IF and WBS tended to be too liberal, on the other hand. We expected considerable deviations concerning the EBS (cf. Abadie and Imbens, 2008), but to our surprise, the bias observed in the coverages of the CIs dissolved for the greater part w.r.t. the CBs.

The Hodgkin's disease data was finally re-analysed on the basis of PS matching, and the resulting confidence regions turned out to be notably wider than those derived earlier. We also discovered large differences between the outcomes that were associated with the different resampling methods. To be more precise, the DR technique and the clustered variance estimator led to wider confidence regions than the remaining approaches.

5.2. Discussion

Event-driven trials with staggered entry are frequently employed in clinical practice. It is all the more important to raise awareness of the dependence that is inherent to the data collected in such trials, since the validity of the standard survival techniques hinges on the condition of independent censoring and it is not obvious whether that condition is met in the setting at hand. Theorem 3.1 ensures that event-driven trials with staggered entry indeed entail independent censoring. Thus, we have justified the use of the common statistical methodology for the evaluation of these studies.

The proof of the mentioned theorem requires that no information about the calendar times of the events is used for the analysis, because this would allow to retrace the order of the occurrences and accordingly, the underlying intensities would be deranged. Examples where this constraint is disregarded can be found in several studies conducted during the COVID-19 pandemic. To differentiate between the diverse conditions prevailing between 2019 and 2023, clinical trials were often subdivided into pre-, during, and post-pandemic stages based on calendar times (cf. EMA, 2020; R. D. Meyer et al., 2020). The potential consequences of such an approach became apparent in our simulations: Even though the bias of the estimated Cox regression coefficients for the treatment indicator was only minimal, the estimated HRs for the entry times deviated significantly from their true values so that we obtained flawed Breslow estimates. For practical applications, the corresponding bias in predictions of the survival probabilities may be of greater concern. It is therefore vital to carefully deliberate the choice of the covariates when analysing event-driven trials with staggered entry.

The considerations surrounding Theorem 3.1 further implied that methods relying on random censoring are not valid in the given setting. Our simulations exhibited that under certain conditions, the EBS may in fact lead to CIs with fairly inaccurate coverage probabilities. This issue was only observed under additional pressure exerted through internal left-truncation and in small samples, however, while in reality, trials typically involve more participants. Interim analyses, on the other hand, involve only subsets of the original population, so that the outlined problem may very well be relevant. One should generally stick to survival methods based on martingale properties for valid inference, as such methods depend on the assumption of independent rather than random censoring.

When changing our focus to causal effect estimation, we decided to define the ATE on the basis of the CIF to avoid the issues that arise with causal inferences pertaining to HRs. Note that the concomitant interpretation of the treatment effect does not depend directly on the HR although the underlying hazards are specified by Cox models (see e.g. Hernán, 2010; Martinussen and Vansteelandt, 2013; Aalen, Cook, and Røysland, 2015 for more insights into the limitations of HRs in the context of causal investigations).

By our definition of the ATE, we capture the total impact of an exposure on the event of interest, but cannot distinguish between direct and indirect effects. Approaches that are able to do so have been suggested by Rubin (2006), Stensrud et al. (2022), and Martinussen and Stensrud (2023). Those methods are generally based on untestable preconditions, though. We focus for now on situations where the interest lies on the total effect.

The estimation of the ATE and the justification of the corresponding resampling techniques require several assumptions, including the validity of the identifiability conditions, the absence of ties, conditional independence between the event and censoring times, as well as proportional hazards w.r.t. each cause. We have restricted our deliberations to the case where these conditions are fulfilled. For a relaxation of the proportional hazards assumption, one may consider alternatives to the Cox model, such as an additive model (Aalen, 1989) or the additive-multiplicative Cox-Aalen model (Scheike and Zhang, 2002). Using a different hazards model necessitates to re-evaluate the proofs in Subsection 4.1.1 as well as the considerations in Section 4.2 w.r.t. the new definition of the CIFs, however. Similarly, the use of the Fine-Gray model is possible, but likewise requires the reassessment of our proofs (Fine and Gray, 1999). The assumption-lean Cox regression proposed by Vansteelandt et al. (2024) is another reasonable option that offers greater flexibility when modelling the association between covariates and event times.

Inference for the g-formula estimator of the ATE is in practice almost always made by means of the EBS. The limitations of this approach w.r.t. dependent data have already been discussed (see also Singh, 1981; Friedrich, Brunner, and Pauly, 2017; Hrba et al., 2022). We have proposed two alternative resampling methods and demonstrated that given i.i.d. data, all three are valid for approximating the asymptotic distribution of the ATE estimator. Our simulation study further explored the performance of these methods in distinct scenarios. Based on the outcomes, we were able to give recommendations on which approach to use in which situation.

The simulations did not replicate the failure of the EBS in light of type II censoring with staggered entry (probably due to the weak dependence of the data), but made it clear that the IF and WBS are considerably faster. What is more, both approaches offer the possibility to implement aspects characteristic for TTE data (such as left-truncation) without further endeavour, since their set-up already accommodates the counting process framework inherent to survival analysis.

We also investigated resampling methods adapted to the PS-matched estimator of the ATE; however, our considerations on this subject should be regarded as preliminary. Research addressing inference for PS-matched TTE data is scarce and there are particularly few sources that focus on endpoints other than the HR.

We found the DR technique and even more so, the clustered variance estimator to yield rather conservative results, especially w.r.t. the time-simultaneous CBs. An explanation for the latter is that the clustered variance approach has been introduced as a method for inference about pointwise estimators, and hence, correlations w.r.t. distinct time points are not taken into account. Other than that, one should note that the execution of the DR technique is relatively slow.

5.3. Outlook

Our considerations and findings throughout this dissertation opened up scope for a number of related research topics.

The first simulation study implied that the intensities in event-driven trials with staggered entry might be deranged even more by additional information on the order of the events. It would be interesting to explore the consequences of disclosing which individuals had entered the study earlier than a particular subject, respectively. Besides, it remains to be checked whether the data involve stronger dependencies for smaller ratios between the number m of observed events and the sample size n than those considered here.

We plan to investigate the effects of deviations from the assumptions imposed w.r.t. the ATE. The use of hazards models other than the Cox model has already been mentioned before. Apart from that, we want to examine the extent to which the outcomes change if the individual preconditions are violated.

In order to permit the application of the considered resampling methods in more general contexts, they shall be extended to allow for left-truncation. (Note however that the matter of left-truncation in causal contexts bears potential for controversy, see Vandenbroucke and Pearce, 2015; Hernán, 2015.)

Therapies and confounders that change over the course of time are another commonly encountered issue that needs to be explored. The naive application of time-dependent Cox models leads to incorrect causal conclusions (Hernán, Brumback, and Robins, 2000). Estimators that can handle time-varying confounding and may hence serve as a foundation for our investigations have for instance been suggested by Hernán, Brumback, and Robins (2000), Keogh et al. (2023), and Rytgaard and van der Laan (2024).

Lastly, the scope of the resampling methods studied here could be expanded considerably by extensions to multistate models, although it is to be expected that the corresponding proofs involve higher levels of complexity. Among the situations that would be covered are illness-death-scenarios, recurrent events, as well as more elaborate settings (cf. e.g. Gran et al., 2015; Bühler, Cook, and Lawless, 2023).

When it comes to the PS-matched ATE estimator, future research will involve refinements of the examined resampling methods that ideally lead to less conservative results. It might also be enlightening to repeat our simulations using other matching methods and higher numbers of matches.

Finally, we seek to complement the findings of this thesis by insights into approaches that allow inference for IPT-weighted and doubly-robust estimators of the ATE. The derivation of the asymptotic distributions for these estimators as well as the justification of the respective resampling methods provide plenty of material for future considerations (see Ozenne, Scheike, et al., 2020 for the corresponding IFs).

Appendix A: Proofs

In this chapter, we provide the pending proofs of the theorems from the main text.

A.1. Independent censoring in event-driven trials with staggered entry

Proof of Theorem 3.1:
To show that $\lambda_i^{\mathscr{G}}(t) = \lambda_i^{\mathscr{F}^c}(t)\ \forall i \in \{1, \dots, n\}, t > 0$, note that

$$\begin{aligned}\lambda_i^{\mathscr{F}^c}(t)\,\mathrm{d}t &= P\Big(T_i \in [t, t+\mathrm{d}t) \mid \big(\mathbb{1}\{T_j \le u\}\big)_{j\in\{1,\dots,n\},\, u<t}\Big) \\ &= \mathbb{1}\{T_i \ge t\} \cdot P\big(T_i \in [t, t+\mathrm{d}t) \mid T_i \ge t\big)\end{aligned}$$

by the independence of the survival times, and

$$\begin{aligned}\lambda_i^{\mathscr{G}}(t)\,\mathrm{d}t &= P\Big(T_i \in [t, t+\mathrm{d}t) \mid \big(\mathbb{1}\{T_j \le u\}, \mathbb{1}\{C_j \ge u\}\big)_{j\in\{1,\dots,n\},\, u<t}\Big) \\ &= \mathbb{1}\{T_i \ge t\} \\ &\quad \cdot P\Big(T_i \in [t, t+\mathrm{d}t) \mid T_i \ge t, \big(\mathbb{1}\{T_j \le u\}, \mathbb{1}\{C_k \ge u\}\big)_{\substack{j\in\{1,\dots,n\}\setminus\{i\},\\ k\in\{1,\dots,n\},\, u<t}}\Big).\end{aligned}$$

(We omit any baseline information possibly included in the filtrations.)

Let w.l.o.g. $i = 1$. Using the calendar times R_i and Q_i to represent T_i and C_i (without conditioning on them individually), we need to prove that

$$\begin{aligned}&P\big(R_1 - Q_1 \in [t, t+\mathrm{d}t) \mid R_1 - Q_1 \ge t\big) \\ &= P\big(R_1 - Q_1 \in [t, t+\mathrm{d}t) \mid R_1 - Q_1 \ge t, (*)\big),\end{aligned}$$

where

$$(*) = \big(\mathbb{1}\{R_j - Q_j \le u\}, \mathbb{1}\{R_{(m)} - Q_k \ge u\}\big)_{\substack{j\in\{2,\dots,n\},\\ k\in\{1,\dots,n\},\, u<t}}.$$

Hence, we consider the components of $(*)$ and demonstrate that they do not imply $R_1 - Q_1 > t$.

Case 1: $R_{(m)} = R_1$
If $R_{(m)} = R_1$, then the indicator $\mathbb{1}\{R_{(m)} - Q_1 \ge u\}$ in $(*)$ equals $\mathbb{1}\{R_1 - Q_1 \ge u\}$, which is already covered by the condition $R_1 - Q_1 \ge t$ in $\lambda_1^{\mathscr{G}}(t)$, with $t > u$. We may therefore restrict our considerations to the set

$$(*) = \big(\mathbb{1}\{R_j - Q_j \le u\}, \mathbb{1}\{R_1 - Q_j \ge u\}\big)_{j\in\{2,\dots,n\},\, u<t}.$$

Besides, it holds that

$$\begin{aligned} & R_1 - Q_j \geq u \\ \iff \quad & R_1 - Q_1 \geq u + Q_j - Q_1. \end{aligned}$$

Supposing there was a $j \in \{2, \dots, n\}$ and some $u < t$ such that $R_1 - Q_j \geq u$ as well as $u + Q_j - Q_1 > t$, one could conclude that $R_1 - Q_1 > t$, i.e. $\lambda_1^{\mathscr{G}}(t)\,\mathrm{d}t = 0$. The calendar times Q_1 and Q_j are not given, however, and the value of the difference $Q_j - Q_1$ can merely be deduced from the study time variables T_k and C_k, $k \in \{1, \dots, n\}$. We might consider $Q_j - Q_1 = (R_1 - Q_1) - (R_1 - Q_j) = T_1 - C_j$, so that

$$\begin{aligned} & u + Q_j - Q_1 > t \\ \iff \quad & T_1 > t + C_j - u, \end{aligned}$$

but the inequality in the second line is only verifiable if a time point $v > t + C_j - u$ is observed such that $T_1 \geq v$. We have $v > t + R_1 - Q_j - u \geq t$ due to the assumption $R_1 - Q_j \geq u$, though, meaning that v is no element of the past until t. In summary, the knowledge of the censoring times does not affect the intensity $\lambda_1^{\mathscr{G}}$. The expression $(*)$ can thus be further reduced to

$$(*) = \Big(\mathbb{1}\{R_j - Q_j \leq u\}\Big)_{j \in \{2,\dots,n\},\, u<t}.$$

Because of the independence of the survival times, conditioning on this set has no impact on $\lambda_1^{\mathscr{G}}$ either, and we conclude that $\lambda_1^{\mathscr{G}}(t) = \lambda_1^{\mathscr{F}^c}(t)$ in case that $R_{(m)} = R_1$.

Case 2: $R_{(m)} \neq R_1$

The censoring times $C_k = R_{(m)} - Q_k$ do not add any information on $T_1 = R_1 - Q_1$ for any $k \in \{2, \dots, n\}$ if $R_{(m)} \neq R_1$, so that our interest is on

$$(*) = \Big(\mathbb{1}\{R_j - Q_j \leq u\}, \mathbb{1}\{R_{(m)} - Q_1\}\Big)_{j \in \{2,\dots,n\},\, u<t}.$$

If there was a time point $u < t$ with both $R_{(m)} - Q_1 \geq u$ as well as $u + R_1 - R_{(m)} > t$, we would be able to infer that $R_1 - Q_1 > t$, since

$$\begin{aligned} & R_{(m)} - Q_1 \geq u \\ \iff \quad & R_1 - Q_1 \geq u + R_1 - R_{(m)}. \end{aligned}$$

The study time data available for the analysis merely conveys information on $R_1 - R_{(m)}$ via $(R_1 - Q_1) - (R_{(m)} - Q_1) = T_1 - C_1$, however, which yields the inequality

$$
\begin{aligned}
& u + R_1 - R_{(m)} > t \\
\iff \qquad & T_1 > t + C_1 - u.
\end{aligned}
$$

Similarly as before, one would have to observe a time point $v > t + C_1 - u$ with $T_1 \geq v$ to confirm that $T_1 > t + C_1 - u$. But $v > t + R_{(m)} - Q_1 - u \geq t$ by assumption, and thus, the observed past does not include v. The set we need to investigate hence reduces to

$$
(*) = \big(\mathbb{1}\{R_j - Q_j \leq u\}\big)_{j \in \{2,\ldots,n\},\, u<t},
$$

and the independence of the survival times can be exploited once again to conclude that $\lambda_1^{\mathscr{G}}(t) = \lambda_1^{\mathscr{F}^c}(t)$ also if $R_{(m)} \neq R_1$. □

A.2. Approximation of the distribution of $\big(U_n(t)\big)$ by the WBS

Proof of Lemma 4.2:

The counting processes N_{ki}, $k \in \{1,\ldots,K\}$, $i \in \{1,\ldots,n\}$, jump at most once, so

$$
\begin{aligned}
& \max_{i \in \{1,\ldots,n\}} \left| \int_0^{t_r} \widehat{\widetilde{H}}_{k1}(u, t_r)\, \frac{\mathrm{d}N_{ki}(u)}{\sqrt{n}\, S^{(0)}\big(\hat{\boldsymbol{\beta}}_k, u\big)} \right| \\
< & \Big(\exp\big(\hat{\beta}_{kA}\big) + 1\Big) \max_{i \in \{1,\ldots,n\}} \exp\big(\hat{\boldsymbol{\beta}}_{kL}^T \mathbf{Z}_{Li}\big)\, \frac{1}{\sqrt{n}\, \inf_{u \in [0,t_r]} S^{(0)}\big(\hat{\boldsymbol{\beta}}_k, u\big)}.
\end{aligned}
$$

Bear in mind that on $\mathcal{B}_k \times [0,\tau]$, the function $S^{(0)}$ converges uniformly to $s^{(0)}$ (where $\mathcal{B}_k$ is a compact neighbourhood of $\boldsymbol{\beta}_{0k}$). Since $s^{(0)}$ is bounded away from zero, and $\hat{\boldsymbol{\beta}}_k$ is strongly consistent, one obtains that the maximum above almost surely converges to zero for every $r \in \{1,\ldots,l\}$.

We further have

$$
\begin{aligned}
& \max_{i \in \{1,\ldots,n\}} \left| \int_0^{\tau} \frac{1}{\sqrt{n}} \big(\widehat{\widetilde{\mathbf{H}}}_{k2}(t_r)\big)^T\, \widehat{\boldsymbol{\Sigma}}_k^{-1} \Big(\mathbf{Z}_i - \mathbf{E}\big(\hat{\boldsymbol{\beta}}_k, u\big)\Big)\, \mathrm{d}N_{ki}(u) \right| \\
\leq \frac{1}{\sqrt{n}} \max_{i \in \{1,\ldots,n\}} \Bigg\{ & \frac{1}{n} \sum_{j_1=1}^{n} \int_0^{t_r} \Bigg(\left| \Big(\big(1, \mathbf{Z}_{j_1}^T\big) - \Big(\mathbf{E}\big(\hat{\boldsymbol{\beta}}_k, v\big)\Big)^T \Big)\, \widehat{\boldsymbol{\Sigma}}_k^{-1} \int_0^{\tau} \Big(\mathbf{Z}_i - \mathbf{E}\big(\hat{\boldsymbol{\beta}}_k, u\big)\Big) \right. \\
& \qquad \left. \cdot\, \mathrm{d}N_{ki}(u) \right| \exp\big(\hat{\beta}_{kA}\big) \\
& + \left| \Big(\big(0, \mathbf{Z}_{j_1}^T\big) - \Big(\mathbf{E}\big(\hat{\boldsymbol{\beta}}_k, v\big)\Big)^T \Big)\, \widehat{\boldsymbol{\Sigma}}_k^{-1} \int_0^{\tau} \Big(\mathbf{Z}_i - \mathbf{E}\big(\hat{\boldsymbol{\beta}}_k, u\big)\Big) \right. \\
& \qquad \left. \cdot\, \mathrm{d}N_{ki}(u) \right| \Bigg) \\
& \cdot \exp\big(\hat{\boldsymbol{\beta}}_{kL}^T \mathbf{Z}_{Lj_1}\big)\, \frac{\mathrm{d}N_k(v)}{n\, S^{(0)}\big(\hat{\boldsymbol{\beta}}_k, v\big)} \Bigg\}
\end{aligned}
$$

$$< \frac{1}{\sqrt{n}} \Bigg(\max_{i, j_1 \in \{1,\ldots,n\}} \sup_{u \in [0,\tau],\, v \in [0,t_r]} \left| \left(\left(1, \mathbf{Z}_{j_1}^T\right) - \left(\mathbf{E}\left(\hat{\boldsymbol{\beta}}_k, v\right) \right)^T \right) \widehat{\boldsymbol{\Sigma}}_k^{-1} \left(\mathbf{Z}_i - \mathbf{E}\left(\hat{\boldsymbol{\beta}}_k, u\right) \right) \right|$$
$$+ \max_{i, j_1 \in \{1,\ldots,n\}} \sup_{u \in [0,\tau],\, v \in [0,t_r]} \left| \left(\left(0, \mathbf{Z}_{j_1}^T\right) - \left(\mathbf{E}\left(\hat{\boldsymbol{\beta}}_k, v\right) \right)^T \right) \widehat{\boldsymbol{\Sigma}}_k^{-1} \left(\mathbf{Z}_i - \mathbf{E}\left(\hat{\boldsymbol{\beta}}_k, u\right) \right) \right| \Bigg)$$
$$\cdot \frac{\max\{\exp\left(\hat{\beta}_{kA}\right), 1\} \max_{j_1 \in \{1,\ldots,n\}} \exp\left(\hat{\boldsymbol{\beta}}_{kL}^T \mathbf{Z}_{Lj_1}\right)}{\inf_{v \in [0,t_r]} S^{(0)}\left(\hat{\boldsymbol{\beta}}_k, v\right)}.$$

The previous considerations and the boundedness of $\mathbf{s}^{(1)}$ as well as $\mathbf{s}^{(2)}$ on $\mathcal{B}_k \times [0,\tau]$ imply that the expression above also vanishes for every $r \in \{1,\ldots,l\}$. This proves Condition (i).

To show that Condition (ii) is fulfilled as well, we consider the time points $0 \leq t_r \leq t_s \leq \tau$. The process N_{ki} jumps only once, and thus,

$$\begin{aligned}
&\sum_{i=1}^n X_{n,i}^{(k)}(t_r)\, X_{n,i}^{(k)}(t_s) \\
&= \frac{1}{n} \sum_{i=1}^n \int_0^{t_r} \widehat{\widetilde{H}}_{k1}(u,t_r)\, \widehat{\widetilde{H}}_{k1}(u,t_s)\, \frac{\mathrm{d}N_{ki}(u)}{\left(S^{(0)}\left(\hat{\boldsymbol{\beta}}_k, u\right)\right)^2} \\
&+ \frac{1}{n} \sum_{i=1}^n \int_0^{t_r} \widehat{\widetilde{H}}_{k1}(u,t_r) \left(\widehat{\widetilde{\mathbf{H}}}_{k2}(t_s)\right)^T \widehat{\boldsymbol{\Sigma}}_k^{-1} \left(\mathbf{Z}_i - \mathbf{E}\left(\hat{\boldsymbol{\beta}}_k, u\right)\right) \frac{\mathrm{d}N_{ki}(u)}{S^{(0)}\left(\hat{\boldsymbol{\beta}}_k, u\right)} \\
&+ \frac{1}{n} \sum_{i=1}^n \int_0^{t_s} \left(\widehat{\widetilde{\mathbf{H}}}_{k2}(t_r)\right)^T \widehat{\boldsymbol{\Sigma}}_k^{-1} \left(\mathbf{Z}_i - \mathbf{E}\left(\hat{\boldsymbol{\beta}}_k, u\right)\right) \widehat{\widetilde{H}}_{k1}(u,t_s)\, \frac{\mathrm{d}N_{ki}(u)}{S^{(0)}\left(\hat{\boldsymbol{\beta}}_k, u\right)} \qquad (1) \\
&+ \frac{1}{n} \sum_{i=1}^n \int_0^{\tau} \left(\widehat{\widetilde{\mathbf{H}}}_{k2}(t_r)\right)^T \widehat{\boldsymbol{\Sigma}}_k^{-1} \left(\mathbf{Z}_i - \mathbf{E}\left(\hat{\boldsymbol{\beta}}_k, u\right)\right) \left(\widehat{\widetilde{\mathbf{H}}}_{k2}(t_s)\right)^T \widehat{\boldsymbol{\Sigma}}_k^{-1} \left(\mathbf{Z}_i - \mathbf{E}\left(\hat{\boldsymbol{\beta}}_k, u\right)\right) \\
&\qquad \cdot \mathrm{d}N_{ki}(u).
\end{aligned}$$

Due to the Doob-Meyer decomposition, the first summand in Expression (1) is equal to

$$\frac{1}{n} \sum_{i=1}^n \int_0^{t_r} \widehat{\widetilde{H}}_{k1}(u,t_r)\, \widehat{\widetilde{H}}_{k1}(u,t_s)\, \frac{\mathrm{d}M_{ki}(u)}{\left(S^{(0)}\left(\hat{\boldsymbol{\beta}}_k, u\right)\right)^2} + \int_0^{t_r} \widehat{\widetilde{H}}_{k1}(u,t_r)\, \widehat{\widetilde{H}}_{k1}(u,t_s)\, \frac{\mathrm{d}A_{0k}(u)}{S^{(0)}\left(\hat{\boldsymbol{\beta}}_k, u\right)}.$$

We may apply Theorem 2.1 to infer that the left-hand side above converges to zero as $n \to \infty$ because of the uniform consistency of $\widehat{\widetilde{H}}_{k1}$ and $S^{(0)}$ as well as the boundedness of $\widetilde{H}_{k1}$ and $s^{(0)}$.

The use of similar arguments on the remaining summands in (1) yields that the sum

$$\sum_{i=1}^n X_{n,i}^{(k)}(t_r)\, X_{n,i}^{(k)}(t_s)$$

converges in probability to

$$\int_0^{t_r} \tilde{h}_{k1}(u,t_r)\,\tilde{h}_{k1}(u,t_s)\,\frac{\mathrm{d}A_{0k}(u)}{s^{(0)}\big(\boldsymbol{\beta}_{0k},u\big)}$$
$$+\int_0^{t_r} \tilde{h}_{k1}(u,t_r)\left(\tilde{\mathbf{h}}_{k2}(t_s)\right)^T\boldsymbol{\Sigma}_k^{-1}\left(\mathbf{s}^{(1)}(\boldsymbol{\beta}_{0k},u)-\mathbf{e}(\boldsymbol{\beta}_{0k},u)\,s^{(0)}(\boldsymbol{\beta}_{0k},u)\right)\frac{\mathrm{d}A_{0k}(u)}{s^{(0)}(\boldsymbol{\beta}_{0k},u)}$$
$$+\int_0^{t_s} \left(\tilde{\mathbf{h}}_{k2}(t_r)\right)^T\boldsymbol{\Sigma}_k^{-1}\left(\mathbf{s}^{(1)}(\boldsymbol{\beta}_{0k},u)-\mathbf{e}(\boldsymbol{\beta}_{0k},u)\,s^{(0)}(\boldsymbol{\beta}_{0k},u)\right)\tilde{h}_{k1}(u,t_s)\,\frac{\mathrm{d}A_{0k}(u)}{s^{(0)}(\boldsymbol{\beta}_{0k},u)}$$
$$+\left(\tilde{\mathbf{h}}_{k2}(t_r)\right)^T\boldsymbol{\Sigma}_k^{-1}\left(\int_0^{\tau}\left(\frac{\mathbf{s}^{(2)}(\boldsymbol{\beta}_{0k},u)}{s^{(0)}(\boldsymbol{\beta}_{0k},u)}-\left(\mathbf{e}(\boldsymbol{\beta}_{0k},u)\right)^{\otimes 2}\right)s^{(0)}(\boldsymbol{\beta}_{0k},u)\;\mathrm{d}A_{0k}(u)\right)$$
$$\cdot\left(\boldsymbol{\Sigma}_k^{-1}\right)^T\tilde{\mathbf{h}}_{k2}(t_s)\,,$$

which equals $\xi^{(k)}(t_r,t_s)$. Hence, Condition (ii) is proven. □

Proof of Lemma 4.3:
It holds that

$$\sqrt{n}\max_{i\in\{1,\ldots,n\}}\left|X_{n,i}^{(k)}(t_s)-X_{n,i}^{(k)}(t_r)\right|$$
$$\leq \max_{i\in\{1,\ldots,n\}}\left\{\int_0^{t_s}\left|\widehat{\widetilde{H}}_{k1}(u,t_s)-\mathbb{1}\{u\leq t_r\}\widehat{\widetilde{H}}_{k1}(u,t_r)\right|\frac{\mathrm{d}N_{ki}(u)}{S^{(0)}\big(\hat{\boldsymbol{\beta}}_k,u\big)}\right.$$
$$\left.+\int_0^{\tau}\left|\left(\widehat{\widetilde{\mathbf{H}}}_{k2}(t_s)-\widehat{\widetilde{\mathbf{H}}}_{k2}(t_r)\right)^T\widehat{\boldsymbol{\Sigma}}_k^{-1}\left(\mathbf{Z}_i-\mathbf{E}\big(\hat{\boldsymbol{\beta}}_k,u\big)\right)\right|\mathrm{d}N_{ki}(u)\right\}$$
$$<\frac{2\left(\exp\left(\hat{\beta}_{kA}\right)+1\right)\max_{i\in\{1,\ldots,n\}}\exp\big(\hat{\boldsymbol{\beta}}_{kL}^T\mathbf{Z}_{Li}\big)}{\inf_{u\in[0,\tau]}S^{(0)}\big(\hat{\boldsymbol{\beta}}_k,u\big)}$$
$$+\max_{i\in\{1,\ldots,n\}}\sup_{u,t_r,t_s\,\in\,[0,\tau]}\left|\left(\widehat{\widetilde{\mathbf{H}}}_{k2}(t_s)-\widehat{\widetilde{\mathbf{H}}}_{k2}(t_r)\right)^T\widehat{\boldsymbol{\Sigma}}_k^{-1}\left(\mathbf{Z}_i-\mathbf{E}\big(\hat{\boldsymbol{\beta}}_k,u\big)\right)\right|,$$

i.e. $\sqrt{n}\,\max_{i\in\{1,\ldots,n\}}\left|X_{n,i}^{(k)}(t_s)-X_{n,i}^{(k)}(t_r)\right|\in O_P(1)$ by the (uniform) consistency and the boundedness of the involved terms. □

Proof of Lemma 4.4:
We first split the expectation in Lemma 4.4 into terms with identical sum indices. Due to Condition (iv) in Theorem 4.4, it follows that an upper bound is given by

$$\max_{j\in\{1,\ldots,n\}}\mathbb{E}\big(G_j^4\mid\mathscr{F}_\tau\big)\sum_{i=1}^{n}\left(X_{n,i}^{(k)}(t_r)-X_{n,i}^{(k)}\big(t_q\big)\right)^2\left(X_{n,i}^{(k)}(t_s)-X_{n,i}^{(k)}(t_r)\right)^2$$
$$+2\max_{j_1\in\{1,\ldots,n\}}\left|\mathbb{E}\big(G_{j_1}^3\mid\mathscr{F}_\tau\big)\right|\max_{j_2\in\{1,\ldots,n\}}\left|\mathbb{E}\big(G_{j_2}\mid\mathscr{F}_\tau\big)\right|\sum_{i_1=1}^{n}\left(X_{n,i_1}^{(k)}(t_r)-X_{n,i_1}^{(k)}\big(t_q\big)\right)^2$$
$$\cdot\left|X_{n,i_1}^{(k)}(t_s)-X_{n,i_1}^{(k)}(t_r)\right|$$
$$\cdot\sum_{i_2=1}^{n}\left|X_{n,i_2}^{(k)}(t_s)-X_{n,i_2}^{(k)}(t_r)\right|$$

$$
\begin{aligned}
&+2 \max_{j_1\in\{1,\dots,n\}} \left|\mathbb{E}\big(G_{j_1} \mid \mathscr{F}_\tau\big)\right| \max_{j_2\in\{1,\dots,n\}} \left|\mathbb{E}\big(G^3_{j_2} \mid \mathscr{F}_\tau\big)\right| \sum_{i_1=1}^{n} \left|X^{(k)}_{n,i_1}(t_r) - X^{(k)}_{n,i_1}\big(t_q\big)\right| \\
&\qquad\qquad \cdot \sum_{i_2=1}^{n} \left|X^{(k)}_{n,i_1}(t_r) - X^{(k)}_{n,i_1}\big(t_q\big)\right| \\
&\qquad\qquad\qquad \cdot \Big(X^{(k)}_{n,i_2}(t_s) - X^{(k)}_{n,i_2}(t_r)\Big)^2 \\
&+ \max_{j\in\{1,\dots,n\}} \Big(\mathbb{E}\big(G^2_j \mid \mathscr{F}_\tau\big)\Big)^2 \sum_{i_1=1}^{n} \Big(X^{(k)}_{n,i_1}(t_r) - X^{(k)}_{n,i_1}\big(t_q\big)\Big)^2 \sum_{i_2=1}^{n} \Big(X^{(k)}_{n,i_2}(t_s) - X^{(k)}_{n,i_2}(t_r)\Big)^2 \\
&\quad +2 \max_{j\in\{1,\dots,n\}} \Big(\mathbb{E}\big(G^2_j \mid \mathscr{F}_\tau\big)\Big)^2 \Big(\sum_{i=1}^{n} \left|X^{(k)}_{n,i}(t_r) - X^{(k)}_{n,i}\big(t_q\big)\right| \left|X^{(k)}_{n,i}(t_s) - X^{(k)}_{n,i}(t_r)\right|\Big)^2 \\
&+ \max_{j_1\in\{1,\dots,n\}} \mathbb{E}\big(G^2_{j_1} \mid \mathscr{F}_\tau\big) \max_{j_2\in\{1,\dots,n\}} \Big(\mathbb{E}\big(G_{j_2} \mid \mathscr{F}_\tau\big)\Big)^2 \sum_{i_1=1}^{n} \Big(X^{(k)}_{n,i_1}(t_r) - X^{(k)}_{n,i_1}\big(t_q\big)\Big)^2 \\
&\qquad\qquad \cdot \Big(\sum_{i_2=1}^{n} \left|X^{(k)}_{n,i_2}(t_s) - X^{(k)}_{n,i_2}(t_r)\right|\Big)^2 \\
&\quad +4 \max_{j\in\{1,\dots,n\}} \Big(\mathbb{E}\big(G_j \mid \mathscr{F}_\tau\big)\Big)^2 \max_{j_2\in\{1,\dots,n\}} \mathbb{E}\big(G^2_{j_2} \mid \mathscr{F}_\tau\big) \sum_{i_1=1}^{n} \left|X^{(k)}_{n,i_1}(t_r) - X^{(k)}_{n,i_1}\big(t_q\big)\right| \\
&\qquad\qquad \cdot \sum_{i_2=1}^{n} \left|X^{(k)}_{n,i_2}(t_r) - X^{(k)}_{n,i_2}\big(t_q\big)\right| \\
&\qquad\qquad\qquad \cdot \left|X^{(k)}_{n,i_2}(t_s) - X^{(k)}_{n,i_2}(t_r)\right| \\
&\qquad\qquad \cdot \sum_{i_3=1}^{n} \left|X^{(k)}_{n,i_3}(t_s) - X^{(k)}_{n,i_3}(t_r)\right| \\
&\quad + \max_{j_1\in\{1,\dots,n\}} \Big(\mathbb{E}\big(G_{j_1} \mid \mathscr{F}_\tau\big)\Big)^2 \max_{j_2\in\{1,\dots,n\}} \mathbb{E}\big(G^2_{j_2} \mid \mathscr{F}_\tau\big) \Big(\sum_{i_1=1}^{n} \left|X^{(k)}_{n,i_1}(t_r) - X^{(k)}_{n,i_1}\big(t_q\big)\right|\Big)^2 \\
&\qquad\qquad \cdot \sum_{i_2=1}^{n} \Big(X^{(k)}_{n,i_2}(t_s) - X^{(k)}_{n,i_2}(t_r)\Big)^2 \\
&+ \max_{j\in\{1,\dots,n\}} \Big(\mathbb{E}\big(G_j \mid \mathscr{F}_\tau\big)\Big)^4 \Big(\sum_{i_1=1}^{n} \left|X^{(k)}_{n,i_1}(t_r) - X^{(k)}_{n,i_1}\big(t_q\big)\right|\Big)^2 \Big(\sum_{i_2=1}^{n} \left|X^{(k)}_{n,i_2}(t_s) - X^{(k)}_{n,i_2}(t_r)\right|\Big)^2.
\end{aligned}
$$

Let this bound be denoted by (2).

An informal representation of the first summand in (2) results from the (uniform) consistency and boundedness of the involved terms:

$$
\begin{aligned}
&\max_{j\in\{1,\dots,n\}} \mathbb{E}\big(G^4_j \mid \mathscr{F}_\tau\big)\, \frac{1}{n^2} \sum_{i=1}^{n} \Big(\int_0^{t_r} \mathrm{d}N_{ki}(u)\cdot O_P(1) + \int_0^{\tau} \mathrm{d}N_{ki}(u)\cdot O_P(1)\Big)^2 \\
&\qquad\qquad \cdot \Big(\int_0^{t_s} \mathrm{d}N_{ki}(u)\cdot O_P(1) + \int_0^{\tau} \mathrm{d}N_{ki}(u)\cdot O_P(1)\Big)^2 \\
&\qquad = \max_{j\in\{1,\dots,n\}} \mathbb{E}\big(G^4_j \mid \mathscr{F}_\tau\big)\, \frac{1}{n}\cdot O_P(1).
\end{aligned}
$$

The equality in the last line holds because N_{ki} is a one-jump process. By means of Condition (iii) in Theorem 4.4, we infer that the first summand of Expression (2) can be neglected if $n \to \infty$.

Next, note that $\sum_{i=1}^{n} \left| X_{n,i}^{(k)}(t_s) - X_{n,i}^{(k)}(t_r) \right| \leq \sqrt{n} \left(\sum_{i=1}^{n} \left(X_{n,i}^{(k)}(t_s) - X_{n,i}^{(k)}(t_r) \right)^2 \right)^{1/2}$ according to Hölder's inequality. It follows that the second summand in (2) is bounded by

$$
\begin{aligned}
&2\sqrt{n} \max_{j_1 \in \{1,\ldots,n\}} \left| \mathbb{E}\big(G_{j_1}^3 \mid \mathscr{F}_\tau\big) \right| \max_{j_2 \in \{1,\ldots,n\}} \left| \mathbb{E}\big(G_{j_2} \mid \mathscr{F}_\tau\big) \right| \\
&\quad \cdot \max_{(t_0,t_p) \in \{(t_q,t_r),(t_r,t_s)\}} \left(\sum_{i=1}^{n} \left(X_{n,i}^{(k)}\big(t_p\big) - X_{n,i}^{(k)}(t_o) \right)^2 \right)^{3/2} \max_{i \in \{1,\ldots,n\}} \left| X_{n,i}^{(k)}(t_s) - X_{n,i}^{(k)}(t_r) \right|.
\end{aligned}
$$

(We used here that the function $f(x) = x^{3/2}$ increases monotonically in x, so that the maximum w.r.t. the time points $\big(t_o, t_p\big)$ can be placed outside of the brackets.) The application of Jensen's inequality to Condition (iii) (considering the convex function $f(x) = x^{4/3}$), together with Condition (i) as well as Lemma 4.3, eventually yields

$$
\max_{(t_0,t_p) \in \{(t_q,t_r),(t_r,t_s)\}} \left(\sum_{i=1}^{n} \left(X_{n,i}^{(k)}\big(t_p\big) - X_{n,i}^{(k)}(t_o) \right)^2 \right)^{3/2} \cdot O_P(1) \tag{3}
$$

as an upper bound of the second summand in (2). Keep in mind that the $O_P(1)$-term does not depend on the time points t_q, t_r, t_s (cf. Lemma 4.3)!

It turns out that similar considerations (involving Hölder's inequality, Lemma 4.3, as well as Conditions (i) and (ii) of the Theorem) apply to each of the summands in (2). Thus, Expression (3) forms a general upper bound for the expectation in Lemma 4.4.

The proof is complete if we find a function $L_n^{(k)}$ such that

$$
\sum_{i=1}^{n} \left(X_{n,i}^{(k)}\big(t_p\big) - X_{n,i}^{(k)}(t_o) \right)^2 \leq \left(L_n^{(k)}(t_s) - L_n^{(k)}\big(t_q\big) \right) \cdot O_P(1)
$$

for $\big(t_o, t_p\big) \in \{\big(t_q, t_r\big), (t_r, t_s\}$.

In this regard, note that $(a+b)^2 \leq 2a^2 + 2b^2$ for $a, b \in \mathbb{R}$ and therefore,

$$
\begin{aligned}
&\sum_{i=1}^{n} \left(X_{n,i}^{(k)}\big(t_p\big) - X_{n,i}^{(k)}(t_o) \right)^2 \\
&\leq \frac{2}{n} \sum_{i=1}^{n} \left(\left(\int_0^{t_p} \widehat{\widetilde{H}}_{k1}\big(u, t_p\big) \, \frac{\mathrm{d}N_{ki}(u)}{S^{(0)}\big(\hat{\boldsymbol{\beta}}_k, u\big)} - \int_0^{t_0} \widehat{\widetilde{H}}_{k1}(u, t_o) \, \frac{\mathrm{d}N_{ki}(u)}{S^{(0)}\big(\hat{\boldsymbol{\beta}}_k, u\big)} \right)^2 \right. \\
&\qquad \left. + \left(\int_0^{\tau} \left(\widehat{\widetilde{\mathbf{H}}}_{k2}\big(t_p\big) - \widehat{\widetilde{\mathbf{H}}}_{k2}(t_o) \right)^T \widehat{\boldsymbol{\Sigma}}_k^{-1} \left(\mathbf{Z}_i - \mathbf{E}\big(\hat{\boldsymbol{\beta}}_k, u\big) \right) \mathrm{d}N_{ki}(u) \right)^2 \right).
\end{aligned} \tag{4}
$$

The definition of $\widehat{\widetilde{H}}_{k1}$ implies that the first line of Expression (4) is bounded by

$$\frac{2}{n}\sum_{i=1}^{n}\Bigg(\frac{1}{n}\sum_{j=1}^{n}\Big(3\exp\big(\hat{\beta}_{kA}\big)+3\Big)\exp\big(\hat{\boldsymbol{\beta}}_{kL}^{T}\mathbf{Z}_{Lj}\big)\int_{t_o}^{t_p}\frac{\mathrm{d}N_{ki}(u)}{S^{(0)}\big(\hat{\boldsymbol{\beta}}_k,u\big)} + \frac{1}{n}\sum_{j=1}^{n}\Bigg(\Big(\hat{F}_1\big(t_p\mid Z_A{=}1,\mathbf{Z}_{Lj}\big)-\hat{F}_1\big(t_o\mid Z_A{=}1,\mathbf{Z}_{Lj}\big)\Big)\exp\big(\hat{\beta}_{kA}\big) + \Big(\hat{F}_1\big(t_p\mid Z_A{=}0,\mathbf{Z}_{Lj}\big)-\hat{F}_1\big(t_o\mid Z_A{=}0,\mathbf{Z}_{Lj}\big)\Big)\Bigg) \cdot\exp\big(\hat{\boldsymbol{\beta}}_{kL}^{T}\mathbf{Z}_{Lj}\big)\int_{0}^{t_o}\frac{\mathrm{d}N_{ki}(u)}{S^{(0)}\big(\hat{\boldsymbol{\beta}}_k,u\big)}\Bigg)^2.$$

Using again that $(a+b)^2\leq 2a^2+2b^2$ as well as the Cauchy-Schwarz inequality and the fact that N_{ki} jumps only once, we observe that the term above is smaller or equal to

$$\begin{aligned}
&\frac{2}{n}\sum_{i=1}^{n}\Bigg(\frac{2}{n}\sum_{j=1}^{n}\Big(9\exp\big(2\hat{\beta}_{kA}\big)+18\exp\big(\hat{\beta}_{kA}\big)+9\Big)\exp\big(2\hat{\boldsymbol{\beta}}_{kL}^{T}\mathbf{Z}_{Lj}\big)\int_{t_o}^{t_p}\frac{\mathrm{d}N_{ki}(u)}{\Big(S^{(0)}\big(\hat{\boldsymbol{\beta}}_k,u\big)\Big)^2}\\
&\quad+\frac{2}{n}\sum_{j=1}^{n}\Bigg(2\Big(\hat{F}_1\big(t_p\mid Z_A{=}1,\mathbf{Z}_{Lj}\big)-\hat{F}_1\big(t_o\mid Z_A{=}1,\mathbf{Z}_{Lj}\big)\Big)^2\exp\big(2\hat{\beta}_{kA}\big)\\
&\qquad+2\Big(\hat{F}_1\big(t_p\mid Z_A{=}0,\mathbf{Z}_{Lj}\big)-\hat{F}_1\big(t_o\mid Z_A{=}0,\mathbf{Z}_{Lj}\big)\Big)^2\Bigg)\\
&\qquad\cdot\exp\big(2\hat{\boldsymbol{\beta}}_{kL}^{T}\mathbf{Z}_{Lj}\big)\int_{0}^{\tau}\frac{\mathrm{d}N_{ki}(u)}{\Big(S^{(0)}\big(\hat{\boldsymbol{\beta}}_k,u\big)\Big)^2}\Bigg)\\
\leq\ &\frac{2}{n}\sum_{i=1}^{n}\Bigg(\frac{2}{n}\sum_{j=1}^{n}\Big(9\exp\big(2\hat{\beta}_{kA}\big)+18\exp\big(\hat{\beta}_{kA}\big)+9\Big)\exp\big(2\hat{\boldsymbol{\beta}}_{kL}^{T}\mathbf{Z}_{Lj}\big)\int_{t_o}^{t_p}\frac{\mathrm{d}N_{ki}(u)}{\Big(S^{(0)}\big(\hat{\boldsymbol{\beta}}_k,u\big)\Big)^2}\\
&\quad+\frac{2}{n}\sum_{j=1}^{n}\Bigg(2\Big(\Big(\hat{F}_1\big(t_p\mid Z_A{=}1,\mathbf{Z}_{Lj}\big)\Big)^2-\Big(\hat{F}_1\big(t_o\mid Z_A{=}1,\mathbf{Z}_{Lj}\big)\Big)^2\Big)\exp\big(2\hat{\beta}_{kA}\big)\\
&\qquad+2\Big(\Big(\hat{F}_1\big(t_p\mid Z_A{=}0,\mathbf{Z}_{Lj}\big)\Big)^2-\Big(\hat{F}_1\big(t_o\mid Z_A{=}0,\mathbf{Z}_{Lj}\big)\Big)^2\Big)\Bigg)\\
&\qquad\cdot\exp\big(2\hat{\boldsymbol{\beta}}_{kL}^{T}\mathbf{Z}_{Lj}\big)\int_{0}^{\tau}\frac{\mathrm{d}N_{ki}(u)}{\Big(S^{(0)}\big(\hat{\boldsymbol{\beta}}_k,u\big)\Big)^2}\Bigg).
\end{aligned}\tag{5}$$

We exploited the non-negativity of the hindmost term (since $t_o\leq t_p$) to extend the limit of the corresponding integral. Besides, the last inequality follows from $(a-b)^2\leq a^2-b^2$ for $0\leq b\leq a$.

Let us now consider the matrix

$$\widetilde{\boldsymbol{\Sigma}}_{ki}=\int_{0}^{\tau}\Big(\mathbf{Z}_i-\mathbf{E}\big(\hat{\boldsymbol{\beta}}_k,u\big)\Big)^{\otimes 2}\,\mathrm{d}N_{ki}(u).$$

Because N_{ki} is a one-jump process, the matrix $\widetilde{\boldsymbol{\Sigma}}_{ki}$ is symmetric, and real numbers are equal to their transpose, the second part of Expression (4) can be represented by

$$\frac{2}{n}\sum_{i=1}^{n}\Big(\widehat{\widehat{\mathbf{H}}}_{k2}\big(t_p\big)-\widehat{\widehat{\mathbf{H}}}_{k2}(t_o)\Big)^{T}\widehat{\boldsymbol{\Sigma}}_k^{-1}\widetilde{\boldsymbol{\Sigma}}_{ki}\widehat{\boldsymbol{\Sigma}}_k^{-1}\Big(\widehat{\widehat{\mathbf{H}}}_{k2}\big(t_p\big)-\widehat{\widehat{\mathbf{H}}}_{k2}(t_o)\Big).$$

One may further express the difference $\widehat{\widehat{\mathbf{H}}}_{k2}\big(t_p\big)-\widehat{\widehat{\mathbf{H}}}_{k2}(t_o)$ as

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}\int_{t_o}^{t_p}\Big(\widehat{\boldsymbol{\chi}}_{k1,A=1}\big(u,t_p\big)-\widehat{\boldsymbol{\chi}}_{k1,A=0}\big(u,t_p\big)\Big)\,\frac{\mathrm{d}N_{ki}(u)}{S^{(0)}\big(\hat{\boldsymbol{\beta}}_k,u\big)}\\&-\frac{1}{n}\sum_{i=1}^{n}\int_{0}^{t_o}\Big(\widehat{\boldsymbol{\chi}}_{k2,A=1}\big(u,t_o,t_p\big)-\widehat{\boldsymbol{\chi}}_{k2,A=0}\big(u,t_o,t_p\big)\Big)\,\frac{\mathrm{d}N_{ki}(u)}{S^{(0)}\big(\hat{\boldsymbol{\beta}}_k,u\big)},\end{aligned}$$

with the functions $\widehat{\boldsymbol{\chi}}_{k1,a}$ and $\widehat{\boldsymbol{\chi}}_{k2,a}$ determined by

$$\begin{aligned}\widehat{\boldsymbol{\chi}}_{k1,a}(u,t)=\frac{1}{n}\sum_{i=1}^{n}&\Big(\mathbb{1}\{k=1\}\,\hat{S}\big(u-\mid a,\mathbf{Z}_{Li}\big)-\hat{F}_1\big(t\mid a,\mathbf{Z}_{Li}\big)+\hat{F}_1\big(u\mid a,\mathbf{Z}_{Li}\big)\Big)\\&\cdot\Big(\big(a,\mathbf{Z}_{Li}^{T}\big)^{T}-\mathbf{E}\big(\hat{\boldsymbol{\beta}}_k,u\big)\Big)\exp\Big(\hat{\boldsymbol{\beta}}_k^{T}\big(a,\mathbf{Z}_{Li}^{T}\big)^{T}\Big),\\\widehat{\boldsymbol{\chi}}_{k2,a}(u,s,t)=\frac{1}{n}\sum_{i=1}^{n}&\Big(\hat{F}_1\big(t\mid a,\mathbf{Z}_{Li}\big)-\hat{F}_1\big(s\mid a,\mathbf{Z}_{Li}\big)\Big)\Big(\big(a,\mathbf{Z}_{Li}^{T}\big)^{T}-\mathbf{E}\big(\hat{\boldsymbol{\beta}}_k,u\big)\Big)\\&\cdot\exp\Big(\hat{\boldsymbol{\beta}}_k^{T}\big(a,\mathbf{Z}_{Li}^{T}\big)^{T}\Big).\end{aligned}$$

Taking into account that the matrix product $\widehat{\boldsymbol{\Sigma}}_k^{-1}\widetilde{\boldsymbol{\Sigma}}_{ki}\widehat{\boldsymbol{\Sigma}}_k^{-1}$ is positive semidefinite according to the definitions of $\widehat{\boldsymbol{\Sigma}}_k$ and $\widetilde{\boldsymbol{\Sigma}}_{ki}$, we obtain the upper bound

$$\begin{aligned}\frac{2}{n}\sum_{i=1}^{n}\Bigg(&\frac{2}{n}\int_{t_o}^{t_p}\Big(\widehat{\boldsymbol{\chi}}_{k1,A=1}\big(u,t_p\big)-\widehat{\boldsymbol{\chi}}_{k1,A=0}\big(u,t_p\big)\Big)^{T}\widehat{\boldsymbol{\Sigma}}_k^{-1}\widetilde{\boldsymbol{\Sigma}}_{ki}\widehat{\boldsymbol{\Sigma}}_k^{-1}\\&\cdot\Big(\widehat{\boldsymbol{\chi}}_{k1,A=1}\big(u,t_p\big)-\widehat{\boldsymbol{\chi}}_{k1,A=0}\big(u,t_p\big)\Big)\,\frac{\mathrm{d}N_k(u)}{\Big(S^{(0)}\big(\hat{\boldsymbol{\beta}}_k,u\big)\Big)^2}\\&+\frac{2}{n}\int_{0}^{t_o}\Big(\widehat{\boldsymbol{\chi}}_{k2,A=1}\big(u,t_o,t_p\big)-\widehat{\boldsymbol{\chi}}_{k2,A=0}\big(u,t_o,t_p\big)\Big)^{T}\widehat{\boldsymbol{\Sigma}}_k^{-1}\widetilde{\boldsymbol{\Sigma}}_{ki}\widehat{\boldsymbol{\Sigma}}_k^{-1}\\&\cdot\Big(\widehat{\boldsymbol{\chi}}_{k2,A=1}\big(u,t_o,t_p\big)-\widehat{\boldsymbol{\chi}}_{k2,A=0}\big(u,t_o,t_p\big)\Big)\,\frac{\mathrm{d}N_k(u)}{\Big(S^{(0)}\big(\hat{\boldsymbol{\beta}}_k,u\big)\Big)^2}\Bigg)\end{aligned}$$

for the second part of (4), since N_{ki} jumps at most once, and because

$$(\mathbf{a}-\mathbf{b})^{T}\mathbf{A}\,(\mathbf{a}-\mathbf{b})\;\leq\;2\mathbf{a}^{T}\mathbf{A}\,\mathbf{a}+2\mathbf{b}^{T}\mathbf{A}\,\mathbf{b}$$

as well as

$$\Big(\sum_{i_1=1}^{n}\mathbf{a}_{i_1}\Big)^{T}\mathbf{A}\Big(\sum_{i_2=1}^{n}\mathbf{a}_{i_2}\Big)\;\leq\;n\sum_{i=1}^{n}\mathbf{a}_i^{T}\mathbf{A}\,\mathbf{a}_i$$

for a positive (semi-)definite matrix $\mathbf{A}$ and suitable vectors $\mathbf{a}, \mathbf{b}, \mathbf{a}_1, \ldots, \mathbf{a}_n$. This bound may be refined even further, resulting in

$$\begin{aligned}
&\frac{2}{n}\sum_{i=1}^{n}\Bigg(\frac{4}{n^2}\int_{t_o}^{t_p}\sum_{j=1}^{n}\Big(\mathbb{1}\{k=1\}\hat{S}\big(u- \mid Z_A{=}1, \mathbf{Z}_{Lj}\big) - \hat{F}_1\big(t_p \mid Z_A{=}1, \mathbf{Z}_{Lj}\big) \\
&\qquad\qquad + \hat{F}_1\big(u \mid Z_A{=}1, \mathbf{Z}_{Lj}\big)\Big)^2 \\
&\qquad \cdot \Big(\big(1, \mathbf{Z}_{Lj}^T\big)^T - \mathbf{E}\big(\hat{\boldsymbol{\beta}}_k, u\big)\Big)^T \widehat{\boldsymbol{\Sigma}}_k^{-1}\widetilde{\boldsymbol{\Sigma}}_{ki}\widehat{\boldsymbol{\Sigma}}_k^{-1}\Big(\big(1, \mathbf{Z}_{Lj}^T\big)^T - \mathbf{E}\big(\hat{\boldsymbol{\beta}}_k, u\big)\Big) \\
&\qquad \cdot \exp\big(2\hat{\boldsymbol{\beta}}_k\big(1, \mathbf{Z}_{Lj}^T\big)^T\big)\,\frac{\mathrm{d}N_k(u)}{\Big(S^{(0)}\big(\hat{\boldsymbol{\beta}}_k, u\big)\Big)^2} \\
&+ \frac{4}{n^2}\int_{t_o}^{t_p}\sum_{j=1}^{n}\Big(\mathbb{1}\{k=1\}\hat{S}\big(u- \mid Z_A{=}0, \mathbf{Z}_{Lj}\big) - \hat{F}_1\big(t_p \mid Z_A{=}0, \mathbf{Z}_{Lj}\big) \\
&\qquad\qquad + \hat{F}_1\big(u \mid Z_A{=}0, \mathbf{Z}_{Lj}\big)\Big)^2 \\
&\qquad \cdot \Big(\big(0, \mathbf{Z}_{Lj}^T\big)^T - \mathbf{E}\big(\hat{\boldsymbol{\beta}}_k, u\big)\Big)^T \widehat{\boldsymbol{\Sigma}}_k^{-1}\widetilde{\boldsymbol{\Sigma}}_{ki}\widehat{\boldsymbol{\Sigma}}_k^{-1}\Big(\big(0, \mathbf{Z}_{Lj}^T\big)^T - \mathbf{E}\big(\hat{\boldsymbol{\beta}}_k, u\big)\Big) \\
&\qquad \cdot \exp\big(2\hat{\boldsymbol{\beta}}_k\big(0, \mathbf{Z}_{Lj}^T\big)^T\big)\,\frac{\mathrm{d}N_k(u)}{\Big(S^{(0)}\big(\hat{\boldsymbol{\beta}}_k, u\big)\Big)^2} \\
&+ \frac{4}{n^2}\int_{0}^{t_o}\sum_{j=1}^{n}\Big(\hat{F}_1\big(t_p \mid Z_A{=}1, \mathbf{Z}_{Lj}\big) - \hat{F}_1\big(t_o \mid Z_A{=}1, \mathbf{Z}_{Lj}\big)\Big)^2 \\
&\qquad \cdot \Big(\big(1, \mathbf{Z}_{Lj}^T\big)^T - \mathbf{E}\big(\hat{\boldsymbol{\beta}}_k, u\big)\Big)^T \widehat{\boldsymbol{\Sigma}}_k^{-1}\widetilde{\boldsymbol{\Sigma}}_{ki}\widehat{\boldsymbol{\Sigma}}_k^{-1}\Big(\big(1, \mathbf{Z}_{Lj}^T\big)^T - \mathbf{E}\big(\hat{\boldsymbol{\beta}}_k, u\big)\Big) \\
&\qquad \cdot \exp\big(2\hat{\boldsymbol{\beta}}_k\big(1, \mathbf{Z}_{Lj}^T\big)^T\big)\,\frac{\mathrm{d}N_k(u)}{\Big(S^{(0)}\big(\hat{\boldsymbol{\beta}}_k, u\big)\Big)^2} \\
&+ \frac{4}{n^2}\int_{0}^{t_o}\sum_{j=1}^{n}\Big(\hat{F}_1\big(t_p \mid Z_A{=}0, \mathbf{Z}_{Lj}\big) - \hat{F}_1\big(t_o \mid Z_A{=}0, \mathbf{Z}_{Lj}\big)\Big)^2 \\
&\qquad \cdot \Big(\big(0, \mathbf{Z}_{Lj}^T\big)^T - \mathbf{E}\big(\hat{\boldsymbol{\beta}}_k, u\big)\Big)^T \widehat{\boldsymbol{\Sigma}}_k^{-1}\widetilde{\boldsymbol{\Sigma}}_{ki}\widehat{\boldsymbol{\Sigma}}_k^{-1}\Big(\big(0, \mathbf{Z}_{Lj}^T\big)^T - \mathbf{E}\big(\hat{\boldsymbol{\beta}}_k, u\big)\Big) \\
&\qquad \cdot \exp\big(2\hat{\boldsymbol{\beta}}_k\big(0, \mathbf{Z}_{Lj}^T\big)^T\big)\,\frac{\mathrm{d}N_k(u)}{\Big(S^{(0)}\big(\hat{\boldsymbol{\beta}}_k, u\big)\Big)^2}\Bigg)
\end{aligned}$$

by the inequalities $(\mathbf{a}-\mathbf{b})^T\mathbf{A}\,(\mathbf{a}-\mathbf{b}) \leq 2\mathbf{a}^T\mathbf{A}\,\mathbf{a} + 2\mathbf{b}^T\mathbf{A}\,\mathbf{b}$ and $\sum_{i_1=1}^{n}\sum_{i_2=1}^{n}\mathbf{a}_{i_1}^T\mathbf{A}\,\mathbf{a}_{i_2}$ $\leq n\sum_{i=1}^{n}\mathbf{a}_i^T\mathbf{A}\mathbf{a}_i$. Lastly, we use that $(a-b)^2 \leq a^2 - b^2$ for $0 \leq b \leq a$ to see that the second part of Expression 4 is smaller or equal to

$$\begin{aligned}
&\frac{2}{n}\sum_{i=1}^{n}\Bigg(\frac{4}{n^2}\int_{t_q}^{t_s}\sum_{j=1}^{n}\Big(\big(1, \mathbf{Z}_{Lj}^T\big)^T - \mathbf{E}\big(\hat{\boldsymbol{\beta}}_k, u\big)\Big)^T \widehat{\boldsymbol{\Sigma}}_k^{-1}\widetilde{\boldsymbol{\Sigma}}_{ki}\widehat{\boldsymbol{\Sigma}}_k^{-1}\Big(\big(1, \mathbf{Z}_{Lj}^T\big)^T - \mathbf{E}\big(\hat{\boldsymbol{\beta}}_k, u\big)\Big) \\
&\qquad \cdot \exp\big(2\hat{\boldsymbol{\beta}}_k\big(1, \mathbf{Z}_{Lj}^T\big)^T\big)\,\frac{\mathrm{d}N_k(u)}{\Big(S^{(0)}\big(\hat{\boldsymbol{\beta}}_k, u\big)\Big)^2}
\end{aligned}$$

$$
\begin{aligned}
&+ \frac{4}{n^2} \int_{t_q}^{t_s} \sum_{j=1}^{n} \Big(\big(0, \mathbf{Z}_{Lj}^T\big)^T - \mathbf{E}\big(\hat{\boldsymbol{\beta}}_k, u\big) \Big)^T \widehat{\boldsymbol{\Sigma}}_k^{-1} \widetilde{\boldsymbol{\Sigma}}_{ki} \widehat{\boldsymbol{\Sigma}}_k^{-1} \Big(\big(0, \mathbf{Z}_{Lj}^T\big)^T - \mathbf{E}\big(\hat{\boldsymbol{\beta}}_k, u\big) \Big) \\
&\qquad \cdot \exp\Big(2\hat{\boldsymbol{\beta}}_k \big(0, \mathbf{Z}_{Lj}^T\big)^T\Big) \frac{\mathrm{d}N_k(u)}{\Big(S^{(0)}\big(\hat{\boldsymbol{\beta}}_k, u\big)\Big)^2} \\
&+ \frac{4}{n^2} \int_{0}^{\tau} \sum_{j=1}^{n} \Big(\Big(\hat{F}_1\big(t_s \mid Z_A = 1, \mathbf{Z}_{Lj}\big)\Big)^2 - \Big(\hat{F}_1\big(t_q \mid Z_A = 1, \mathbf{Z}_{Lj}\big)\Big)^2 \Big) \\
&\qquad \cdot \Big(\big(1, \mathbf{Z}_{Lj}^T\big)^T - \mathbf{E}\big(\hat{\boldsymbol{\beta}}_k, u\big) \Big)^T \widehat{\boldsymbol{\Sigma}}_k^{-1} \widetilde{\boldsymbol{\Sigma}}_{ki} \widehat{\boldsymbol{\Sigma}}_k^{-1} \Big(\big(1, \mathbf{Z}_{Lj}^T\big)^T - \mathbf{E}\big(\hat{\boldsymbol{\beta}}_k, u\big) \Big) \\
&\qquad \cdot \exp\Big(2\hat{\boldsymbol{\beta}}_k \big(1, \mathbf{Z}_{Lj}^T\big)^T\Big) \frac{\mathrm{d}N_k(u)}{\Big(S^{(0)}\big(\hat{\boldsymbol{\beta}}_k, u\big)\Big)^2} \\
&+ \frac{4}{n^2} \int_{0}^{\tau} \sum_{j=1}^{n} \Big(\Big(\hat{F}_1\big(t_s \mid Z_A = 0, \mathbf{Z}_{Lj}\big)\Big)^2 - \Big(\hat{F}_1\big(t_q \mid Z_A = 0, \mathbf{Z}_{Lj}\big)\Big)^2 \Big) \\
&\qquad \cdot \Big(\big(0, \mathbf{Z}_{Lj}^T\big)^T - \mathbf{E}\big(\hat{\boldsymbol{\beta}}_k, u\big) \Big)^T \widehat{\boldsymbol{\Sigma}}_k^{-1} \widetilde{\boldsymbol{\Sigma}}_{ki} \widehat{\boldsymbol{\Sigma}}_k^{-1} \Big(\big(0, \mathbf{Z}_{Lj}^T\big)^T - \mathbf{E}\big(\hat{\boldsymbol{\beta}}_k, u\big) \Big) \\
&\qquad \cdot \exp\Big(2\hat{\boldsymbol{\beta}}_k \big(0, \mathbf{Z}_{Lj}^T\big)^T\Big) \frac{\mathrm{d}N_k(u)}{\Big(S^{(0)}\big(\hat{\boldsymbol{\beta}}_k, u\big)\Big)^2} \Bigg).
\end{aligned}
$$

Bear in mind that $\widehat{\boldsymbol{\Sigma}}_k^{-1} \widetilde{\boldsymbol{\Sigma}}_{ki} \widehat{\boldsymbol{\Sigma}}_k^{-1}$ is positive semidefinite, which is why it is possible to extend the integral limits.

One may finally conclude by the bounds in Term (5) and above that

$$
\sum_{i=1}^{n} \Big(X_{n,i}^{(k)}\big(t_p\big) - X_{n,i}^{(k)}(t_o) \Big)^2 \leq 36 \Big(L_n^{(k)}(t_s) - L_n^{(k)}\big(t_q\big) \Big)^{3/2}
$$

for $k \in \{1, \ldots, K\}$, $\big(t_o, t_p\big) \in \{\big(t_q, t_r\big), (t_r, t_s)\}$, and the function

$$
\begin{aligned}
L_n^{(k)}(t) = \frac{1}{n} \sum_{j=1}^{n} \Bigg(& \Big(\exp\big(2\hat{\beta}_{kA}\big) + 2\exp\big(\hat{\beta}_{kA}\big) + 1 \Big) \exp\big(2\hat{\boldsymbol{\beta}}_{kL}^T \mathbf{Z}_{Lj}\big) \int_0^t \frac{\mathrm{d}N_k(u)}{n\Big(S^{(0)}\big(\hat{\boldsymbol{\beta}}_k, u\big)\Big)^2} \\
& + \Big(\Big(\hat{F}_1\big(t \mid Z_A = 1, \mathbf{Z}_{Lj}\big)\Big)^2 \exp\big(2\hat{\beta}_{kA}\big) + \Big(\hat{F}_1\big(t \mid Z_A = 0, \mathbf{Z}_{Lj}\big)\Big)^2 \Big) \\
& \qquad \cdot \exp\big(2\hat{\boldsymbol{\beta}}_{kL}^T \mathbf{Z}_{Lj}\big) \int_0^{\tau} \frac{\mathrm{d}N_k(u)}{n\Big(S^{(0)}\big(\hat{\boldsymbol{\beta}}_k, u\big)\Big)^2} \\
& + \int_0^t \Big(\big(1, \mathbf{Z}_{Lj}^T\big)^T - \mathbf{E}\big(\hat{\boldsymbol{\beta}}_k, u\big) \Big)^T \widehat{\boldsymbol{\Sigma}}_k^{-1} \Big(\frac{1}{n} \sum_{i=1}^{n} \widetilde{\boldsymbol{\Sigma}}_{ki} \Big) \widehat{\boldsymbol{\Sigma}}_k^{-1} \\
& \qquad \cdot \Big(\big(1, \mathbf{Z}_{Lj}^T\big)^T - \mathbf{E}\big(\hat{\boldsymbol{\beta}}_k, u\big) \Big) \\
& \qquad \cdot \exp\Big(2\hat{\boldsymbol{\beta}}_k \big(1, \mathbf{Z}_{Lj}^T\big)^T\Big) \frac{\mathrm{d}N_k(u)}{n\Big(S^{(0)}\big(\hat{\boldsymbol{\beta}}_k, u\big)\Big)^2}
\end{aligned}
$$

$$
\begin{aligned}
&+ \int_0^t \Big(\big(0, \mathbf{Z}_{Lj}^T\big)^T - \mathbf{E}\big(\hat{\boldsymbol{\beta}}_k, u\big) \Big)^T \widehat{\boldsymbol{\Sigma}}_k^{-1} \Big(\frac{1}{n} \sum_{i=1}^n \widetilde{\boldsymbol{\Sigma}}_{ki} \Big) \widehat{\boldsymbol{\Sigma}}_k^{-1} \\
&\qquad \cdot \Big(\big(0, \mathbf{Z}_{Lj}^T\big)^T - \mathbf{E}\big(\hat{\boldsymbol{\beta}}_k, u\big) \Big) \\
&\qquad \cdot \exp\Big(2\hat{\boldsymbol{\beta}}_k \big(0, \mathbf{Z}_{Lj}^T\big)^T \Big) \frac{\mathrm{d}N_k(u)}{n \Big(S^{(0)}\big(\hat{\boldsymbol{\beta}}_k, u\big) \Big)^2} \\
&+ \Big(\hat{F}_1\big(t \mid Z_A = 1, \mathbf{Z}_{Lj}\big) \Big)^2 \\
&\quad \cdot \int_0^\tau \Big(\big(1, \mathbf{Z}_{Lj}^T\big)^T - \mathbf{E}\big(\hat{\boldsymbol{\beta}}_k, u\big) \Big)^T \widehat{\boldsymbol{\Sigma}}_k^{-1} \Big(\frac{1}{n} \sum_{i=1}^n \widetilde{\boldsymbol{\Sigma}}_{ki} \Big) \widehat{\boldsymbol{\Sigma}}_k^{-1} \\
&\qquad \cdot \Big(\big(1, \mathbf{Z}_{Lj}^T\big)^T - \mathbf{E}\big(\hat{\boldsymbol{\beta}}_k, u\big) \Big) \\
&\qquad \cdot \exp\Big(2\hat{\boldsymbol{\beta}}_k \big(1, \mathbf{Z}_{Lj}^T\big)^T \Big) \frac{\mathrm{d}N_k(u)}{n \Big(S^{(0)}\big(\hat{\boldsymbol{\beta}}_k, u\big) \Big)^2} \\
&+ \Big(\hat{F}_1\big(t \mid Z_A = 0, \mathbf{Z}_{Lj}\big) \Big)^2 \\
&\quad \cdot \int_0^\tau \Big(\big(0, \mathbf{Z}_{Lj}^T\big)^T - \mathbf{E}\big(\hat{\boldsymbol{\beta}}_k, u\big) \Big)^T \widehat{\boldsymbol{\Sigma}}_k^{-1} \Big(\frac{1}{n} \sum_{i=1}^n \widetilde{\boldsymbol{\Sigma}}_{ki} \Big) \widehat{\boldsymbol{\Sigma}}_k^{-1} \\
&\qquad \cdot \Big(\big(0, \mathbf{Z}_{Lj}^T\big)^T - \mathbf{E}\big(\hat{\boldsymbol{\beta}}_k, u\big) \Big) \\
&\qquad \cdot \exp\Big(2\hat{\boldsymbol{\beta}}_k \big(0, \mathbf{Z}_{Lj}^T\big)^T \Big) \frac{\mathrm{d}N_k(u)}{n \Big(S^{(0)}\big(\hat{\boldsymbol{\beta}}_k, u\big) \Big)^2} \Bigg).
\end{aligned}
$$

Hence,

$$
\begin{aligned}
&\mathbb{E}\left(\left(\sum_{i=1}^n X_{n,i}^{(k)}(t_r) G_i - \sum_{i=1}^n X_{n,i}^{(k)}\big(t_q\big) G_i \right)^2 \left(\sum_{i=1}^n X_{n,i}^{(k)}(t_s) G_i - \sum_{i=1}^n X_{n,i}^{(k)}(t_r) G_i \right)^2 \,\Big|\, \mathscr{F}_\tau \right) \\
&\qquad \leq \Big(L_n^{(k)}(t_s) - L_n^{(k)}\big(t_q\big) \Big)^{3/2} \cdot O_P(1).
\end{aligned}
$$

□

Appendix B: Further simulation results

This chapter gathers additional results obtained from the simulations described in Chapters 3 and 4.

B.1. Impact of conditioning on calendar times in event-driven trials with staggered entry

The following figures and tables pertain to the simulation study presented in Subsection 3.2.1. Note that iterations with less than two observed events in one treatment group are excluded. The shadow lines in the shadow plots are further restricted to a random sample of size 2,000 for greater clarity.

Figure B.1: Shadow plots of the Breslow estimators in the exponential scenario with HR 1, $n = 600$, and $m = 300$.

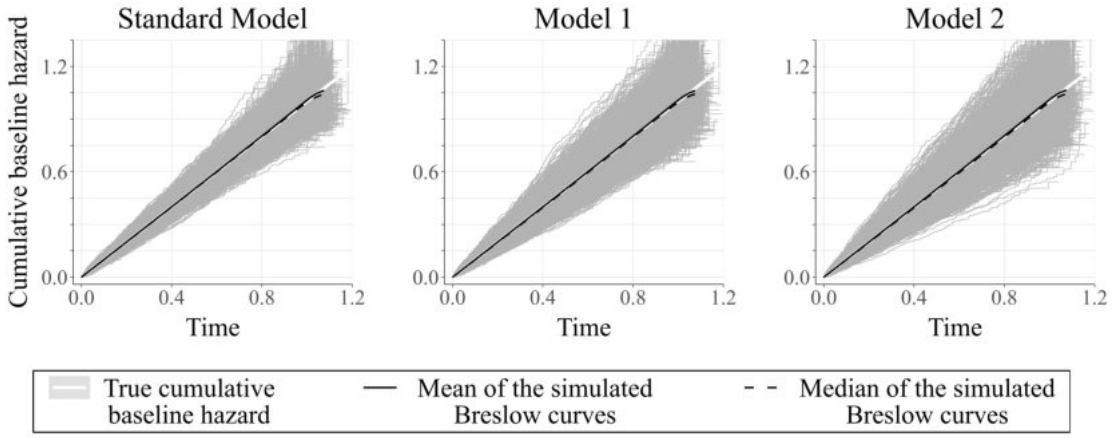


Figure B.2: Shadow plots of the Breslow estimators in the exponential scenario with HR 1, $n = 300$, and $m = 150$.

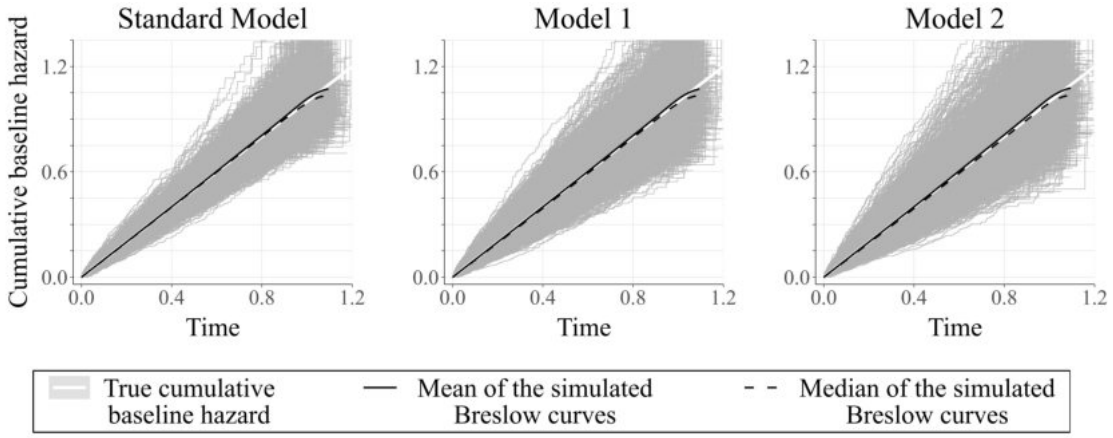


Figure B.3: Shadow plots of the Breslow estimators in the exponential scenario with HR 1, $n = 50$, and $m = 25$.

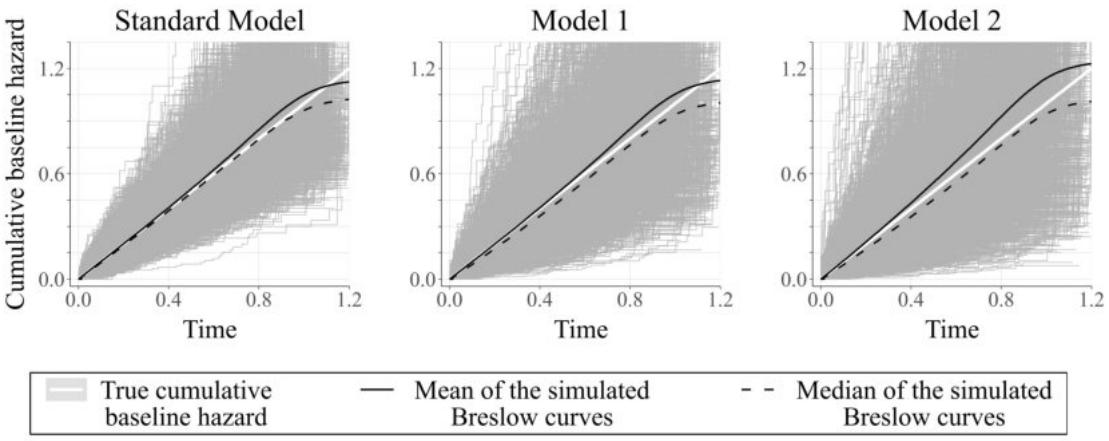


Figure B.4: Shadow plots of the Breslow estimators in the exponential scenario with HR 1, $n = 26$, and $m = 13$ (4 iterations excluded).

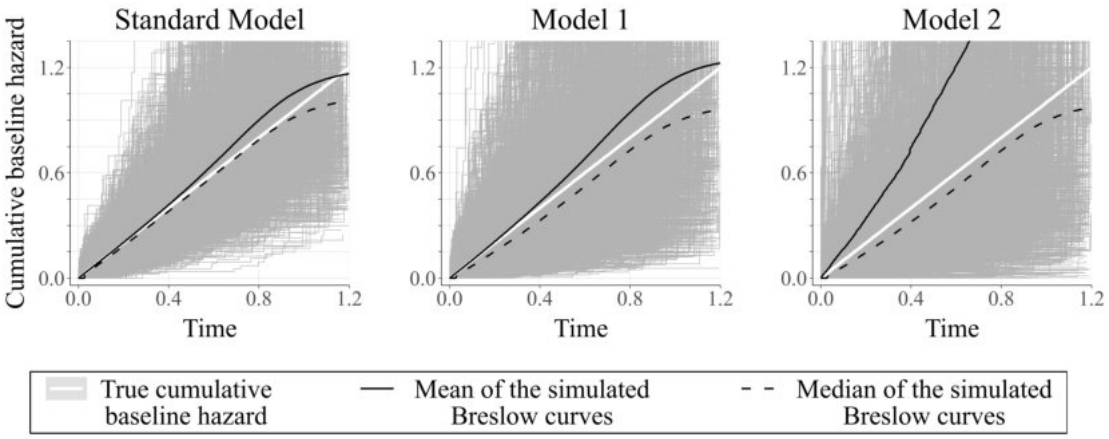


Table B.1: Bias of the estimated log-HRs in the exponential scenarios with HR 1.

n	m	Measure of bias	Standard Model	Model 1		Model 2		
			Z_A	Z_A	Q	Z_A	Q	Z_N
600	300	mean bias	0.00004	0.00002	0.01829	0.00003	0.02890	-0.00001
		median bias	-0.00005	-0.00024	0.01719	-0.00012	0.02548	-0.00002
		RMSE	0.11584	0.11605	0.32065	0.11632	9.49880	0.01096
		coverage	0.95071	0.95024	0.94974	0.95006	0.94991	0.95011
300	150	mean bias	0.00009	0.00012	0.03566	0.00009	-0.00368	0.00009
		median bias	-0.00028	0.00029	0.03259	-0.00042	-0.04291	0.00014
		RMSE	0.16430	0.16484	0.45827	0.16560	9.50765	0.02190
		coverage	0.95055	0.95009	0.94771	0.94934	0.94410	0.94479
50	25	mean bias	0.00130	0.00142	0.21394	0.00199	0.01578	0.00261
		median bias	0.00137	0.00097	0.19848	0.00093	0.04921	0.00244
		RMSE	0.41187	0.42149	1.23124	0.43517	10.80612	0.14689
		coverage	0.95511	0.95193	0.94771	0.94823	0.94410	0.94479
50	10[a]	mean bias	0.00665	0.00687	2.07118	0.00701	0.42306	0.00687
		median bias	0.00233	0.00297	1.86923	0.00450	0.58871	0.00625
		RMSE	0.65962	0.67233	7.12208	0.69416	57.51430	0.25053
		coverage	0.98131	0.97872	0.95275	0.97418	0.95167	0.95235
26	13[b]	mean bias	-0.00299	-0.00279	0.41800	-0.00200	0.09130	0.00789
		median bias	-0.00130	-0.00008	0.37943	-0.00086	0.10810	0.00847
		RMSE	0.59677	0.62824	1.92377	0.67787	12.82265	0.32802
		coverage	0.95929	0.95326	0.94861	0.94575	0.94263	0.94247

[a] 1,032 excluded iterations.

[b] 4 excluded iterations.

Figure B.5: Shadow plots of the Breslow estimators in the Weibull scenario with HR 1, $n = 600$, and $m = 300$.

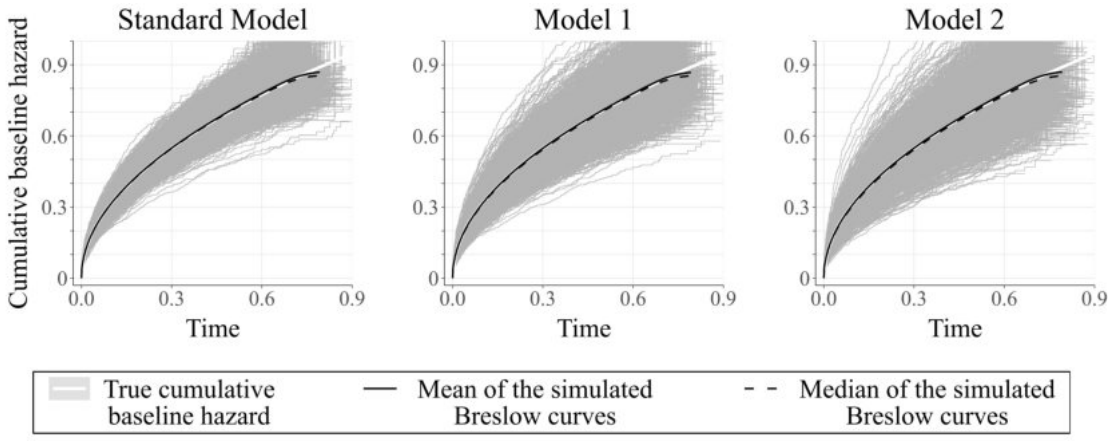


Figure B.6: Shadow plots of the Breslow estimators in the Weibull scenario with HR 1, $n = 300$, and $m = 150$.

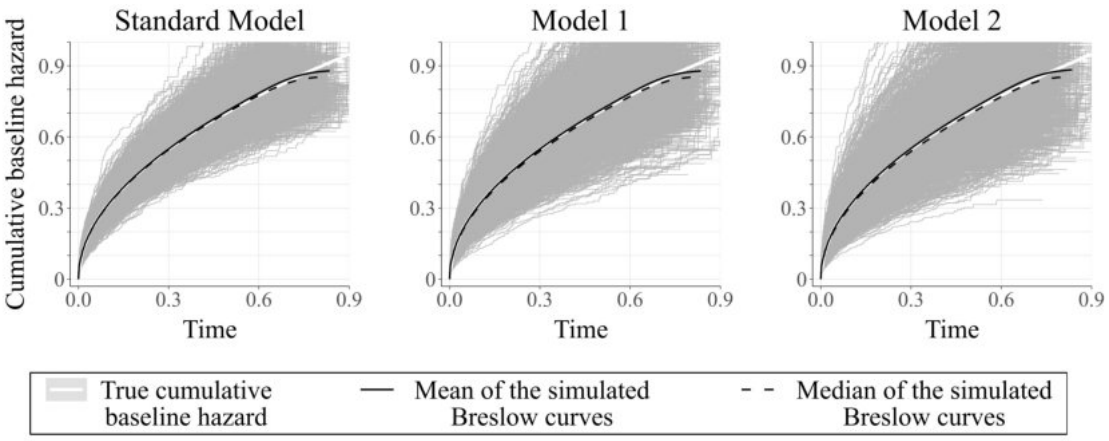


Figure B.7: Shadow plots of the Breslow estimators in the Weibull scenario with HR 1, $n = 50$, and $m = 25$.

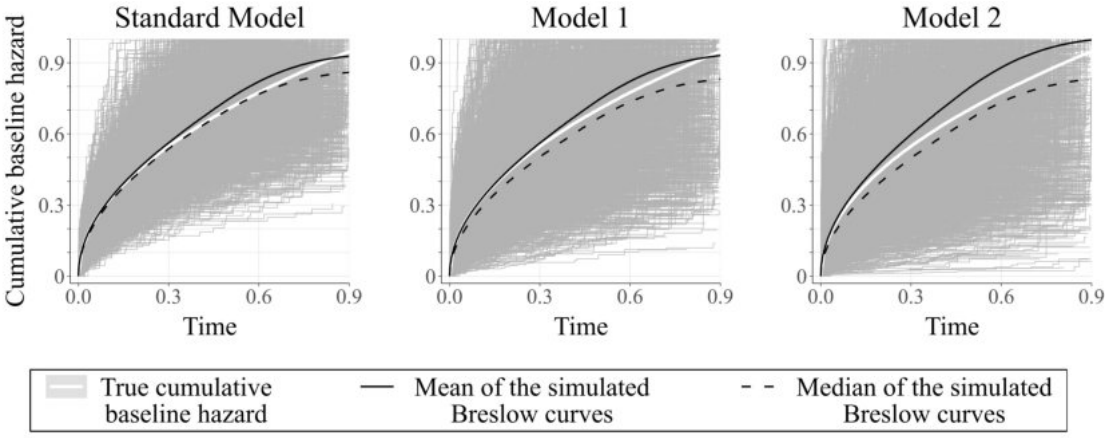


Figure B.8: Shadow plots of the Breslow estimators in the Weibull scenario with HR 1, $n = 26$, and $m = 13$ (4 iterations excluded).

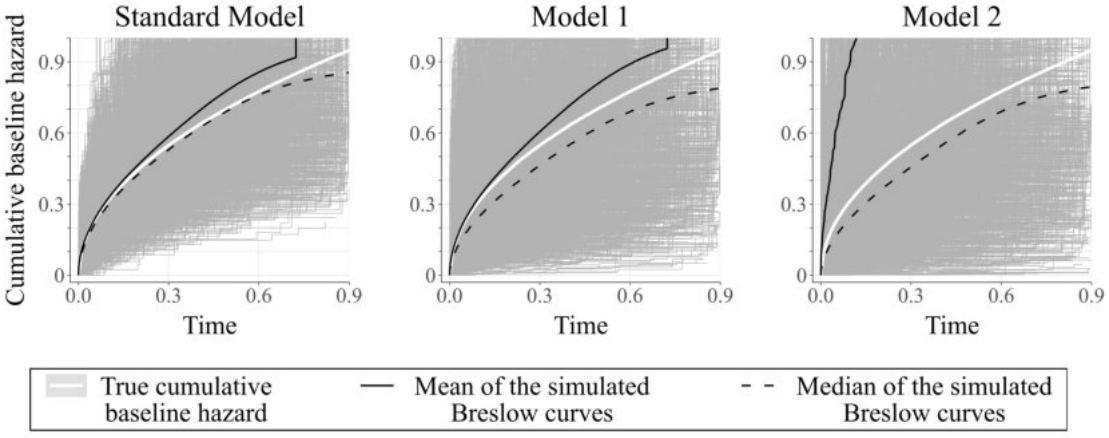


Table B.2: Bias of the estimated log-HRs in the Weibull scenarios with HR 1.

n	m	Measure of bias	Standard Model	Model 1		Model 2		
			Z_A	Z_A	Q	Z_A	Q	Z_N
600	300	mean bias	0.00028	0.00028	0.02759	0.00029	0.02417	0.00000
		median bias	-0.00024	-0.00011	0.02578	-0.00014	-0.01447	0.00003
		RMSE	0.11546	0.11568	0.44059	0.11598	13.56160	0.01084
		coverage	0.95043	0.95025	0.94939	0.95013	0.94874	0.94886
300	150	mean bias	0.00053	0.00049	0.04761	0.00049	0.01555	0.00005
		median bias	0.00015	0.00035	0.04574	0.00045	-0.02485	0.00011
		RMSE	0.16399	0.16461	0.62953	0.16529	13.71267	0.02190
		coverage	0.95000	0.94944	0.94886	0.94896	0.94887	0.94856
50	25	mean bias	0.00262	0.00282	0.29790	0.00274	0.05363	0.00234
		median bias	0.00139	0.00155	0.27817	0.00285	-0.01638	0.00275
		RMSE	0.41109	0.42120	1.68760	0.43426	15.46159	0.14584
		coverage	0.95354	0.95114	0.94829	0.94806	0.94451	0.94459
26	13[a]	mean bias	-0.00238	-0.00256	0.62494	-0.00235	0.15095	0.00854
		median bias	-0.00052	0.00042	0.57390	0.00165	0.14026	0.00848
		RMSE	0.58901	0.62311	2.68667	0.67014	18.34748	0.32545
		coverage	0.96170	0.95446	0.94564	0.94719	0.94168	0.94124

[a] 4 excluded iterations.

Figure B.9: Shadow plots of the Breslow estimators in the exponential scenario with HR 0.8, $n = 600$, and $m = 300$.

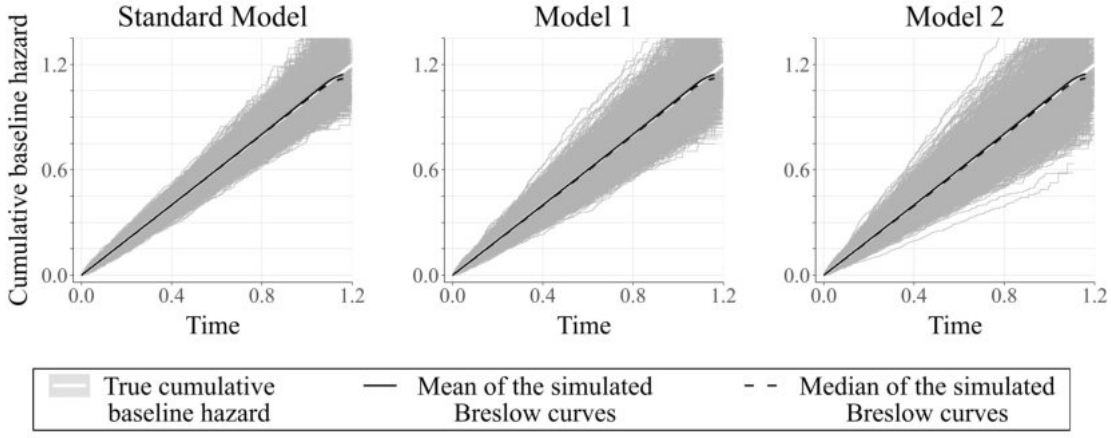


Figure B.10: Shadow plots of the Breslow estimators in the exponential scenario with HR 0.8, $n = 300$, and $m = 150$.

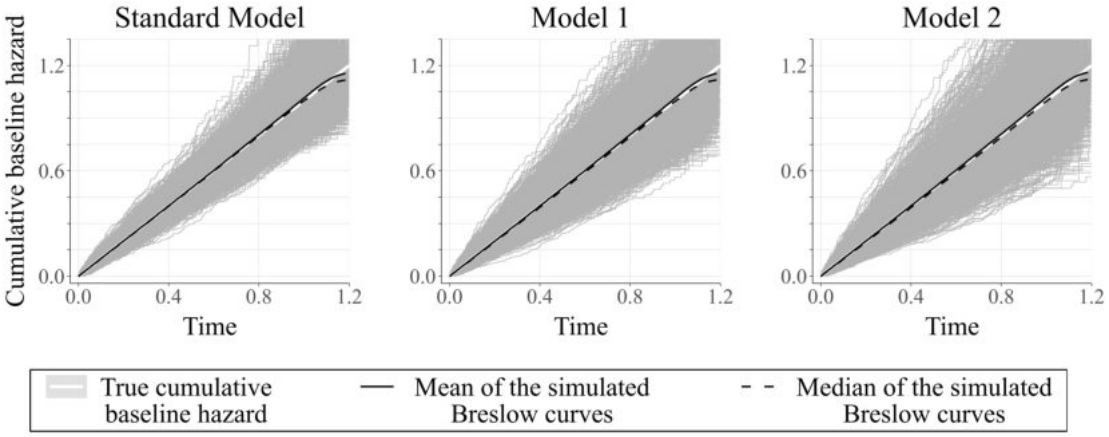


Figure B.11: Shadow plots of the Breslow estimators in the exponential scenario with HR 0.8, $n = 50$, and $m = 25$.

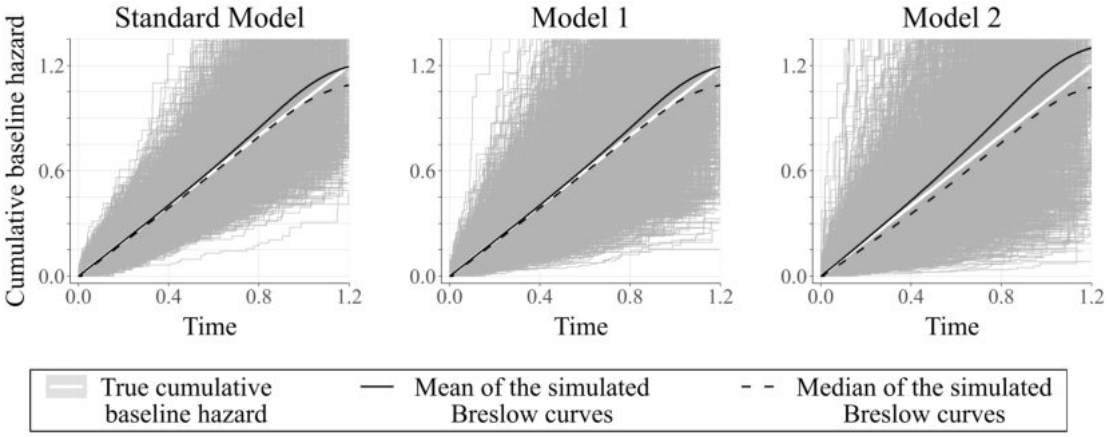


Figure B.12: Shadow plots of the Breslow estimators in the exponential scenario with HR 0.8, $n = 50$, and $m = 10$ (1,500 iterations excluded).

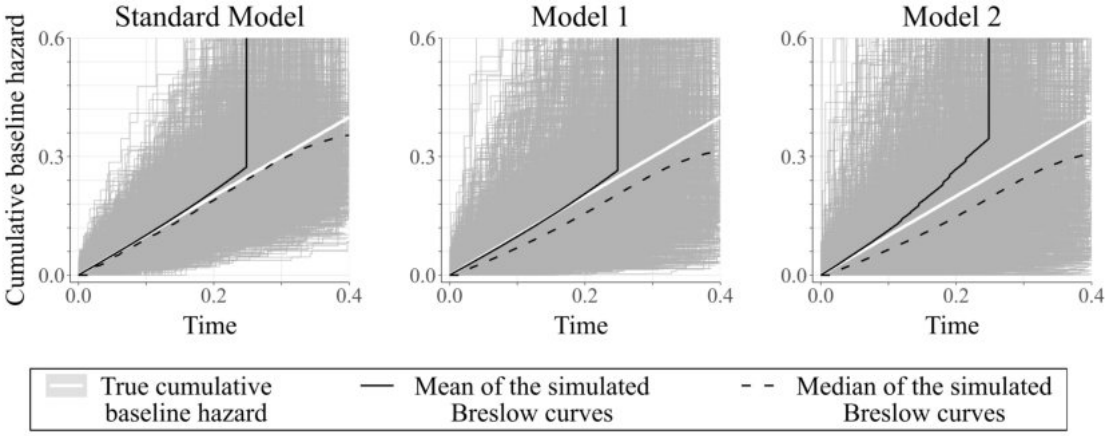


Figure B.13: Shadow plots of the Breslow estimators in the exponential scenario with HR 0.8, $n = 26$, and $m = 13$ (8 iterations excluded).

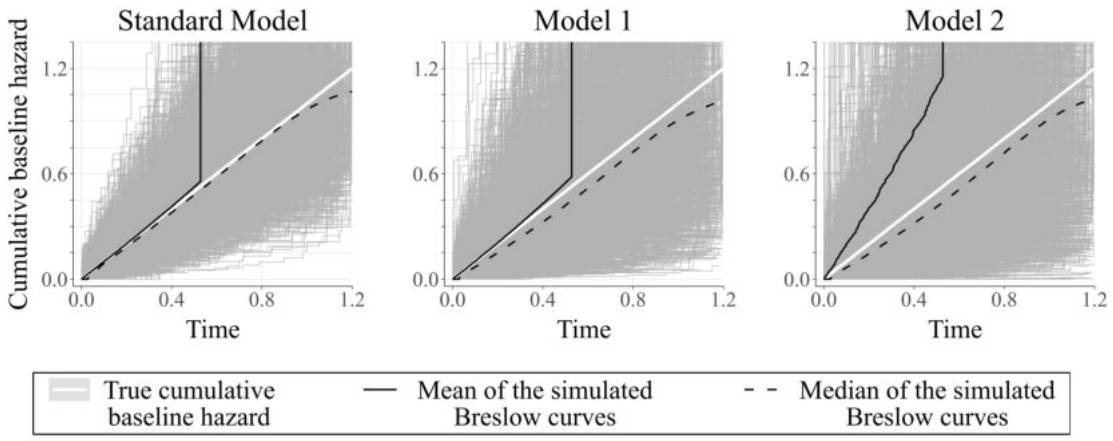


Table B.3: Bias of the estimated log-HRs in the exponential scenarios with HR 0.8.

n	m	Measure of bias	Standard Model	Model 1		Model 2		
			Z_A	Z_A	Q	Z_A	Q	Z_N
600	300	mean bias	-0.00020	-0.00041	0.01795	-0.00064	0.02610	-0.00001
		median bias	0.00015	-0.00018	0.01608	-0.00044	0.04292	-0.00002
		RMSE	0.11617	0.11639	0.31725	0.11665	9.47582	0.01093
		coverage	0.95057	0.94997	0.94963	0.94965	0.95017	0.95013
300	150	mean bias	-0.00049	-0.00086	0.03487	-0.00135	-0.00800	0.00010
		median bias	-0.00059	-0.00074	0.03312	-0.00120	-0.02660	0.00013
		RMSE	0.16469	0.16526	0.45277	0.16602	9.48315	0.02185
		coverage	0.95075	0.95033	0.94898	0.94933	0.94993	0.95001
50	25	mean bias	-0.00287	-0.00536	0.20705	-0.00853	-0.00183	0.00278
		median bias	0.00053	-0.00165	0.19503	-0.00397	0.00976	0.00275
		RMSE	0.41308	0.42282	1.21526	0.43651	10.79102	0.14679
		coverage	0.95474	0.95180	0.94821	0.94859	0.94408	0.94381
50	10[a]	mean bias	0.00826	0.00659	1.98319	0.00327	0.37544	0.00680
		median bias	0.01074	0.00970	1.80658	0.00469	0.52921	0.00698
		RMSE	0.65726	0.66995	6.93843	0.69177	56.92291	0.24801
		coverage	0.98213	0.98010	0.95256	0.97625	0.95309	0.95315
26	13[b]	mean bias	-0.01234	-0.01818	0.40668	-0.02730	0.06872	0.00825
		median bias	-0.00500	-0.01112	0.37047	-0.01715	0.11217	0.00759
		RMSE	0.60647	0.63856	1.89449	0.68757	12.77146	0.32692
		coverage	0.95860	0.95357	0.94772	0.94621	0.94319	0.94267

[a] 1,500 excluded iterations.

[b] 8 excluded iterations.

Figure B.14: Shadow plots of the Breslow estimators in the Weibull scenario with HR 0.8, $n = 600$, and $m = 300$.

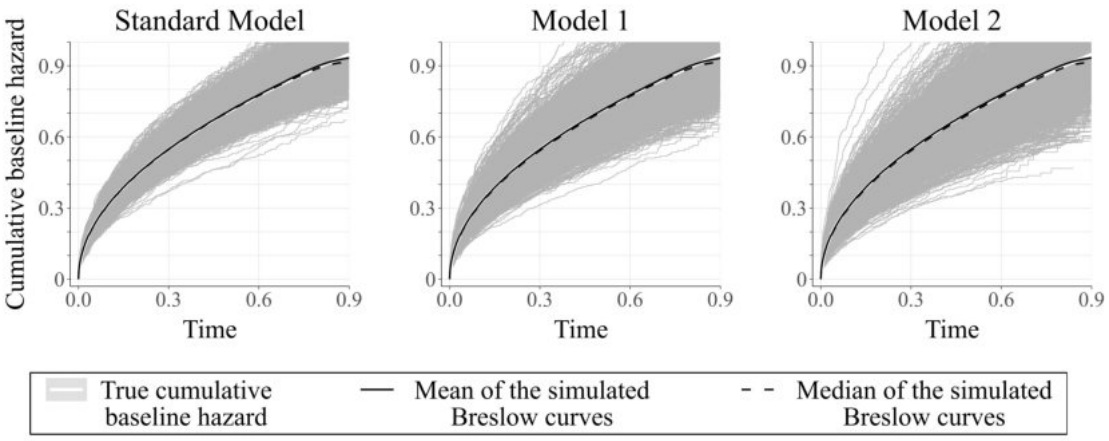


Figure B.15: Shadow plots of the Breslow estimators in the Weibull scenario with HR 0.8, $n = 300$, and $m = 150$.

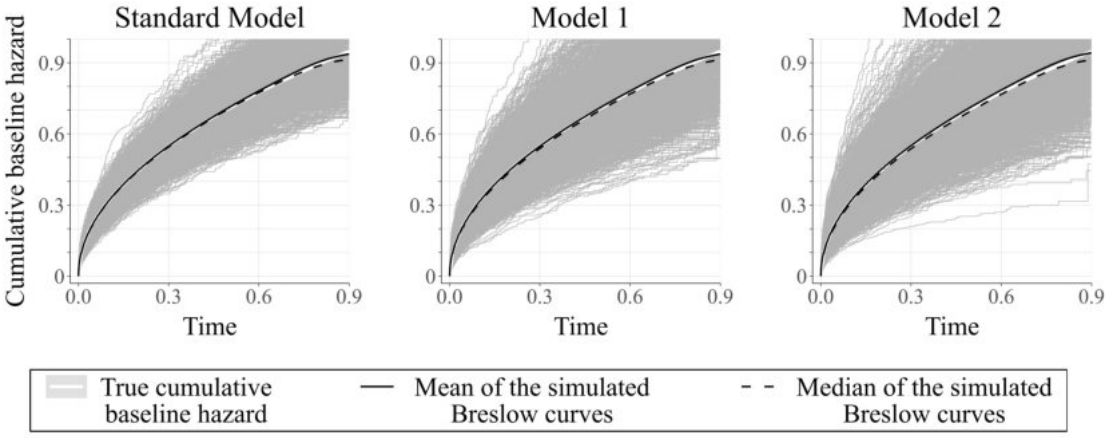


Figure B.16: Shadow plots of the Breslow estimators in the Weibull scenario with HR 0.8, $n = 50$, and $m = 25$.

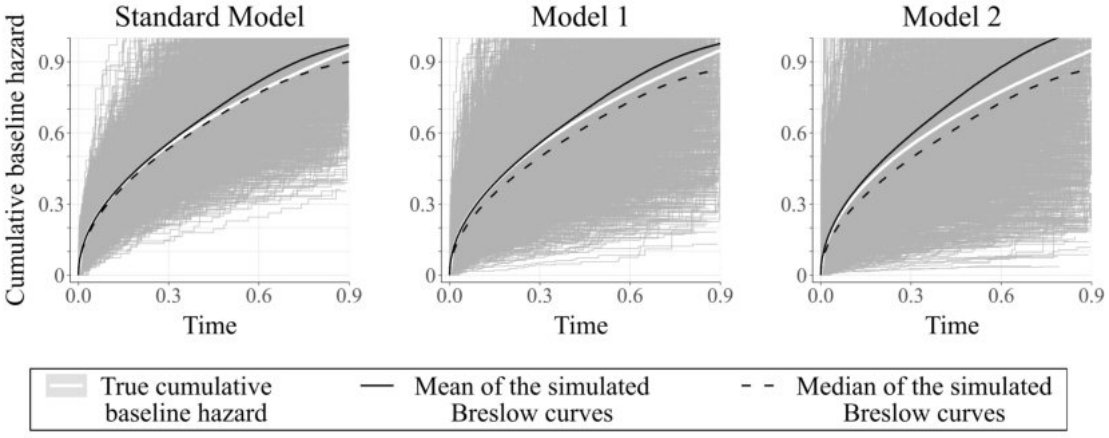


Figure B.17: Shadow plots of the Breslow estimators in the Weibull scenario with HR 0.8, $n = 50$, and $m = 10$ (1,479 iterations excluded).

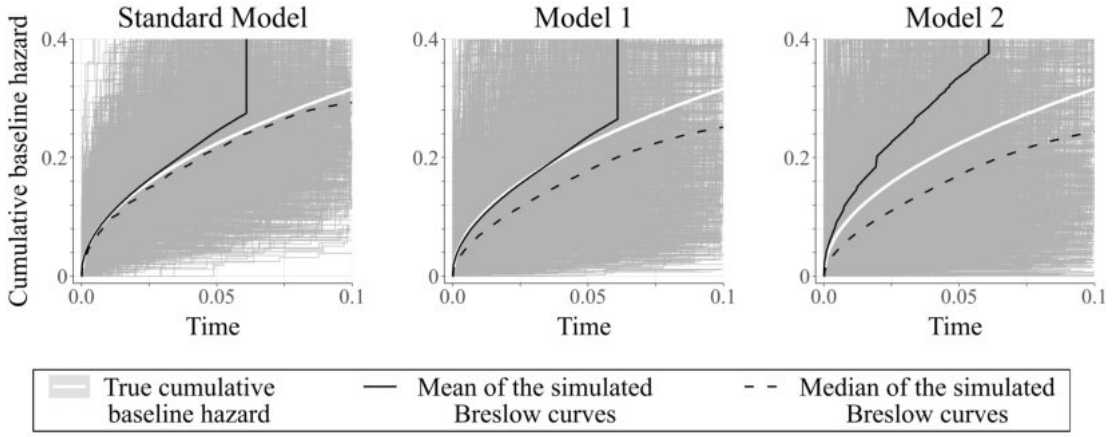


Figure B.18: Shadow plots of the Breslow estimators in the Weibull scenario with HR 0.8, $n = 26$, and $m = 13$ (10 iterations excluded).

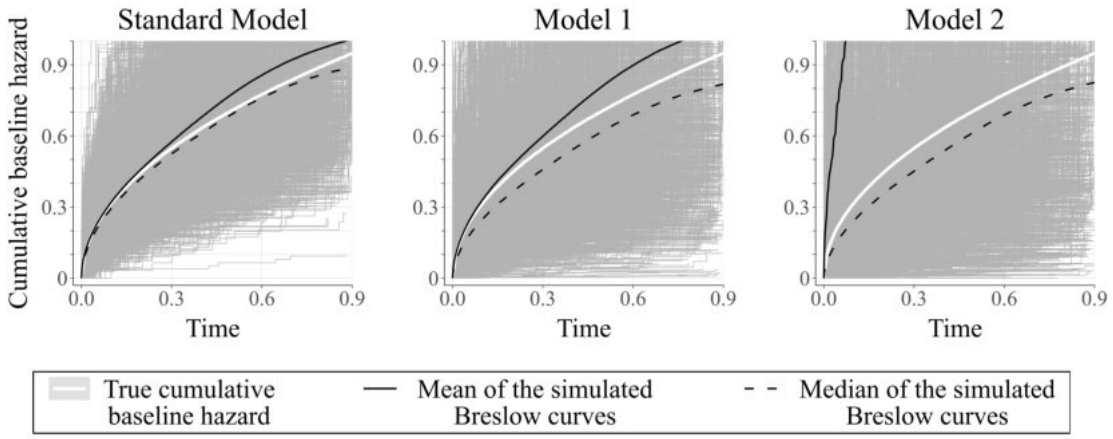


Table B.4: Bias of the estimated log-HRs in the Weibull scenarios with HR 0.8.

n	m	Measure of bias	Standard Model	Model 1		Model 2		
			Z_A	Z_A	Q	Z_A	Q	Z_N
600	300	mean bias	-0.00010	-0.00029	0.02771	-0.00052	0.03908	-0.00001
		median bias	0.00012	0.00002	0.02800	-0.00023	0.02652	0.00000
		RMSE	0.11590	0.11612	0.43619	0.11642	13.54622	0.01083
		coverage	0.95110	0.95058	0.94930	0.95020	0.94904	0.94918
300	150	mean bias	-0.00012	-0.00059	0.04525	-0.00106	0.00331	0.00007
		median bias	0.00033	-0.00019	0.04399	-0.00076	-0.00699	0.00007
		RMSE	0.16448	0.16511	0.62319	0.16582	13.69249	0.02187
		coverage	0.95079	0.95046	0.94887	0.94943	0.94852	0.94828
50	25	mean bias	-0.00085	-0.00347	0.28606	-0.00706	0.03714	0.00240
		median bias	0.00224	-0.00019	0.27068	-0.00249	-0.00769	0.00281
		RMSE	0.41229	0.42242	1.66060	0.43547	15.38364	0.14519
		coverage	0.95425	0.95197	0.94810	0.94909	0.94449	0.94472
50	10[a]	mean bias	0.00458	0.00269	8.62521	-0.00098	0.18233	0.00850
		median bias	0.00690	0.00203	7.83206	-0.00135	2.01290	0.00661
		RMSE	0.65566	0.66871	29.28074	0.68966	252.89386	0.24643
		coverage	0.98415	0.98220	0.95069	0.97793	0.95056	0.95042
26	13[b]	mean bias	-0.00934	-0.01632	0.59292	-0.02540	0.08218	0.00916
		median bias	0.00031	-0.00475	0.55326	-0.01416	0.10825	0.00863
		RMSE	0.58632	0.62084	2.63070	0.66823	18.22146	0.32374
		coverage	0.96068	0.95518	0.94594	0.94768	0.94169	0.94164

[a] 1,479 excluded iterations.

[b] 10 excluded iterations.

Figure B.19: Shadow plots of the Breslow estimators in the exponential scenario with HR 1.25, $n = 600$, and $m = 300$.

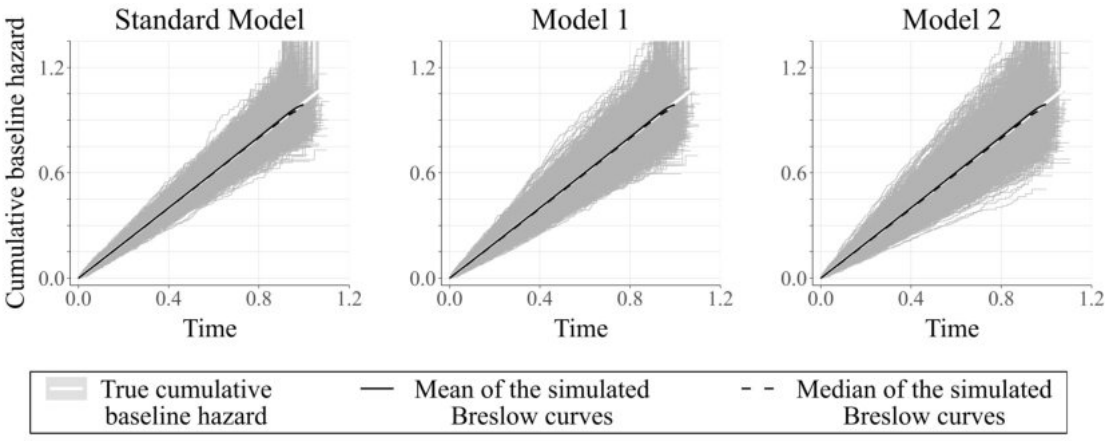


Figure B.20: Shadow plots of the Breslow estimators in the exponential scenario with HR 1.25, $n = 300$, and $m = 150$.

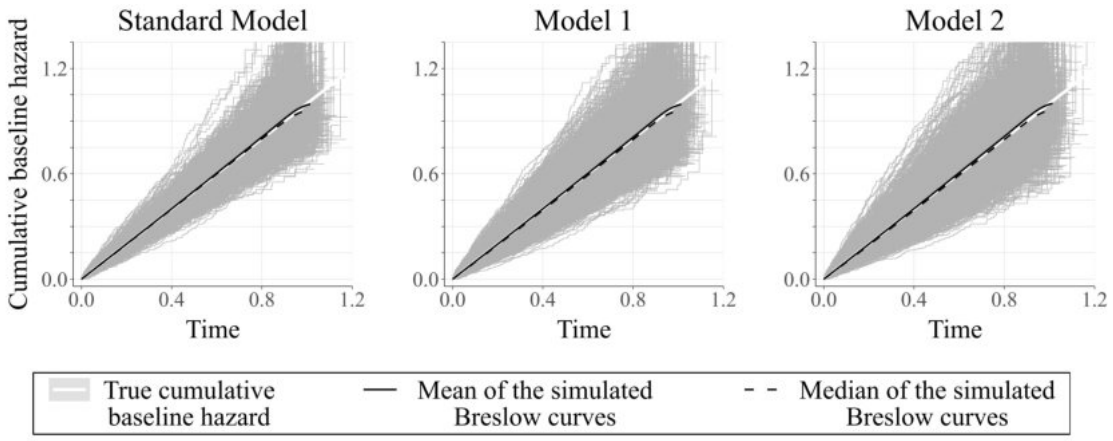


Figure B.21: Shadow plots of the Breslow estimators in the exponential scenario with HR 1.25, $n = 50$, and $m = 25$.

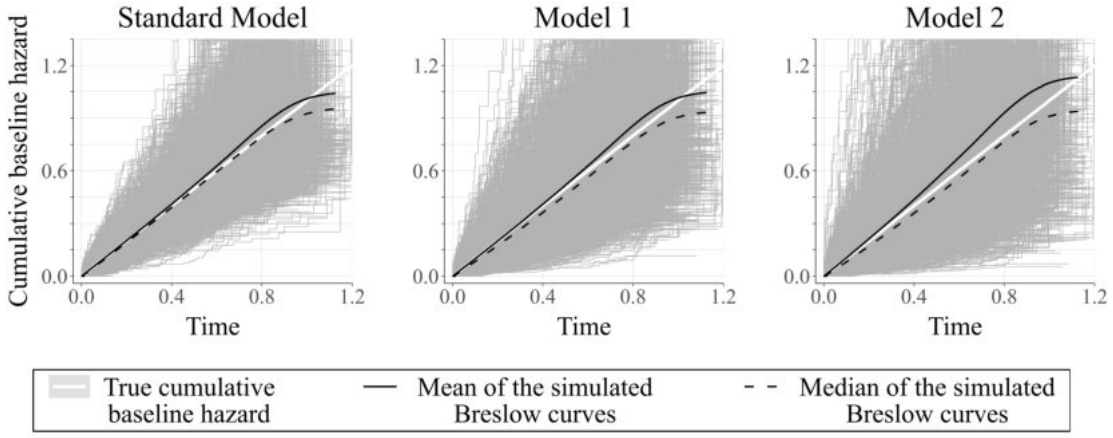


Figure B.22: Shadow plots of the Breslow estimators in the exponential scenario with HR 1.25, $n = 50$, and $m = 10$ (1,527 iterations excluded).

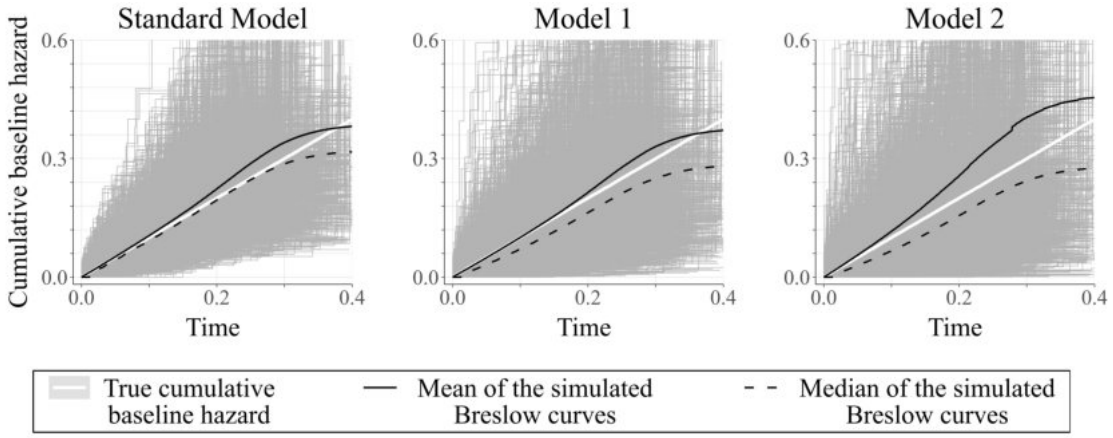


Figure B.23: Shadow plots of the Breslow estimators in the exponential scenario with HR 1.25, $n = 26$, and $m = 13$ (10 iterations excluded).

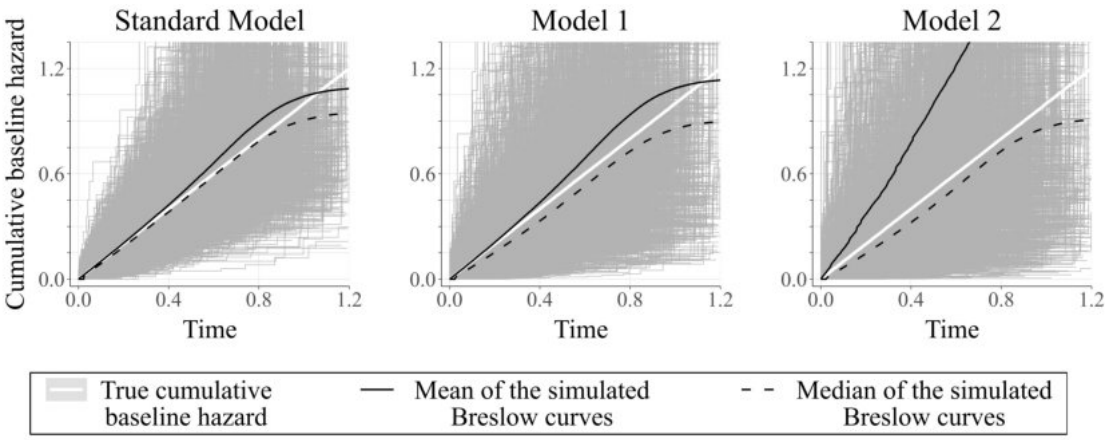


Table B.5: Bias of the estimated log-HRs in the exponential scenarios with HR 1.25.

n	m	Measure of bias	Standard Model	Model 1		Model 2		
			Z_A	Z_A	Q	Z_A	Q	Z_N
600	300	mean bias	0.00020	0.00037	0.01902	0.00061	0.02171	0.00000
		median bias	-0.00011	0.00000	0.01776	0.00025	0.02054	0.00001
		RMSE	0.11606	0.11627	0.32469	0.11654	9.52392	0.01099
		coverage	0.95043	0.94998	0.94975	0.94984	0.94955	0.94961
300	150	mean bias	0.00047	0.00088	0.03641	0.00134	0.00260	0.00008
		median bias	-0.00068	-0.00071	0.03351	-0.00043	-0.01229	0.00010
		RMSE	0.16504	0.16558	0.46434	0.16637	9.54886	0.02200
		coverage	0.94988	0.94998	0.94977	0.94921	0.94968	0.94994
50	25	mean bias	0.00482	0.00757	0.22011	0.01192	0.02180	0.00258
		median bias	0.00145	0.00326	0.20047	0.00696	0.02705	0.00238
		RMSE	0.41379	0.42340	1.24904	0.43727	10.84812	0.14739
		coverage	0.95424	0.95137	0.94820	0.94778	0.94441	0.94475
50	10[a]	mean bias	0.00232	0.00435	2.15942	0.00831	0.33045	0.00763
		median bias	0.00235	0.00552	1.94781	0.00897	0.72940	0.00654
		RMSE	0.65883	0.67132	7.32056	0.69386	58.21950	0.25331
		coverage	0.98237	0.98070	0.95249	0.97651	0.95152	0.95189
26	13[b]	mean bias	0.00506	0.01083	0.43388	0.02132	0.09288	0.00816
		median bias	-0.00397	0.00174	0.38380	0.01045	0.10516	0.00972
		RMSE	0.61441	0.64598	1.96279	0.69642	12.95524	0.33095
		coverage	0.95949	0.95349	0.94867	0.94639	0.94193	0.94202

[a] 1,527 excluded iterations.

[b] 10 excluded iterations.

Figure B.24: Shadow plots of the Breslow estimators in the Weibull scenario with HR 1.25, $n = 600$, and $m = 300$.

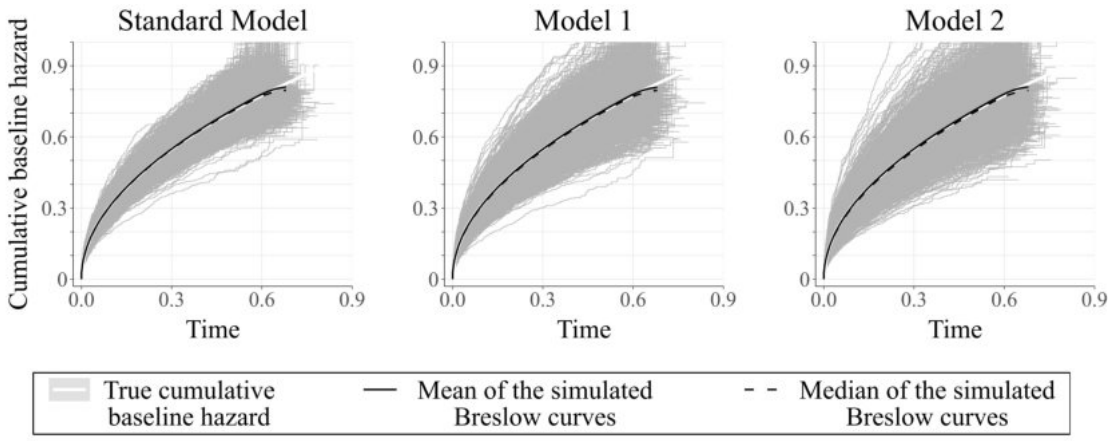


Figure B.25: Shadow plots of the Breslow estimators in the Weibull scenario with HR 1.25, $n = 300$, and $m = 150$.

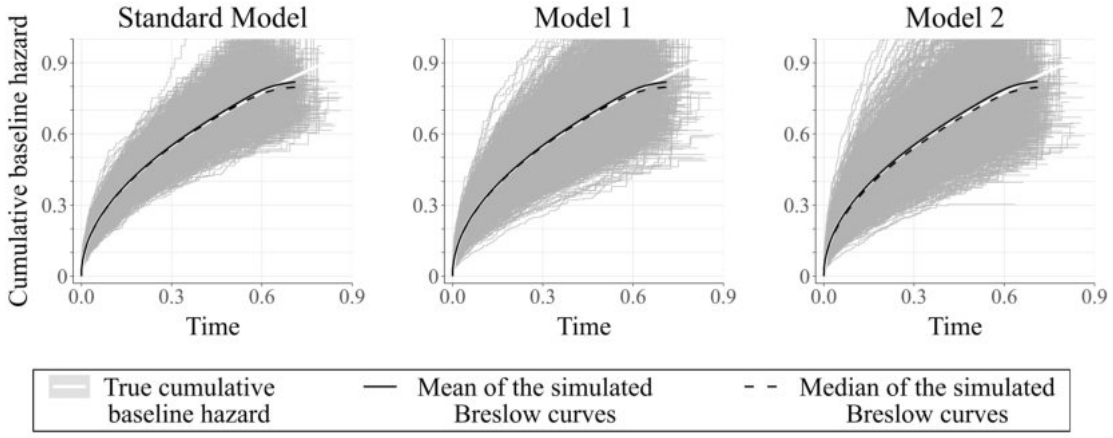


Figure B.26: Shadow plots of the Breslow estimators in the Weibull scenario with HR 1.25, $n = 50$, and $m = 25$.

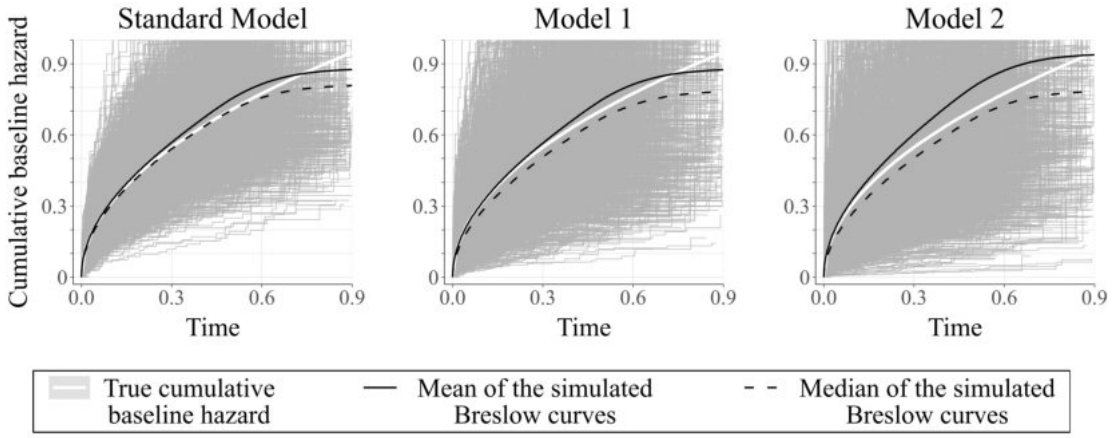


Figure B.27: Shadow plots of the Breslow estimators in the Weibull scenario with HR 1.25, $n = 50$, and $m = 10$ (1,505 iterations excluded).

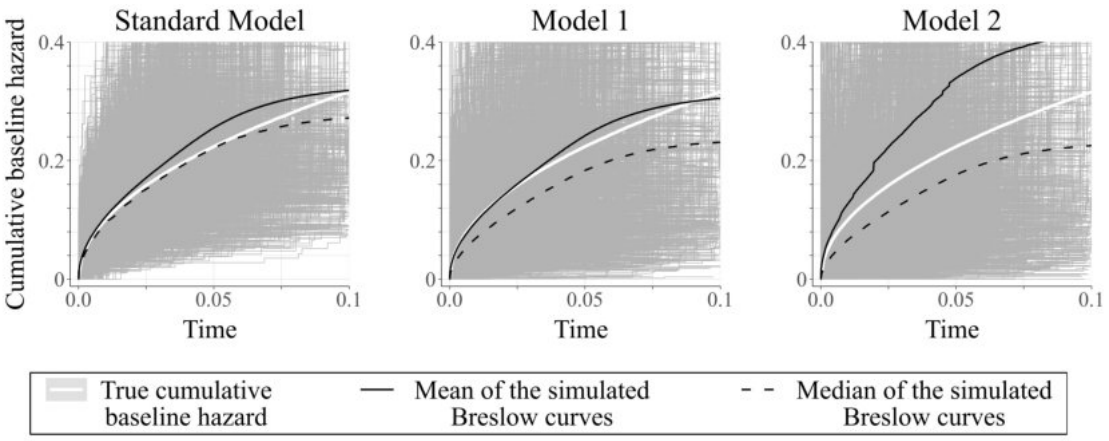


Figure B.28: Shadow plots of the Breslow estimators in the Weibull scenario with HR 1.25, $n = 26$, and $m = 13$ (9 iterations excluded).

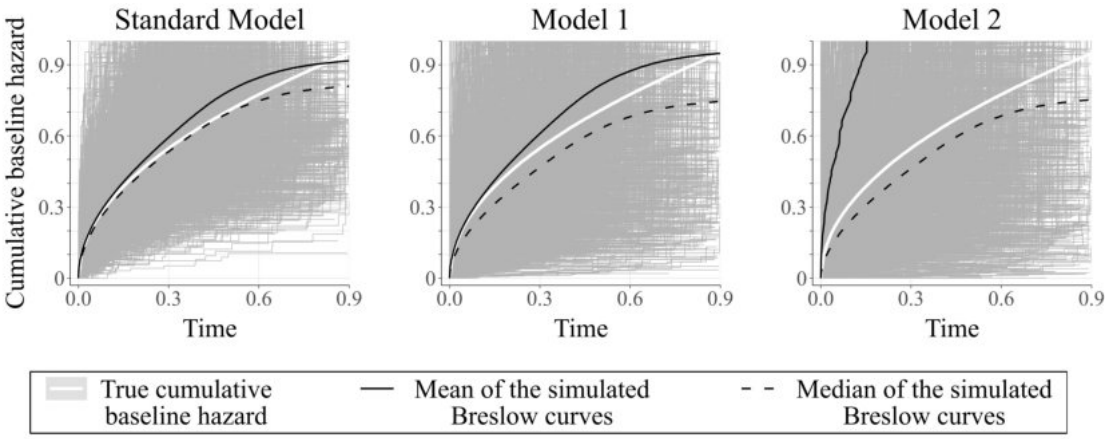


Table B.6: Bias of the estimated log-HRs in the Weibull scenarios with HR 1.25.

n	m	Measure of bias	Standard Model	Model 1		Model 2		
			Z_A	Z_A	Q	Z_A	Q	Z_N
600	300	mean bias	0.00066	0.00087	0.02928	0.00111	0.01256	0.00001
		median bias	0.00035	0.00044	0.02791	0.00081	-0.00599	0.00002
		RMSE	0.11570	0.11590	0.44663	0.11620	13.59952	0.01087
		coverage	0.95087	0.95076	0.94909	0.95057	0.94968	0.94929
300	150	mean bias	0.00103	0.00140	0.05000	0.00188	0.03304	0.00003
		median bias	0.00067	0.00104	0.04736	0.00125	-0.00369	0.00006
		RMSE	0.16437	0.16497	0.63853	0.16566	13.73754	0.02194
		coverage	0.95115	0.95093	0.94960	0.95015	0.94855	0.94859
50	25	mean bias	0.00609	0.00904	0.31736	0.01264	0.08680	0.00222
		median bias	0.00386	0.00749	0.29321	0.01077	0.07795	0.00235
		RMSE	0.41260	0.42282	1.72339	0.43638	15.59063	0.14689
		coverage	0.95420	0.95133	0.94788	0.94844	0.94314	0.94360
50	10[a]	mean bias	-0.00127	0.00114	9.94427	0.00442	1.98879	0.00818
		median bias	-0.00290	0.00032	8.89267	0.00297	3.68433	0.00626
		RMSE	0.65148	0.66436	31.32874	0.68721	258.51183	0.25134
		coverage	0.98362	0.98158	0.95099	0.97740	0.95084	0.95077
26	13[b]	mean bias	0.00634	0.01270	0.65890	0.02297	0.19613	0.00852
		median bias	0.00142	0.00750	0.59814	0.01483	0.21581	0.00772
		RMSE	0.58809	0.62278	2.75339	0.67134	18.63140	0.32983
		coverage	0.96001	0.95415	0.94573	0.94724	0.94071	0.94067

[a] 1,505 excluded iterations.

[b] 9 excluded iterations.

Figure B.29: Shadow plots of the Breslow estimators in the randomly censored scenario with $n = 50$ and $m = 25$.

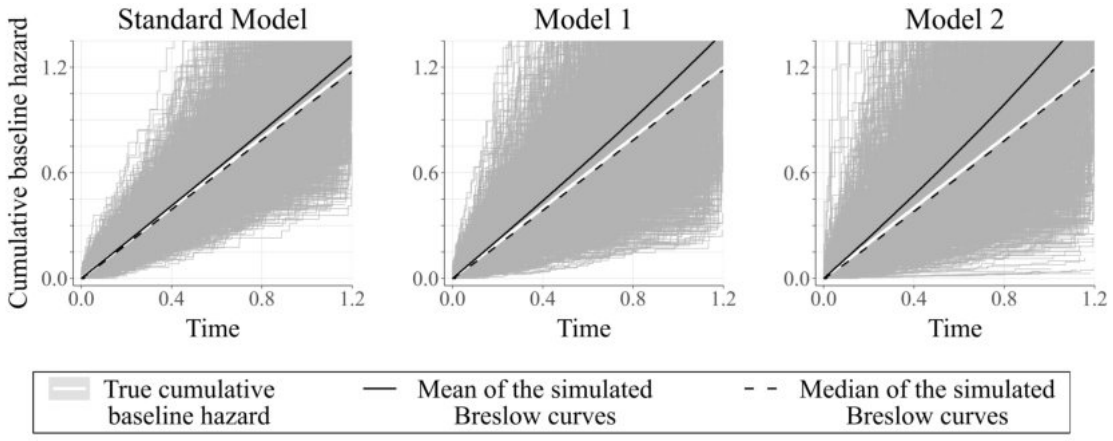


Figure B.30: Shadow plots of the Breslow estimators in the randomly censored scenario with $n = 26$ and $m = 13$ (350 iterations excluded).

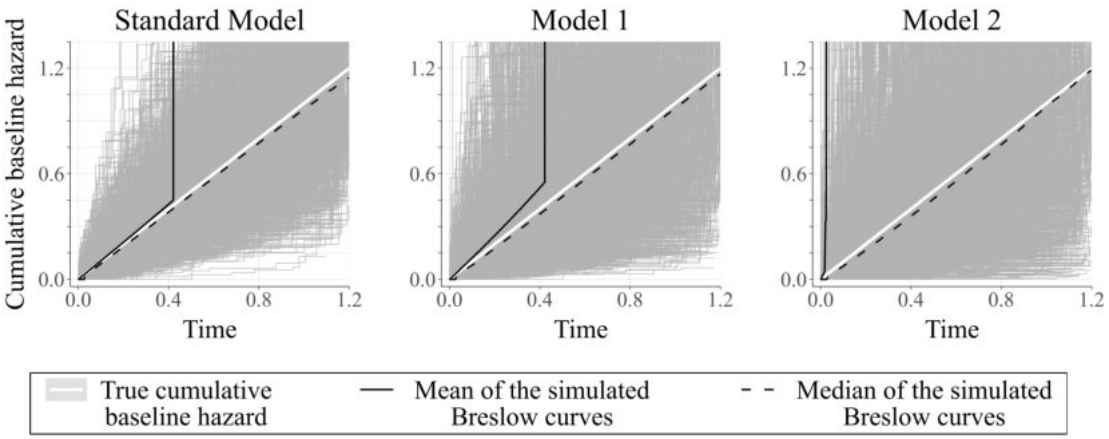


Table B.7: Bias of the estimated log-HRs in the randomly censored scenarios.

n	m	Measure of bias	Standard Model	Model 1		Model 2		
			Z_A	Z_A	Q	Z_A	Q	Z_N
50	25	mean bias	0.00011	-0.00029	0.00226	-0.00021	-0.07464	0.00103
		median bias	-0.00088	-0.00128	0.00421	-0.00139	-0.02309	0.00033
		RMSE	0.43319	0.45225	1.15243	0.47317	11.36095	0.15712
		coverage	0.95185	0.94759	0.94544	0.94272	0.94098	0.94114
50	10[a]	mean bias	0.00134	0.00135	-0.03612	0.00079	-0.00795	-0.00026
		median bias	0.00237	0.00445	-0.01255	0.00546	-0.10112	0.00023
		RMSE	0.80513	0.84112	6.04802	0.88731	61.25004	0.28904
		coverage	0.97681	0.97091	0.95405	0.96482	0.94782	0.94774
26	13[b]	mean bias	0.00138	0.00235	0.00486	0.00334	-0.01111	0.00033
		median bias	0.00410	0.00666	0.00400	0.00711	0.00524	0.00005
		RMSE	0.78476	0.83879	1.85816	0.90509	14.19590	0.37999
		coverage	0.95941	0.95004	0.94555	0.94098	0.93654	0.93654

[a] 2133 excluded iterations.

[b] 350 excluded iterations.

B.2. Resampling-based inference for the ATE using the g-formula

The subsequent table and figures were generated within the scope of the simulation study presented in Subsection 4.1.2.

Table B.8: Frequency of errors and convergence issues for the simulation study that investigates the resampling methods on the basis of the g-formula.

Scenario	n	β_{01A}	Iterations with errors (out of 5,000)	Iterations with convergence issues (out of 5,000)	mean number of invalid EBS samples (out of 1,000)				
					$t=1$	$t=3$	$t=5$	$t=7$	$t=9$
No censoring	50	-2	0	16	46.7	120.0	127.3	133.6	140.6
		0	0	33	82.3	112.5	117.4	123.1	129.1
		2	8	464	85.1	90.3	94.5	97.9	100.7
	75	-2	0	0	2.7	6.3	6.6	6.8	7.1
		0	0	0	10.0	10.9	11.3	11.9	12.5
		2	1	28	84.1	86.5	89.6	92.8	95.9
	100	-2	0	0	0.0	0.1	0.1	0.1	0.1
		0	0	0	0.2	0.2	0.2	0.2	0.3
		2	0	0	22.6	23.1	23.7	24.6	25.5
Light censoring	50	-2	1	100	68.5	169.7	179.0	187.3	196.3
		0	5	137	95.8	134.8	140.5	147.5	154.0
		2	2	547	69.4	73.3	76.7	79.3	81.3
	75	-2	0	0	12.7	26.7	27.7	28.7	30.0
		0	0	0	29.7	33.1	34.1	35.5	37.1
		2	4	108	92.5	95.4	98.5	101.7	104.4
	100	-2	0	0	0.5	1.0	1.0	1.0	1.0
		0	0	0	2.2	2.3	2.4	2.5	2.6
		2	0	4	48.7	49.6	50.9	52.5	53.9
Heavy censoring	50	-2	10	501	1.1	180.4	191.9	200.7	208.2
		0	3	359	112.0	155.5	162.4	171.0	176.6
		2	8	580	80.1	83.7	87.4	90.5	92.0
	75	-2	0	11	43.3	86.7	89.7	93.5	97.7
		0	1	14	62.7	72.1	74.2	77.1	79.2
		2	5	287	80.8	83.2	85.9	88.3	89.4
	100	-2	0	0	5.8	11.0	11.3	11.7	12.2
		0	0	0	14.6	15.4	15.8	16.3	16.8
		2	3	40	77.4	78.8	80.9	83.1	84.3
	200	2	0	0	0.6	0.6	0.6	0.6	0.6

Scenario	n	β_{01A}	Iterations with errors (out of 5,000)	Iterations with convergence issues (out of 5,000)	mean number of invalid EBS samples (out of 1,000)				
					$t=1$	$t=3$	$t=5$	$t=7$	$t=9$
Low treatment probability	50	-2	1	45	97.9	167.4	173.7	182.0	190.7
		0	4	138	103.3	143.0	148.2	154.9	161.3
		2	9	331	134.6	144.5	149.5	155.2	160.0
	75	-2	0	0	12.2	15.6	16.0	16.6	17.4
		0	0	0	30.1	33.5	34.4	35.7	37.2
		2	0	7	64.8	66.0	67.6	69.6	71.5
	100	-2	0	0	0.3	0.4	0.4	0.4	0.4
		0	0	0	2.3	2.3	2.4	2.5	2.6
		2	0	0	10.0	10.1	10.3	10.6	10.9
High treatment probability	50	-2	8	501	34.6	133.5	151.8	160.7	167.7
		0	4	141	104.4	149.0	156.8	164.3	170.9
		2	8	536	66.3	70.3	72.1	72.9	73.3
	75	-2	0	12	21.8	71.0	76.8	79.9	83.1
		0	0	0	29.5	33.5	34.9	36.4	37.9
		2	6	335	73.4	77.3	80.4	82.7	83.9
	100	-2	0	0	4.4	13.1	14.0	14.5	14.9
		0	0	0	2.3	2.4	2.5	2.6	2.7
		2	2	66	80.0	82.9	85.9	88.6	89.9
	200	2	0	0	1.3	1.3	1.3	1.4	1.4
Low variance of the covariates	50	-2	1	63	40.9	131.5	140.7	147.5	155.6
		0	1	77	57.3	104.4	108.9	114.3	119.7
		2	9	510	69.2	73.6	76.9	79.1	80.3
	75	-2	0	0	4.8	14.6	15.3	15.9	16.7
		0	0	0	13.5	17.8	18.3	19.1	20.0
		2	0	87	85.3	88.2	91.3	94.6	97.3
	100	-2	0	0	0.1	0.4	0.5	0.5	0.5
		0	0	0	0.8	1.0	1.0	1.0	1.1
		2	1	5	45.0	46.0	47.2	48.8	50.1

Scenario	n	β_{01A}	Iterations with errors (out of 5,000)	Iterations with convergence issues (out of 5,000)	mean number of invalid EBS samples (out of 1,000)				
					$t=1$[a]	$t=3$[a]	$t=5$[a]	$t=7$[a]	$t=9$[a]
High variance of the covariates	50	-2	3	475	184.2	235.0	245.2	254.9	263.8
		0	11	501	190.2	201.8	211.0	220.3	228.2
		2	9	688	111.5	117.9	126.0	133.5	139.4
	75	-2	0	0	88.2	99.5	102.7	106.5	110.4
		0	0	17	93.5	95.9	99.3	103.4	107.0
		2	1	183	113.4	116.7	120.8	124.4	127.0
	100	-2	0	0	6.9	7.6	7.8	8.0	8.3
		0	0	0	15.9	16.1	16.8	17.5	18.2
		2	0	6	76.2	77.7	80.1	82.6	84.6
Type II censoring	50	-2	0	0	20.7	26.5	27.6	28.7	29.9
		0	0	0	16.8	22.5	23.0	23.3	23.9
		2	0	0	38.3	45.3	45.7	46.0	46.4

[a] For the scenario with type II censoring, we considered $t \in \{2, 4, 6, 8, 10\}$, $t \in \{1, 2, 3, 4, 5\}$, and $t \in \{0.5, 1, 1.5, 2, 2.5\}$ for $\beta_{01A} = -2, 0, 2$, respectively.

Figure B.31: Coverage of the g-formula CIs in the scenario with no censoring and $\beta_{01A} = 0$.

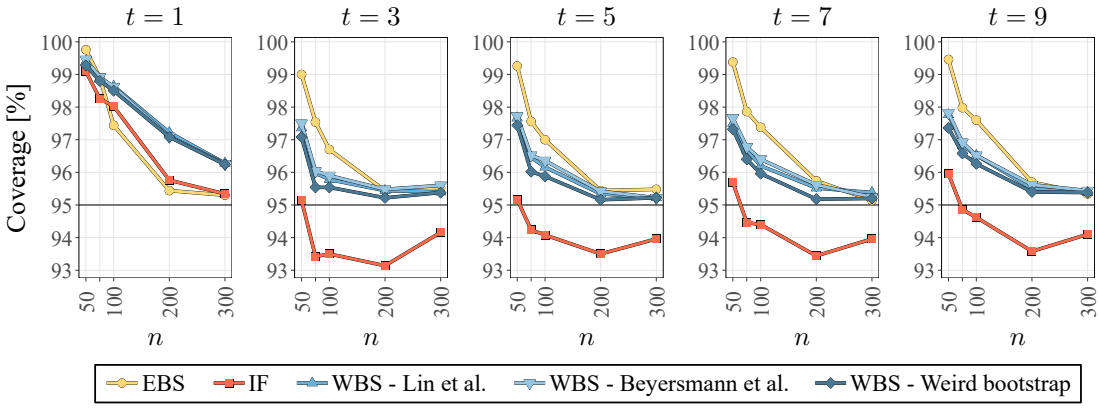


Figure B.32: Coverage of the g-formula CIs in the scenario with no censoring and $\beta_{01A} = 2$.

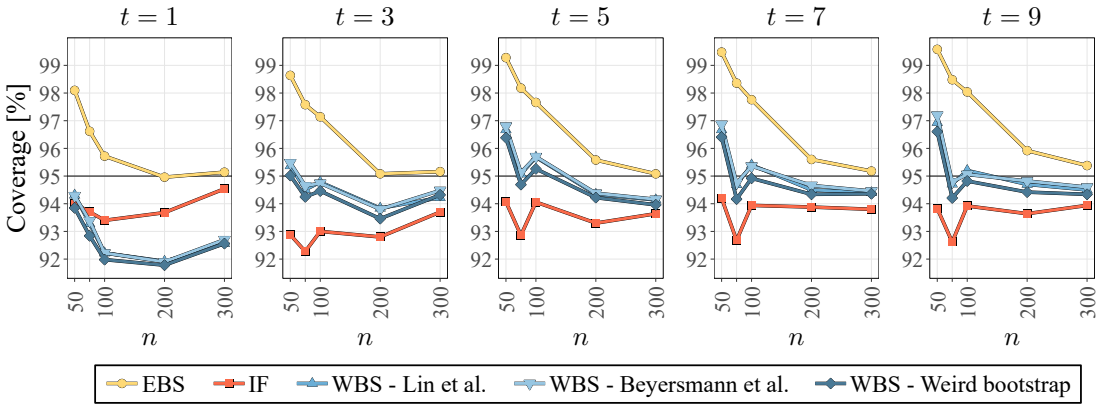


Figure B.33: Coverage of the g-formula CIs in the scenario with light censoring and $\beta_{01A} = -2$.

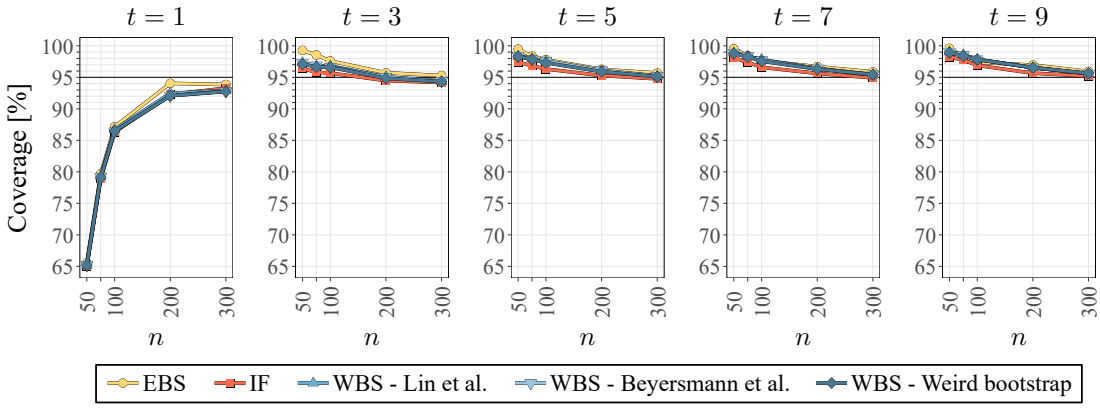


Figure B.34: Coverage of the g-formula CIs in the scenario with light censoring and $\beta_{01A} = 0$.

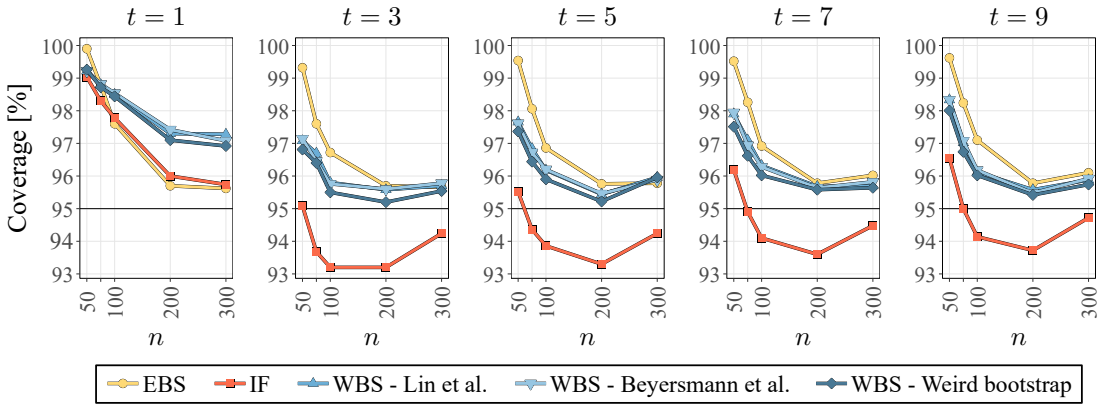


Figure B.35: Coverage of the g-formula CIs in the scenario with heavy censoring and $\beta_{01A} = -2$.

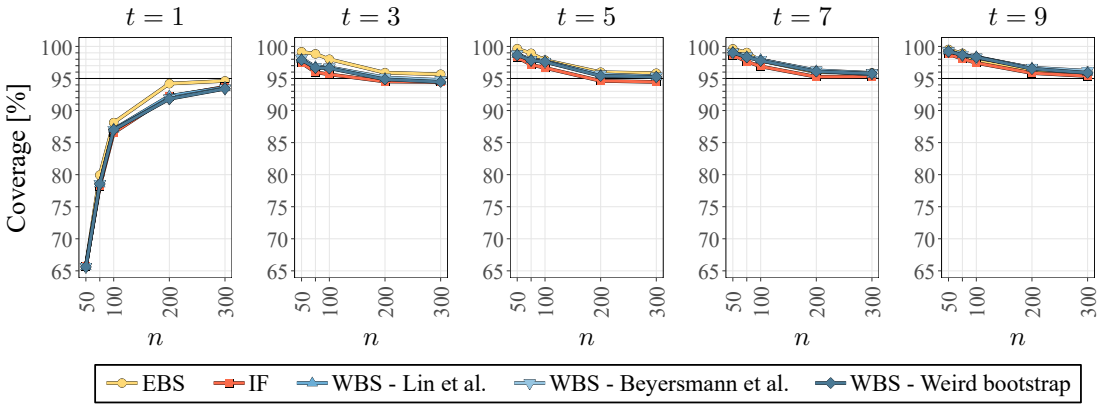


Figure B.36: Coverage of the g-formula CIs in the scenario with heavy censoring and $\beta_{01A} = 0$.

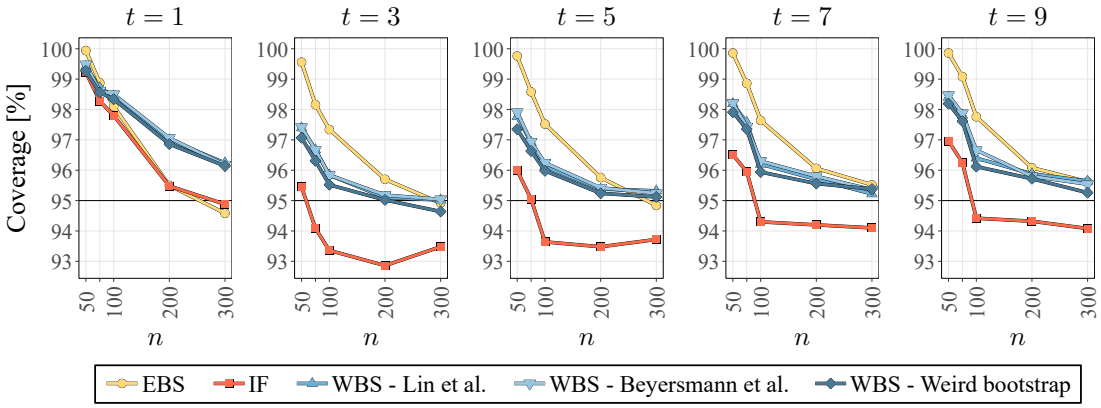


Figure B.37: Coverage of the g-formula CIs in the scenario with heavy censoring and $\beta_{01A} = 2$.

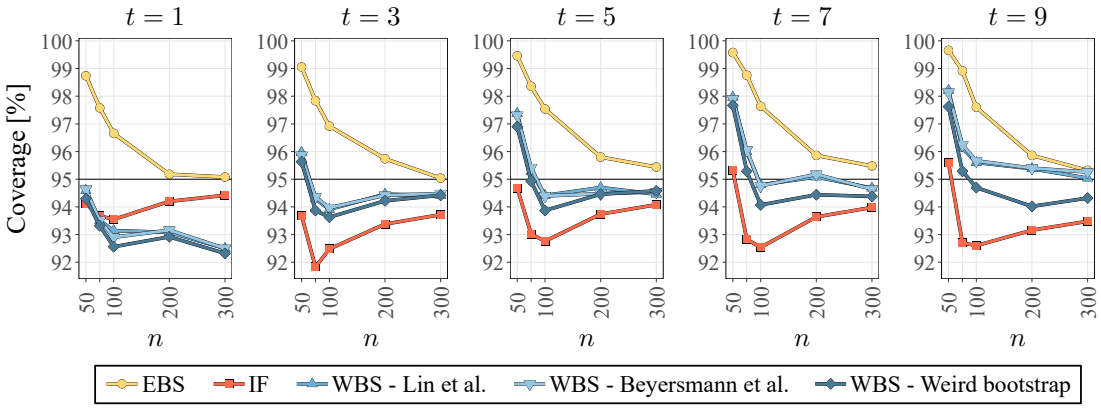


Figure B.38: Coverage of the g-formula CIs in the scenario with low treatment probability and $\beta_{01A} = -2$.

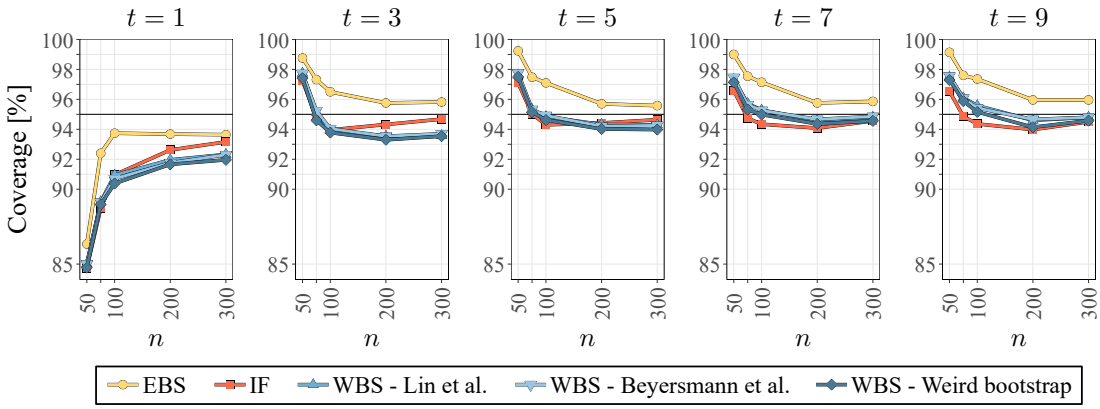


Figure B.39: Coverage of the g-formula CIs in the scenario with low treatment probability and $\beta_{01A} = 0$.

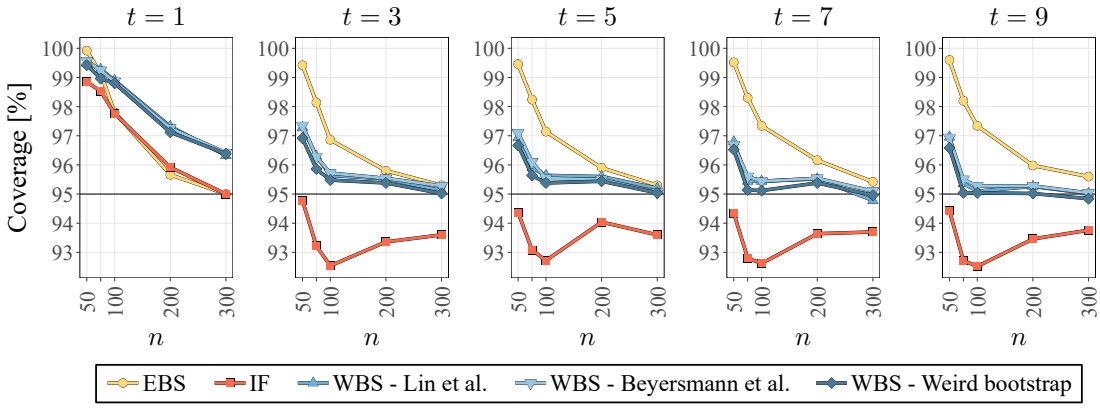


Figure B.40: Coverage of the g-formula CIs in the scenario with low treatment probability and $\beta_{01A} = 2$.

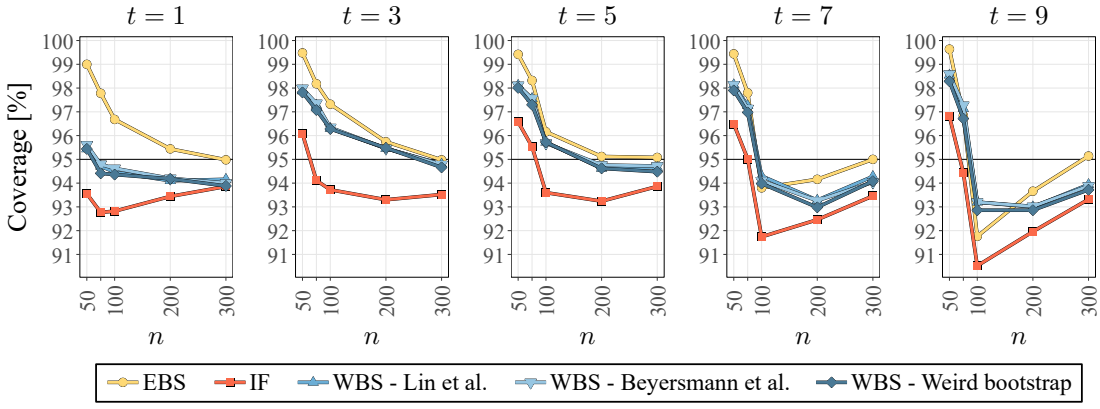


Figure B.41: Coverage of the g-formula CIs in the scenario with high treatment probability and $\beta_{01A} = -2$.

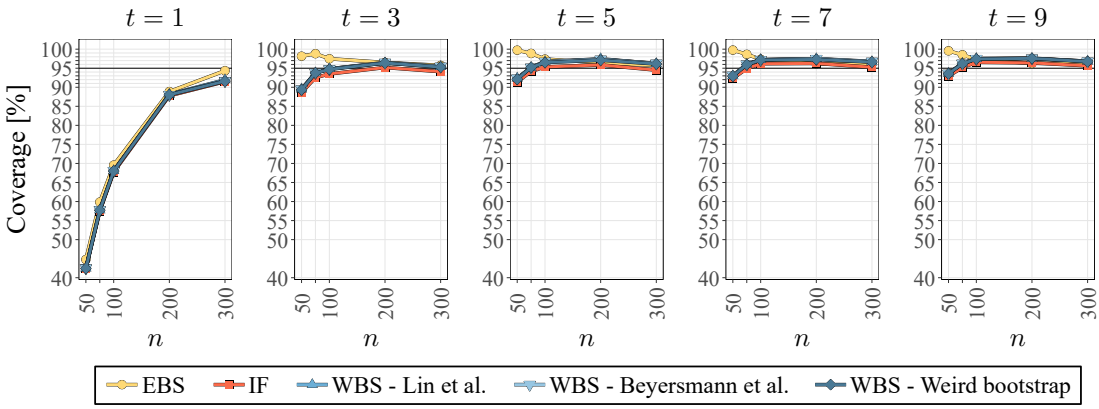


Figure B.42: Coverage of the g-formula CIs in the scenario with high treatment probability and $\beta_{01A} = 2$.

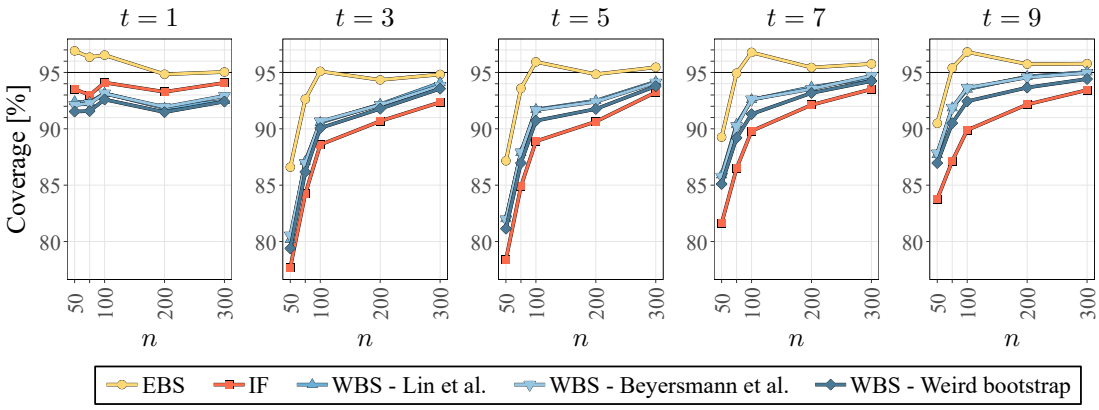


Figure B.43: Coverage of the g-formula CIs in the scenario with low variance of the covariates and $\beta_{01A} = -2$.

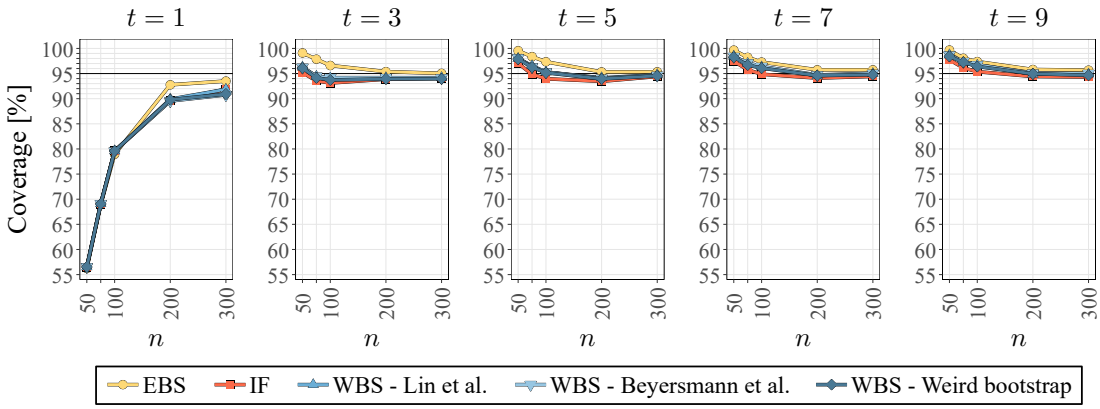


Figure B.44: Coverage of the g-formula CIs in the scenario with low variance of the covariates and $\beta_{01A} = 0$.

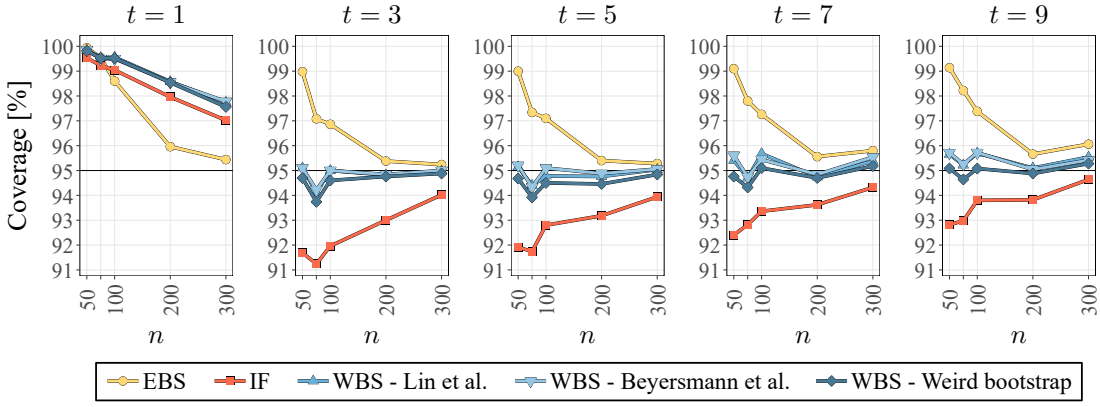


Figure B.45: Coverage of the g-formula CIs in the scenario with low variance of the covariates and $\beta_{01A} = 2$.

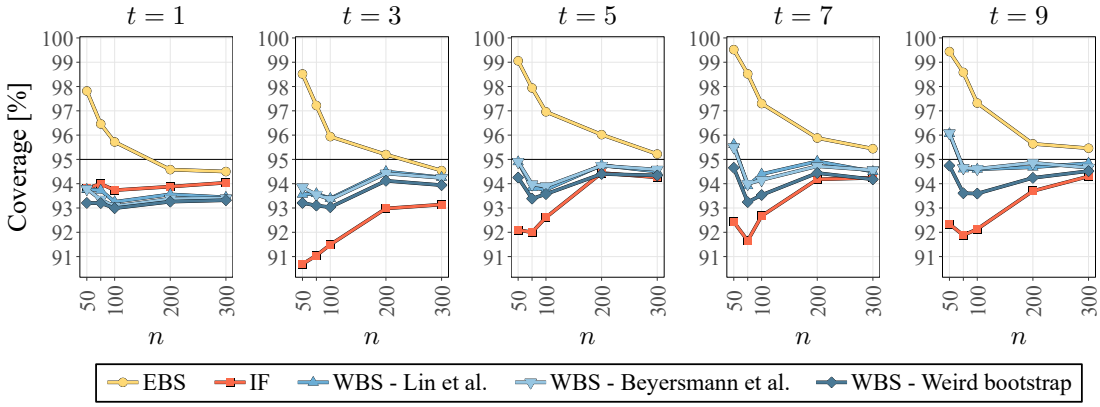


Figure B.46: Coverage of the g-formula CIs in the scenario with high variance of the covariates and $\beta_{01A} = -2$.

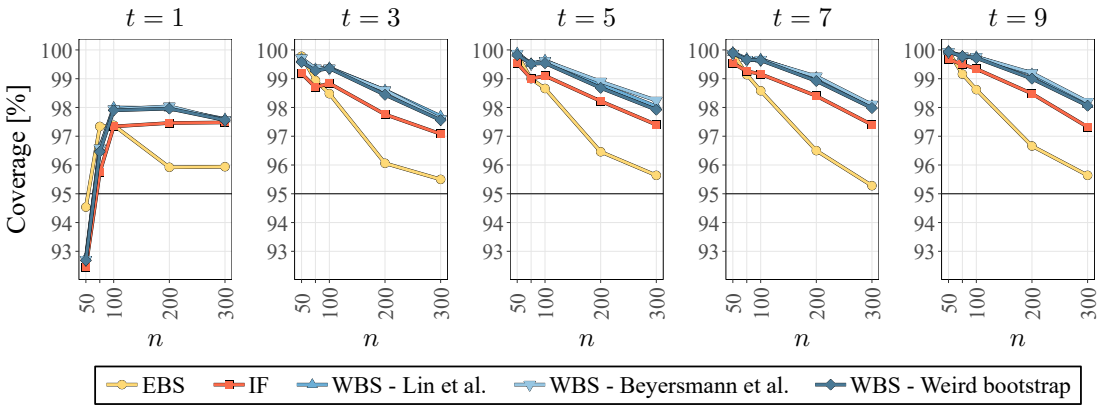


Figure B.47: Coverage of the g-formula CIs in the scenario with high variance of the covariates and $\beta_{01A} = 0$.

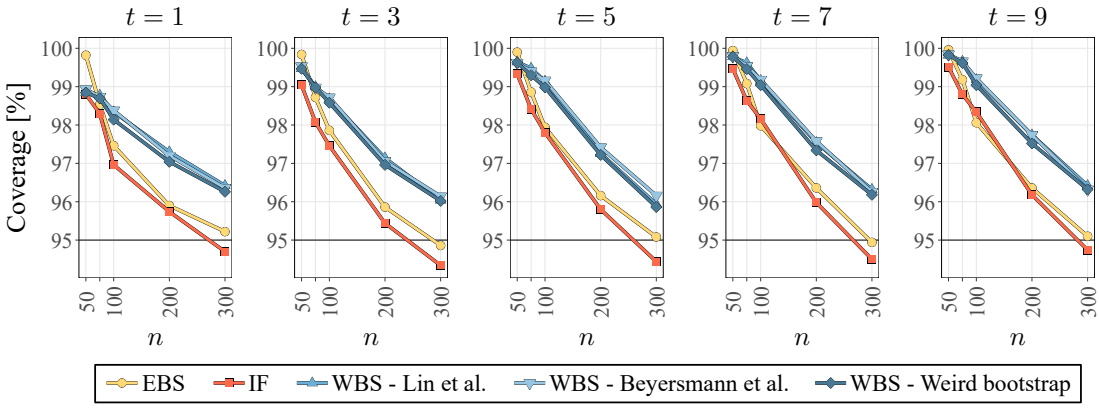


Figure B.48: Coverage of the g-formula CIs in the scenario with high variance of the covariates and $\beta_{01A} = 2$.

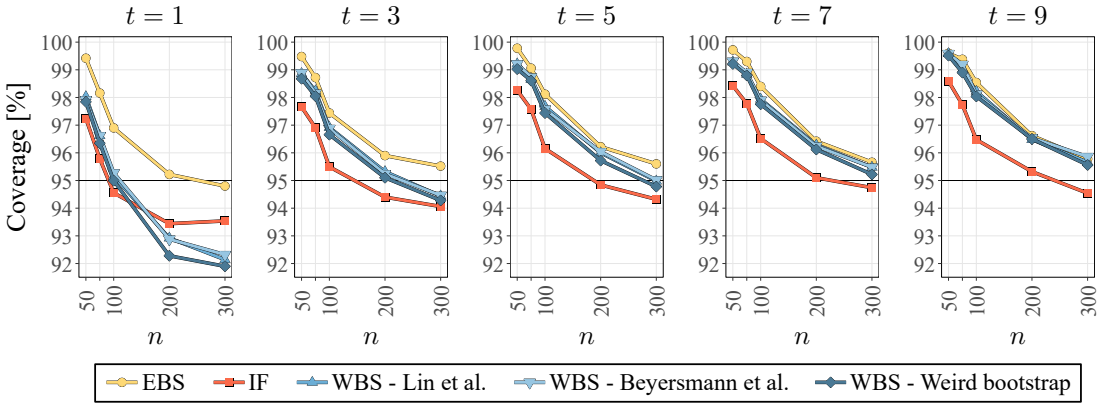


Figure B.49: Coverage of the g-formula CIs in the scenario with type II censoring and $\beta_{01A} = -2$.

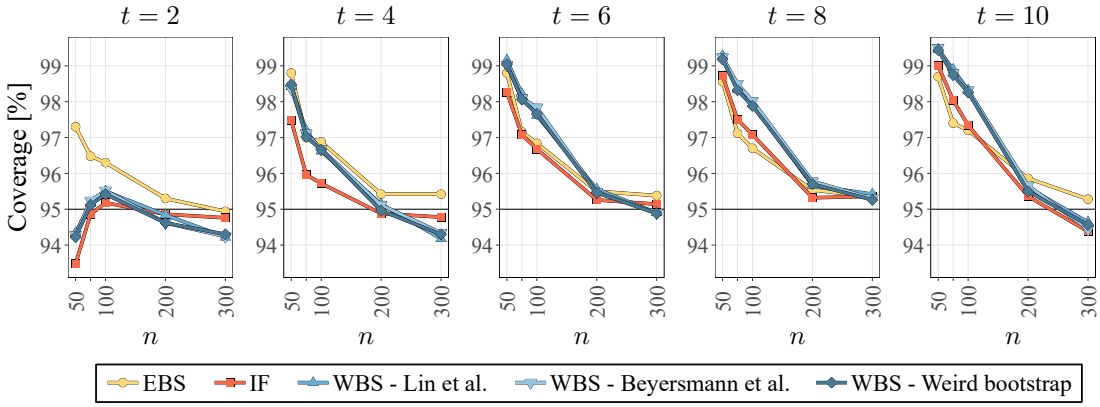


Figure B.50: Coverage of the g-formula CIs in the scenario with type II censoring and $\beta_{01A} = 0$.

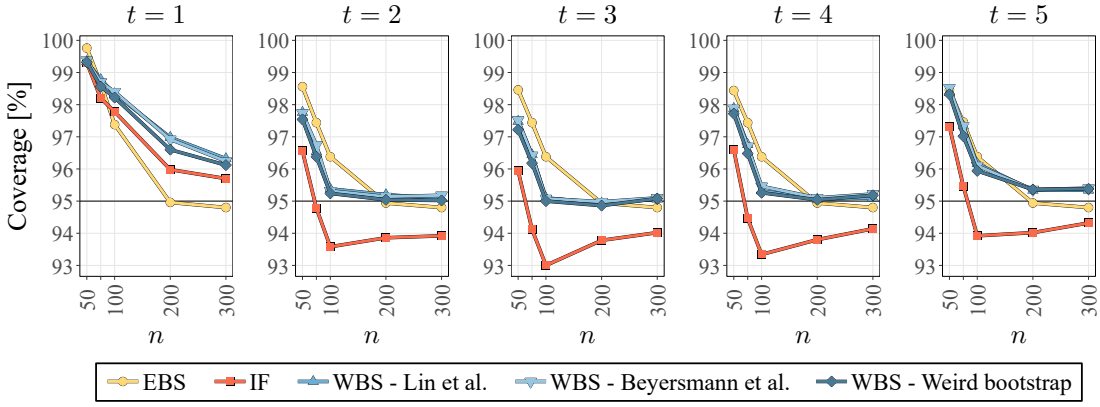


Figure B.51: Coverage of the g-formula CIs in the scenario with type II censoring and $\beta_{01A} = 2$.

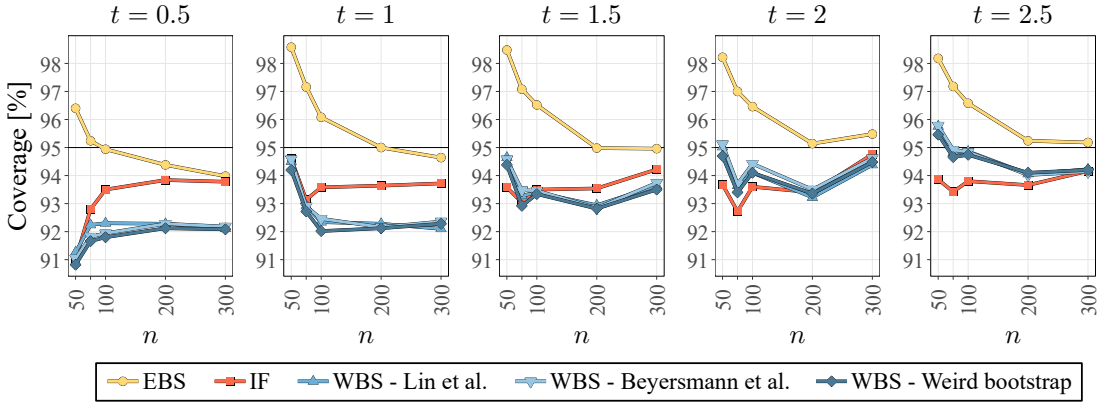


Figure B.52: Coverage of the g-formula CBs in the scenario with no censoring and $\beta_{01A} = -2$.

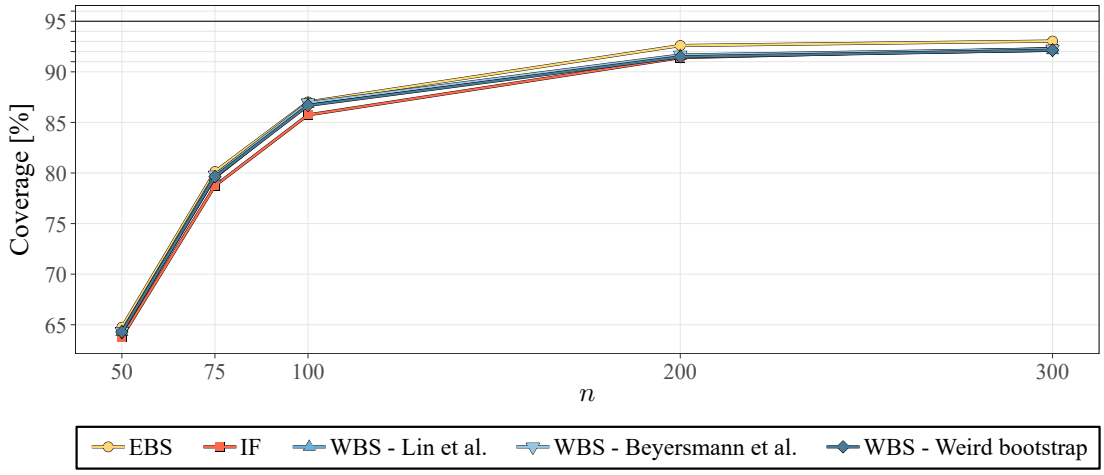


Figure B.53: Coverage of the g-formula CBs in the scenario with no censoring and $\beta_{01A} = 0$.

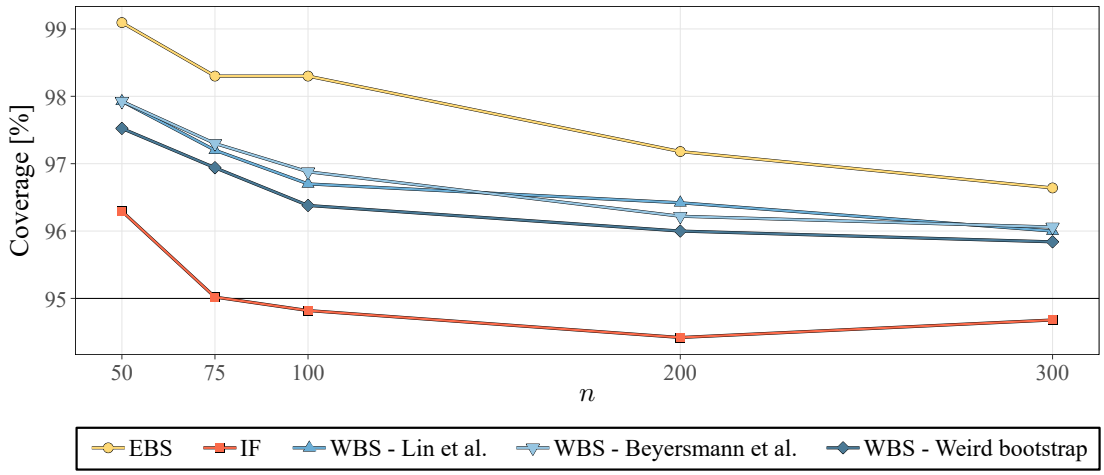


Figure B.54: Coverage of the g-formula CBs in the scenario with no censoring and $\beta_{01A} = 2$.

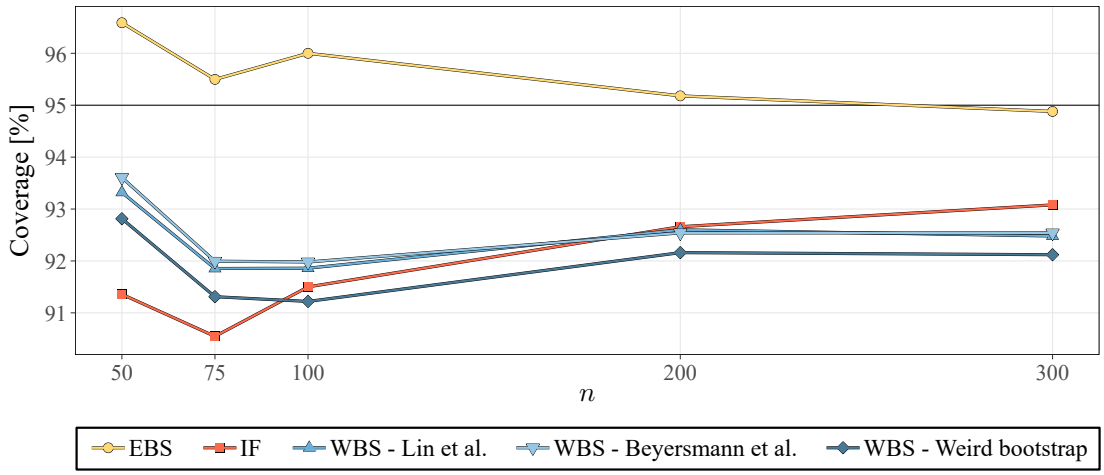


Figure B.55: Coverage of the g-formula CBs in the scenario with light censoring and $\beta_{01A} = -2$.

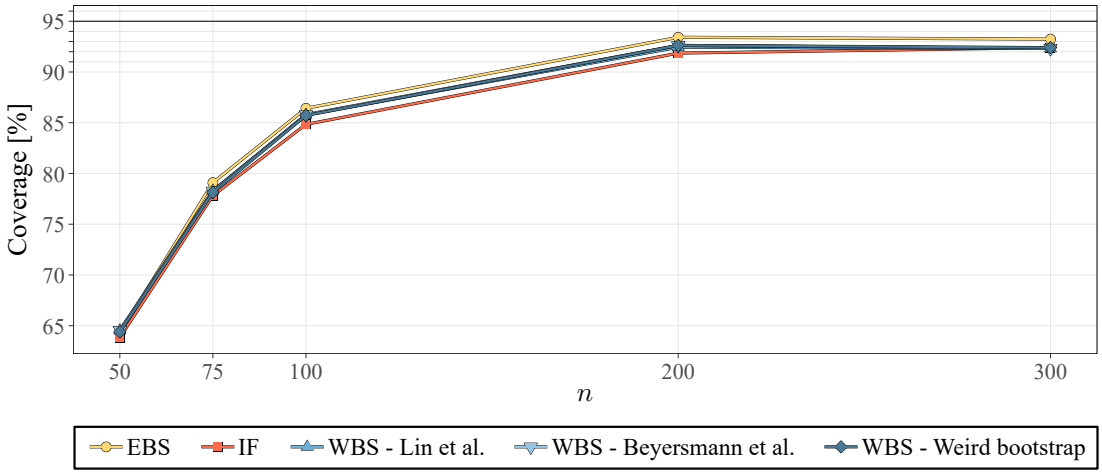


Figure B.56: Coverage of the g-formula CBs in the scenario with light censoring and $\beta_{01A} = 0$.

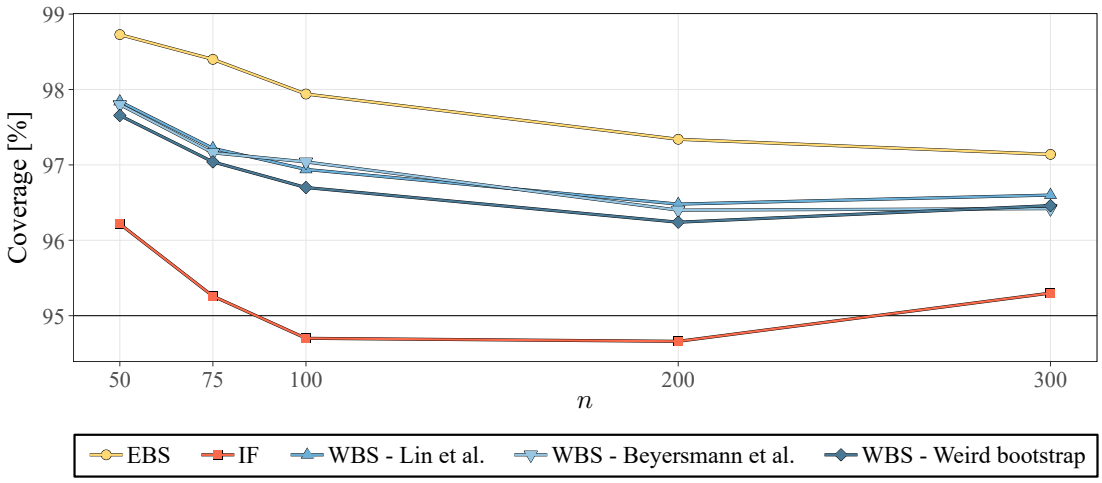


Figure B.57: Coverage of the g-formula CBs in the scenario with light censoring and $\beta_{01A} = 2$.

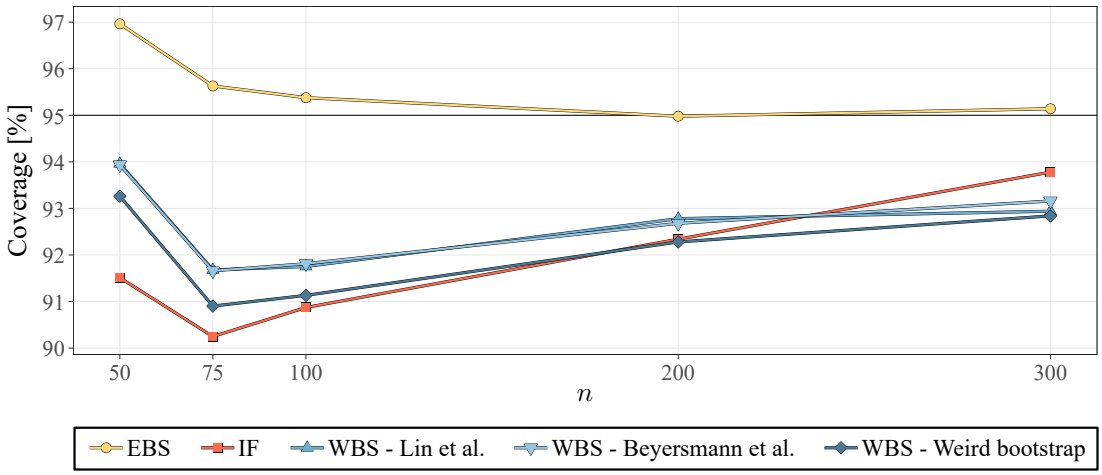


Figure B.58: Coverage of the g-formula CBs in the scenario with heavy censoring and $\beta_{01A} = 0$.

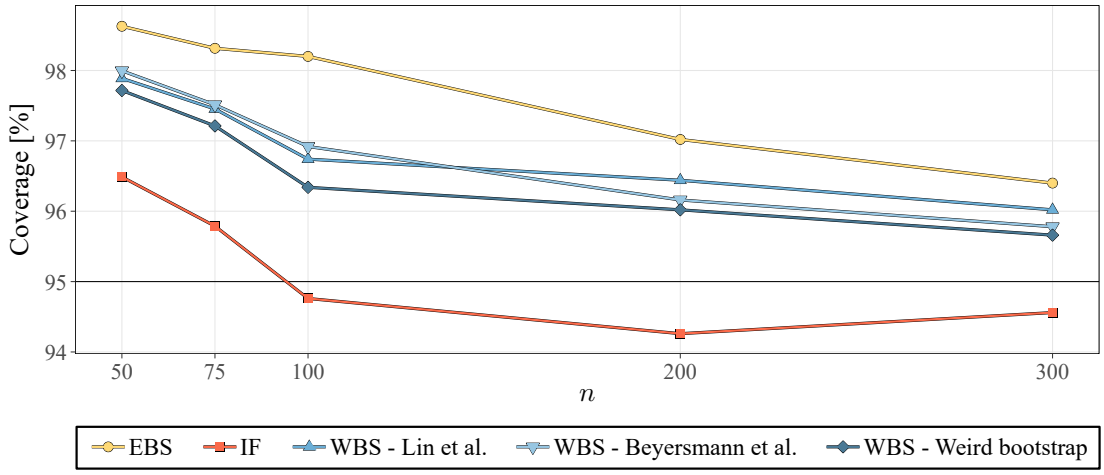


Figure B.59: Coverage of the g-formula CBs in the scenario with heavy censoring and $\beta_{01A} = 2$.

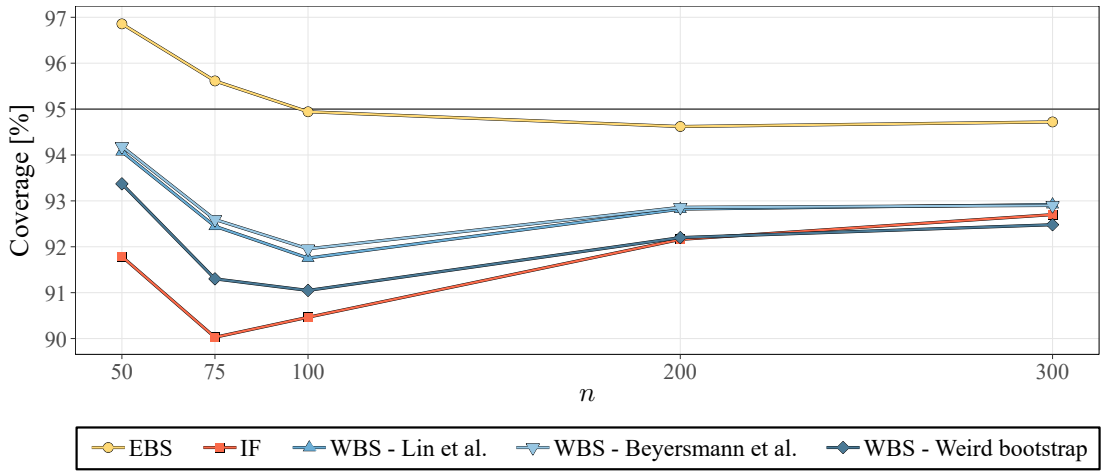


Figure B.60: Coverage of the g-formula CBs in the scenario with low treatment probability and $\beta_{01A} = -2$.

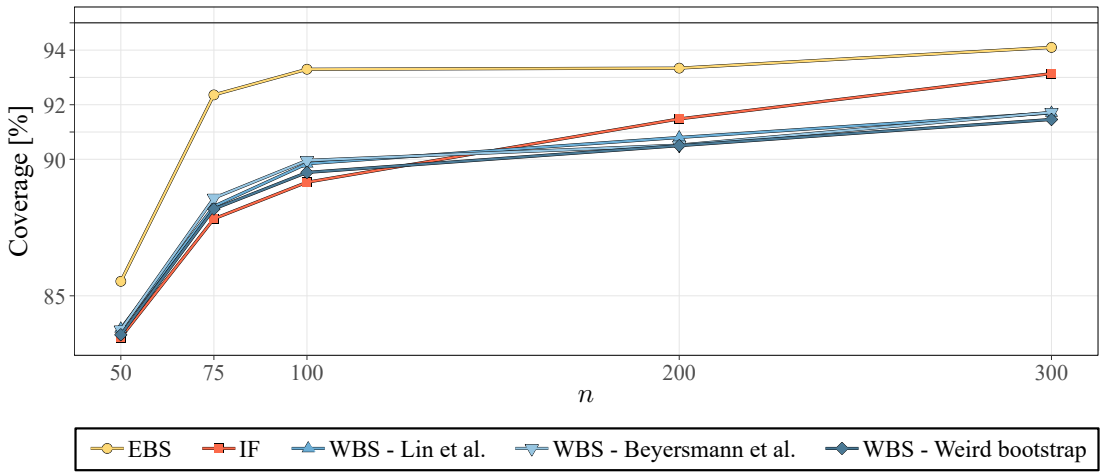


Figure B.61: Coverage of the g-formula CBs in the scenario with low treatment probability and $\beta_{01A} = 0$.

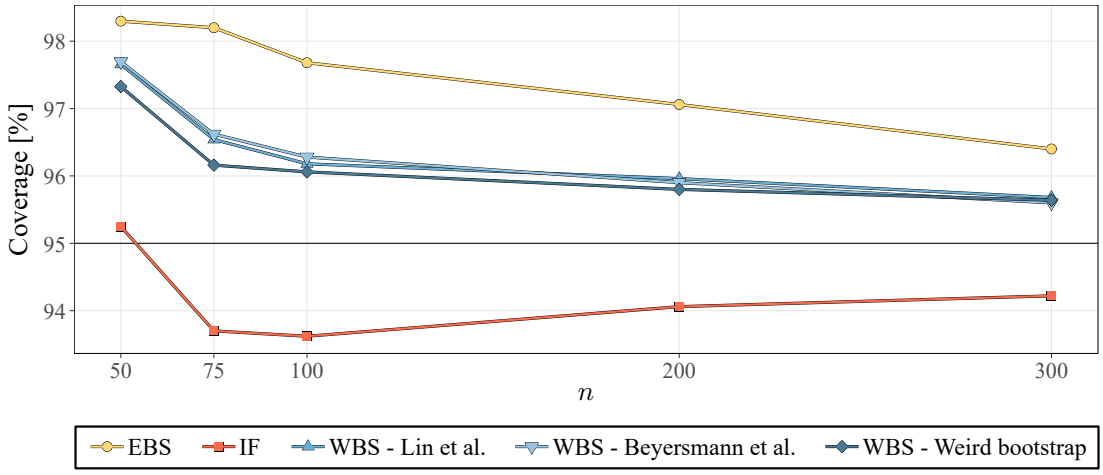


Figure B.62: Coverage of the g-formula CBs in the scenario with low treatment probability and $\beta_{01A} = 2$.

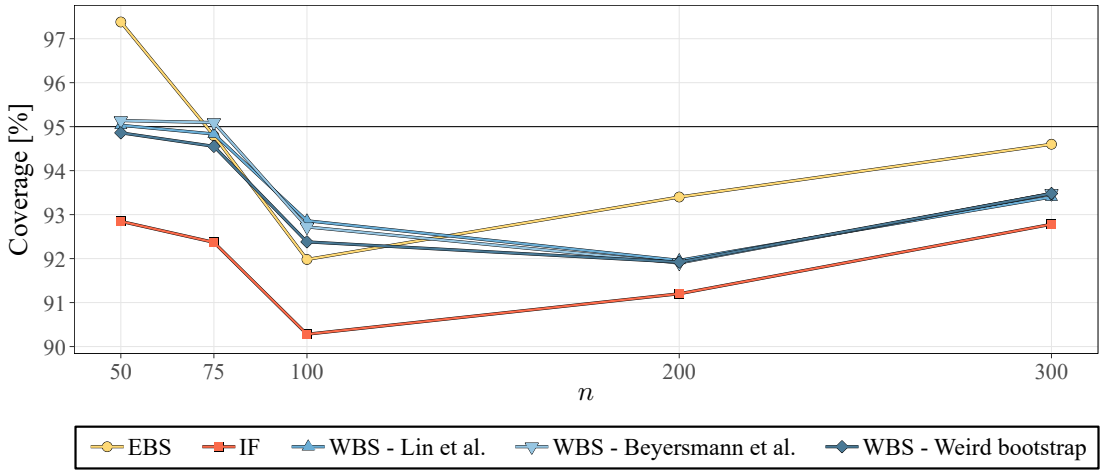


Figure B.63: Coverage of the g-formula CBs in the scenario with high treatment probability and $\beta_{01A} = -2$.

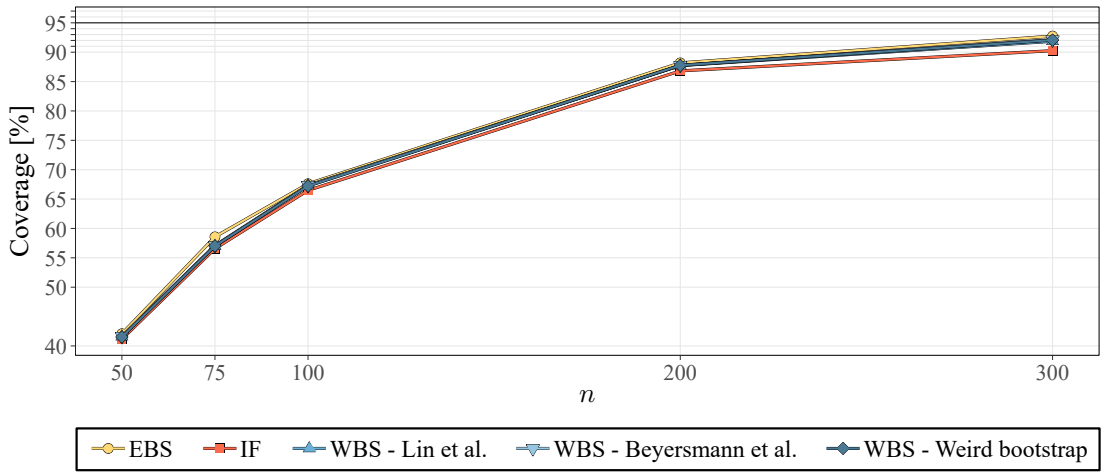


Figure B.64: Coverage of the g-formula CBs in the scenario with high treatment probability and $\beta_{01A} = 0$.

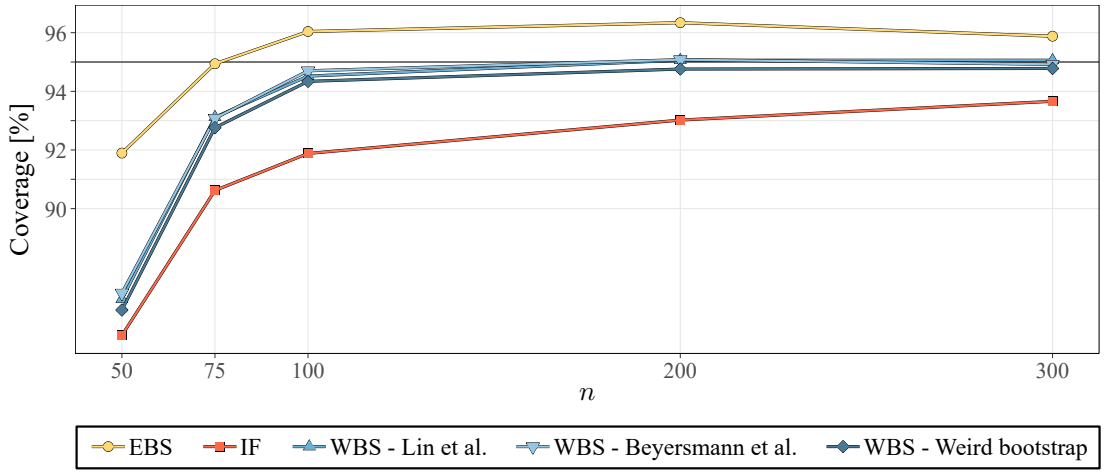


Figure B.65: Coverage of the g-formula CBs in the scenario with high treatment probability and $\beta_{01A} = 2$.

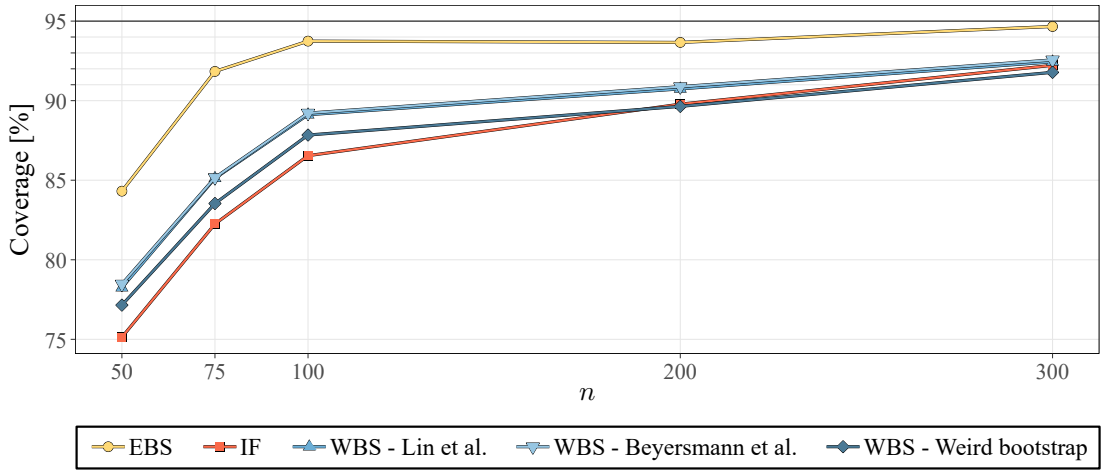


Figure B.66: Coverage of the g-formula CBs in the scenario with low variance of the covariates and $\beta_{01A} = -2$.

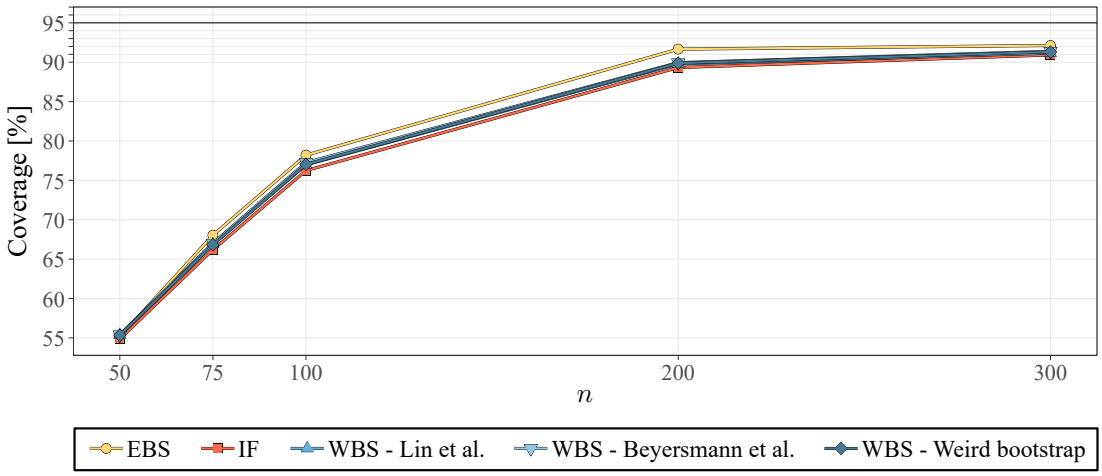


Figure B.67: Coverage of the g-formula CBs in the scenario with low variance of the covariates and $\beta_{01A} = 0$.

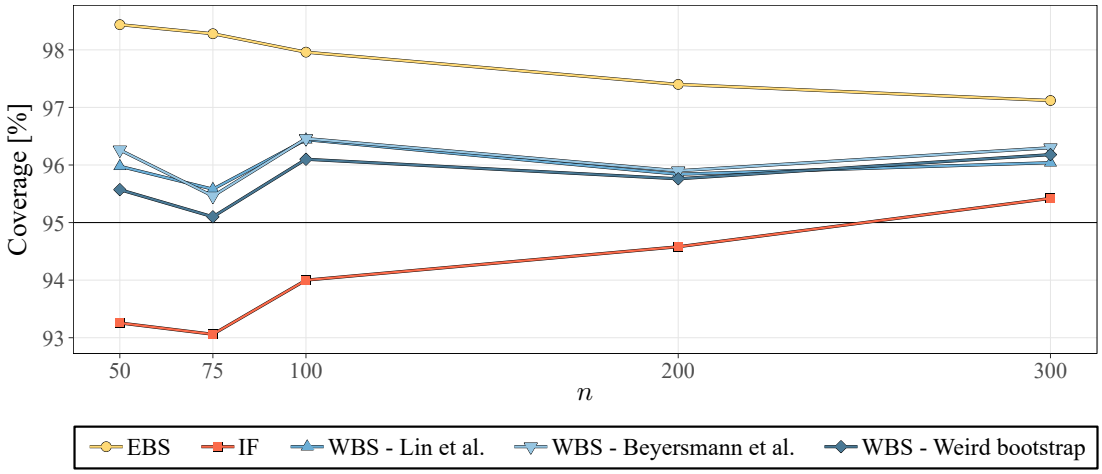


Figure B.68: Coverage of the g-formula CBs in the scenario with high variance of the covariates and $\beta_{01A} = -2$.

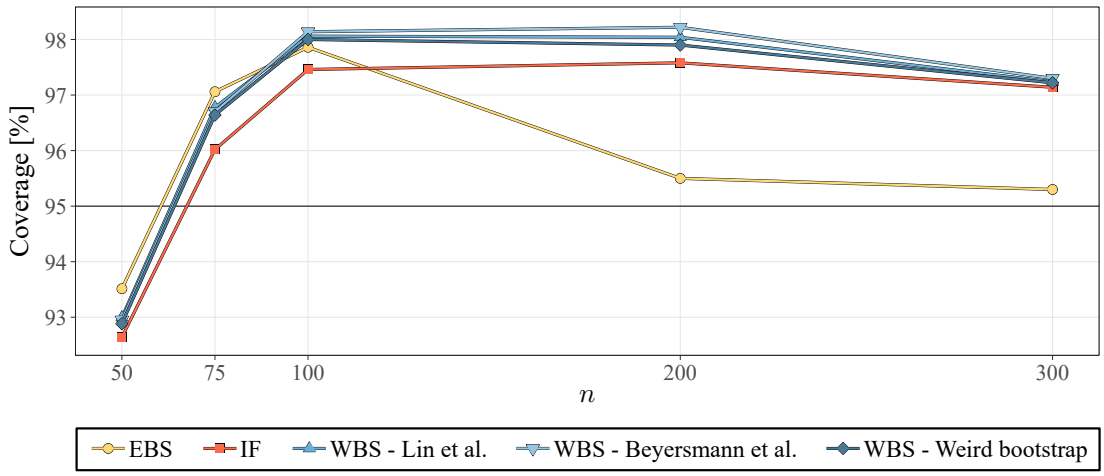


Figure B.69: Coverage of the g-formula CBs in the scenario with high variance of the covariates and $\beta_{01A} = 0$.

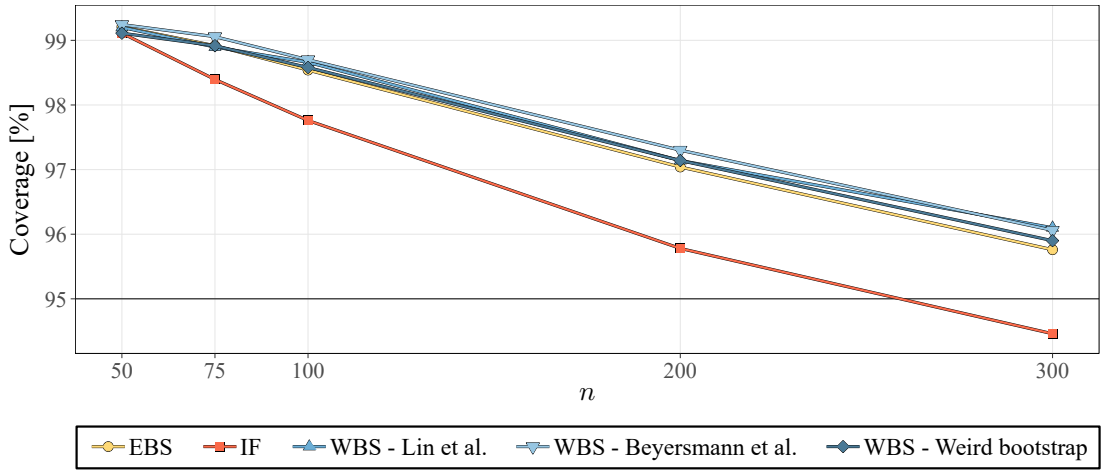


Figure B.70: Coverage of the g-formula CBs in the scenario with high variance of the covariates and $\beta_{01A} = 2$.

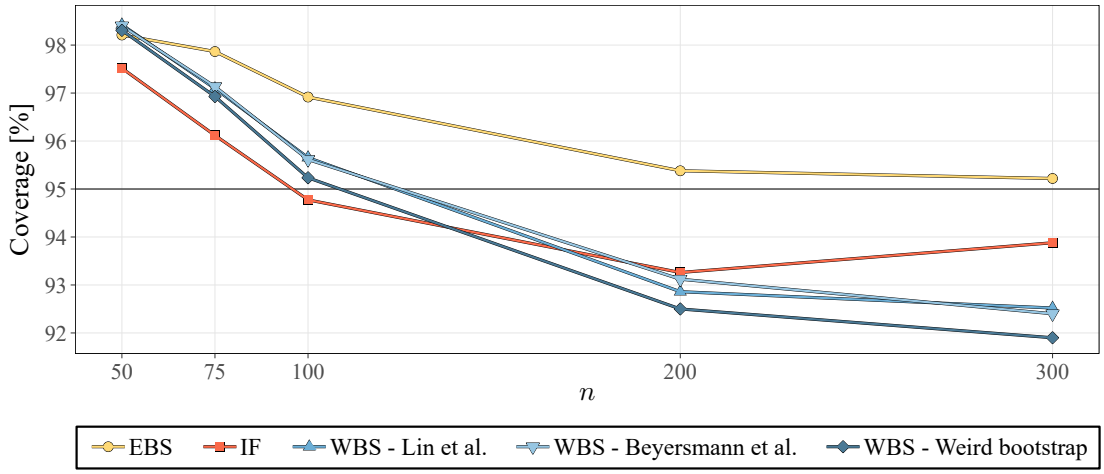


Figure B.71: Coverage of the g-formula CBs in the scenario with type II censoring and $\beta_{01A} = -2$.

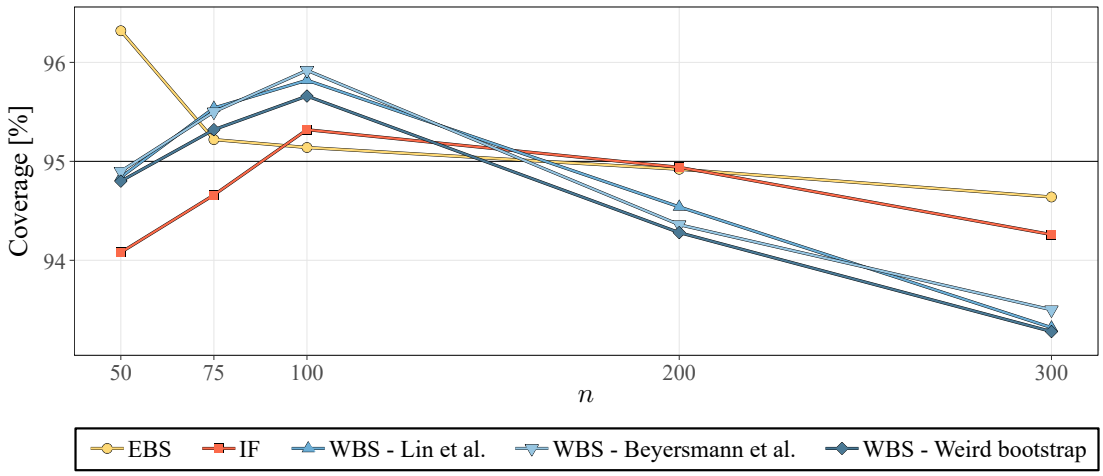


Figure B.72: Coverage of the g-formula CBs in the scenario with type II censoring and $\beta_{01A} = 0$.

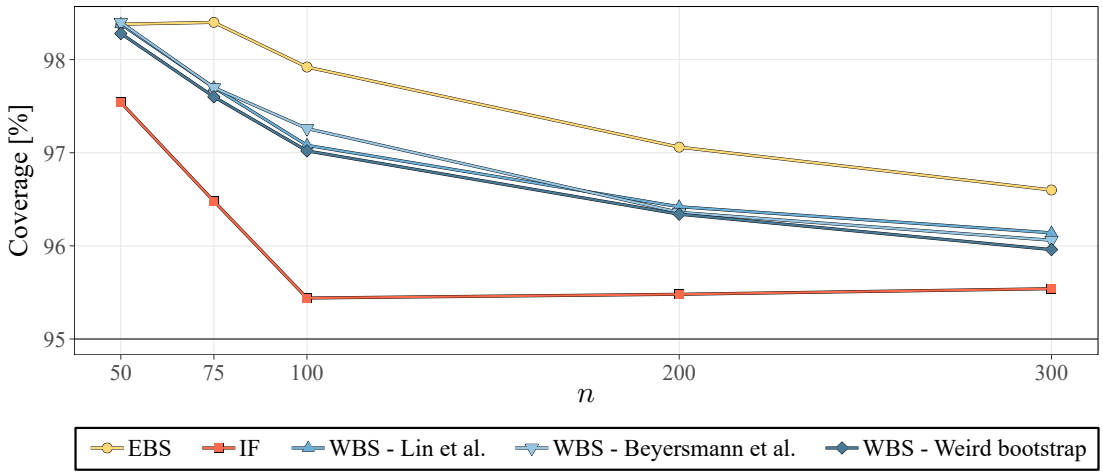
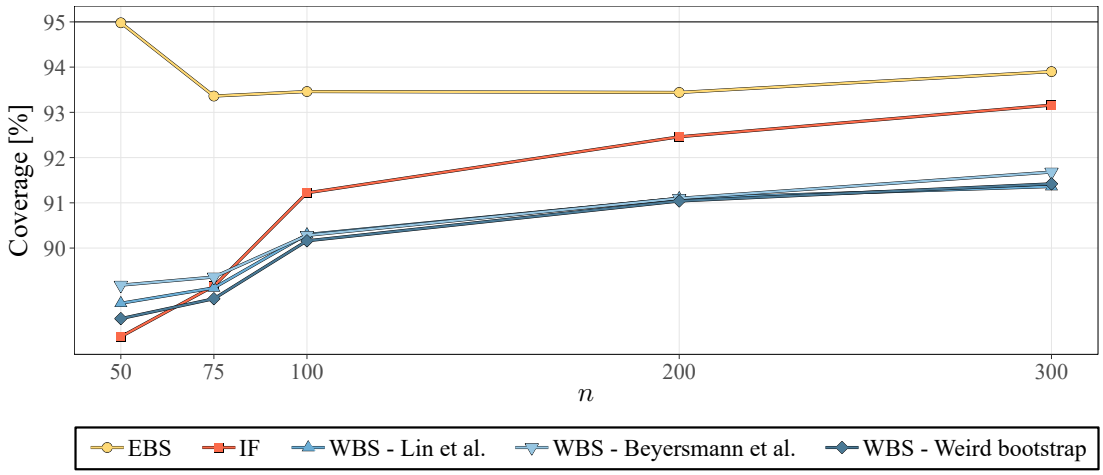


Figure B.73: Coverage of the g-formula CBs in the scenario with type II censoring and $\beta_{01A} = 2$.



In the following, we present the results of the tests on the scaled Schoenfeld residuals that were performed within the scope of the data analysis in Subsection 4.1.3.

Figure B.74: Test of the proportional hazards assumption for the Cox model w.r.t. relapse.

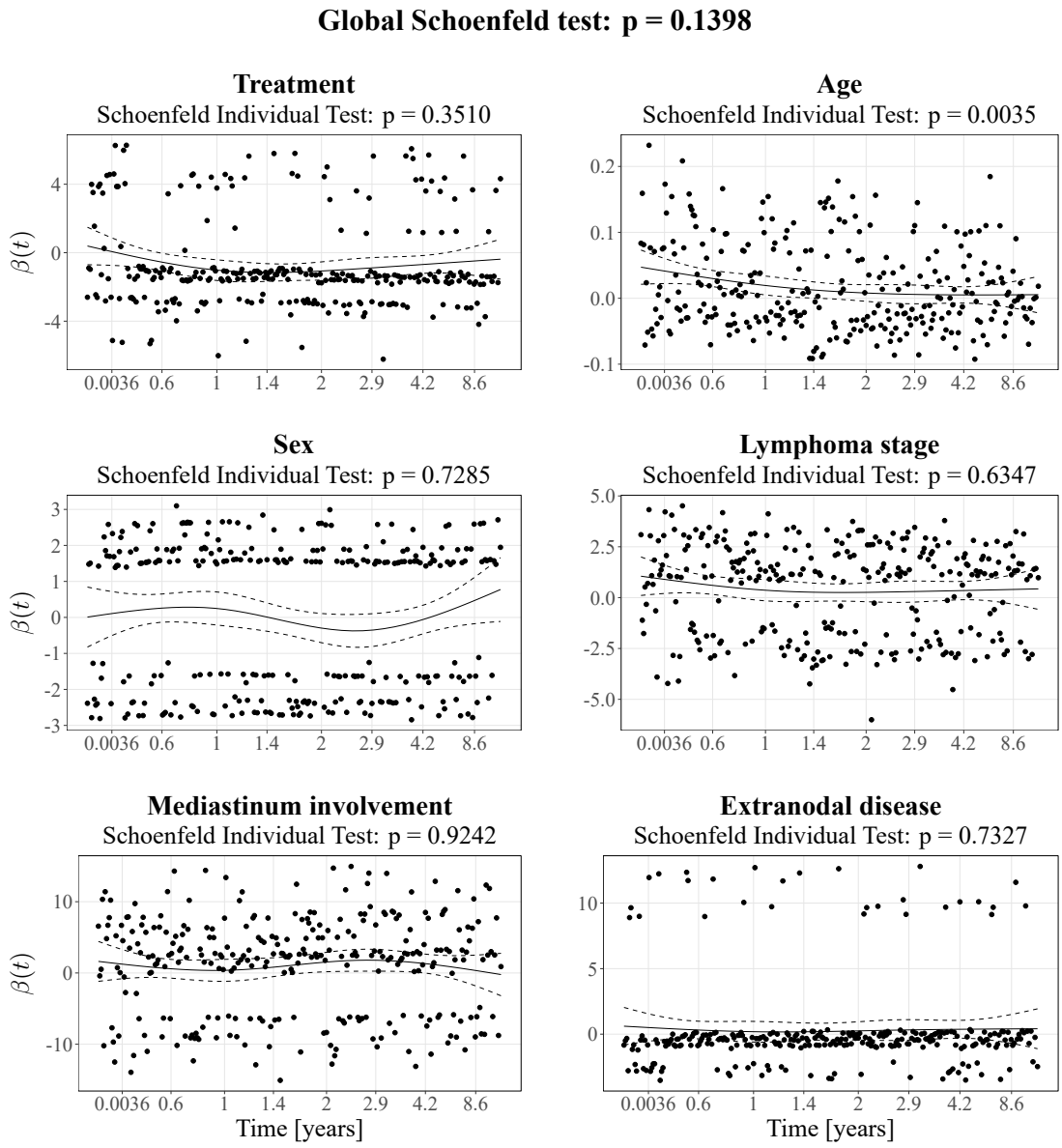
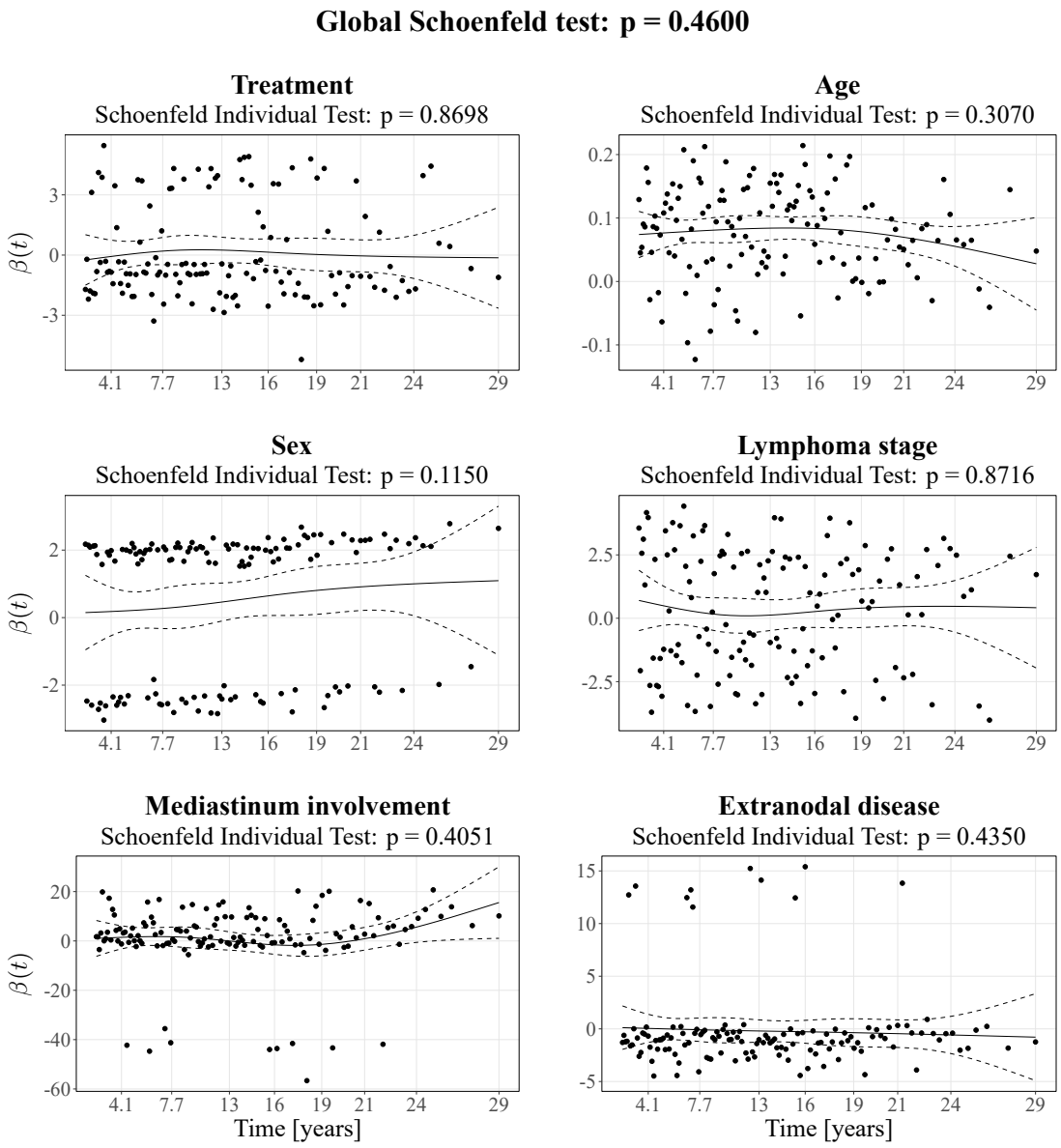


Figure B.75: Test of the proportional hazards assumption for the Cox model w.r.t. death.



B.2. Resampling-based inference for the ATE using PS matching

The table and the figures below complement the outcomes of the simulation study described in Subsection 4.2.1.

Table B.9: Frequency of errors and convergence issues for the simulation study that investigates the resampling methods on the basis of PS matching.

Scenario	n	β_{01A}	Iterations with errors (out of 5,000)	Iterations with convergence issues (out of 5,000)	mean number of invalid EBS samples (out of 1,000)				
					$t=1$	$t=3$	$t=5$	$t=7$	$t=9$
No censoring	50	-2	0	174	249.1	517.8	525.3	525.2	506.5
		0	0	169	449.1	525.5	525.5	525.0	467.3
		2	0	153	525.3	525.7	525.7	516.0	364.1
	75	-2	0	0	38.4	64.7	64.8	64.8	64.4
		0	0	0	61.4	65.1	65.1	65.1	62.7
		2	0	0	65.1	65.1	65.1	64.9	55.1
	100	-2	0	0	2.3	3.3	3.3	3.3	3.3
		0	0	0	3.2	3.3	3.3	3.3	3.2
		2	0	0	3.3	3.3	3.3	3.3	3.0
Light censoring	50	-2	0	164	257.9	520.2	527.0	526.5	467.9
		0	0	188	454.3	532.3	532.3	528.5	381.2
		2	0	123	527.9	528.3	528.2	489.1	232.6
	75	-2	0	2	37.6	63.5	63.7	63.7	61.6
		0	0	2	59.5	63.2	63.2	63.1	54.4
		2	0	1	64.3	64.3	64.3	62.8	38.5
	100	-2	0	0	2.1	3.1	3.1	3.1	3.0
		0	0	0	3.1	3.1	3.1	3.1	2.9
		2	0	0	3.2	3.2	3.2	3.2	2.4
Heavy censoring	50	-2	0	169	246.6	517.1	526.3	518.6	360.3
		0	0	124	453.3	537.4	537.4	504.8	226.3
		2	1	56	531.4	531.9	528.9	385.3	95.2
	75	-2	0	1	39.0	65.8	66.1	66.0	55.9
		0	0	0	60.3	63.9	63.9	62.7	36.6
		2	0	1	67.1	67.1	67.1	56.5	17.0
	100	-2	0	0	2.2	3.2	3.3	3.3	3.0
		0	0	0	2.9	3.0	3.0	3.0	2.2
		2	0	0	3.1	3.1	3.1	2.8	1.2

Scenario	n	β_{01A}	Iterations with errors (out of 5,000)	Iterations with convergence issues (out of 5,000)	mean number of invalid EBS samples (out of 1,000)				
					$t=1$	$t=3$	$t=5$	$t=7$	$t=9$
Low treatment probability	50	-2	0	1,122	566.1	801.1	801.6	799.3	657.5
		0	0	1,067	679.0	801.4	801.4	795.8	588.8
		2	0	1,021	788.0	802.2	802.1	790.4	538.6
	75	-2	0	85	266.2	314.8	314.8	314.7	291.6
		0	0	54	288.1	307.2	307.2	307.1	268.9
		2	0	64	307.3	308.3	308.3	307.8	256.0
	100	-2	0	1	57.2	62.1	62.1	62.1	60.1
		0	0	0	62.1	63.7	63.7	63.7	59.7
		2	0	0	64.3	64.3	64.3	64.3	59.1
High treatment probability	50	-2	3	2,737	269.0	842.9	908.0	911.9	839.4
		0	2	2,443	749.7	913.6	913.6	907.6	670.1
		2	5	883	884.2	884.5	882.7	688.8	190.6
	75	-2	0	635	253.8	592.8	606.6	606.9	594.2
		0	0	599	572.5	610.9	610.9	610.7	530.9
		2	0	262	597.3	597.3	597.3	528.6	187.8
	100	-2	0	93	135.9	277.7	279.8	279.8	278.1
		0	0	83	265.0	271.6	271.6	271.6	253.5
		2	0	42	270.7	270.7	270.7	254.2	109.7
	200	-2	0	0	0.9	1.3	1.3	1.3	1.3
		0	0	0	1.3	1.3	1.3	1.3	1.3
		2	0	0	1.4	1.4	1.4	1.4	1.0
Low variance of the covariates	50	-2	0	30	130.7	338.0	348.1	348.3	330.5
		0	0	41	250.4	356.6	356.7	356.3	292.3
		2	0	28	356.1	357.4	357.4	343.3	192.0
	75	-2	0	0	7.7	15.7	15.8	15.8	15.6
		0	0	0	13.5	16.2	16.2	16.2	15.1
		2	0	0	17.4	17.4	17.4	17.1	12.1
	100	-2	0	0	0.1	0.2	0.2	0.2	0.2
		0	0	0	0.3	0.3	0.3	0.3	0.3
		2	0	0	0.3	0.3	0.3	0.3	0.3

Scenario	n	β_{01A}	Iterations with errors (out of 5,000)	Iterations with convergence issues (out of 5,000)	mean number of invalid EBS samples (out of 1,000)				
					$t=1$[a]	$t=3$[a]	$t=5$[a]	$t=7$[a]	$t=9$[a]
High variance of the covariates	50	-2	0	1,236	699.9	834.6	835.8	812.7	559.0
		0	0	1,019	830.2	839.0	838.7	778.1	402.8
		2	0	663	835.7	835.8	829.9	655.1	221.9
	75	-2	0	86	335.7	360.0	360.0	359.0	305.1
		0	0	88	353.0	353.4	353.4	347.0	238.4
		2	0	61	357.6	357.6	357.4	326.6	156.3
	100	-2	0	2	76.6	79.1	79.1	79.0	72.6
		0	0	1	80.4	80.4	80.4	80.1	64.2
		2	0	5	86.2	86.2	86.2	82.9	47.7
Type II censoring	50	-2	0	116	481.6	528.8	525.2	496.9	256.4
		0	0	163	450.6	526.6	520.5	508.6	377.7
		2	0	140	448.9	489.2	483.9	464.4	332.5
	75	-2	0	1	61.6	63.9	63.9	63.1	40.4
		0	0	1	60.4	63.8	63.7	63.5	55.6
		2	0	1	59.8	60.9	60.8	60.3	50.8
	100	-2	0	0	3.2	3.3	3.3	3.3	2.1
		0	0	0	3.3	3.3	3.3	3.3	2.9
		2	0	0	3.2	3.2	3.2	3.2	2.7

[a] For the scenario with type II censoring, we considered $t \in \{2, 4, 6, 8, 10\}$, $t \in \{1, 2, 3, 4, 5\}$, and $t \in \{0.5, 1, 1.5, 2, 2.5\}$ for $\beta_{01A} = -2, 0, 2$, respectively.

Figure B.76: Coverage of the PS-matched CIs in the scenario with no censoring and $\beta_{01A} = -2$.

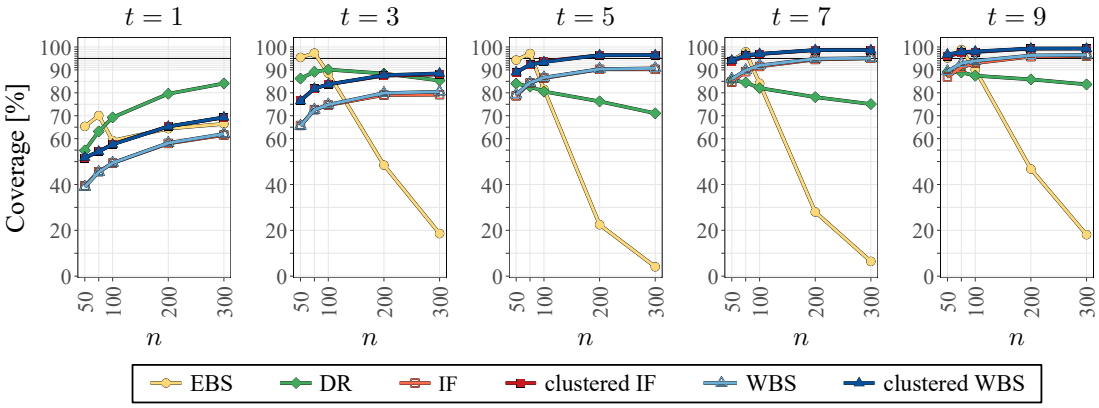


Figure B.77: Coverage of the PS-matched CIs in the scenario with no censoring and $\beta_{01A} = 0$.

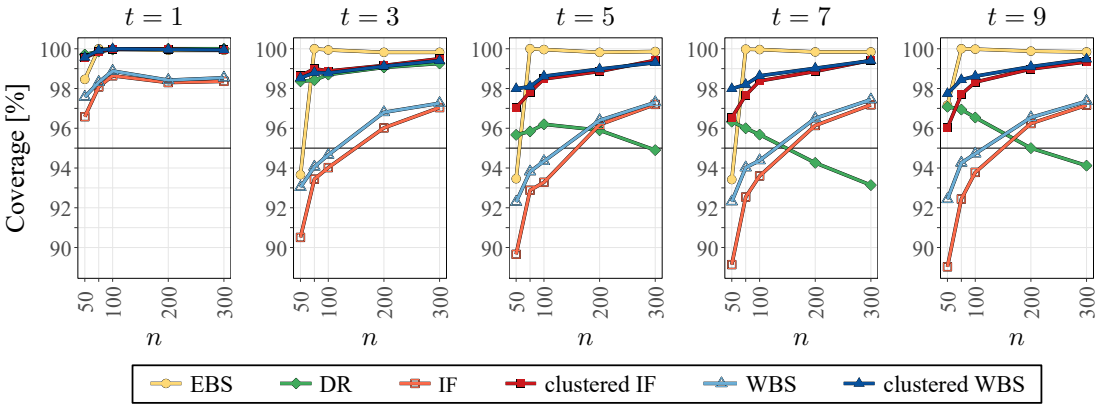


Figure B.78: Coverage of the PS-matched CIs in the scenario with no censoring and $\beta_{01A} = 2$.

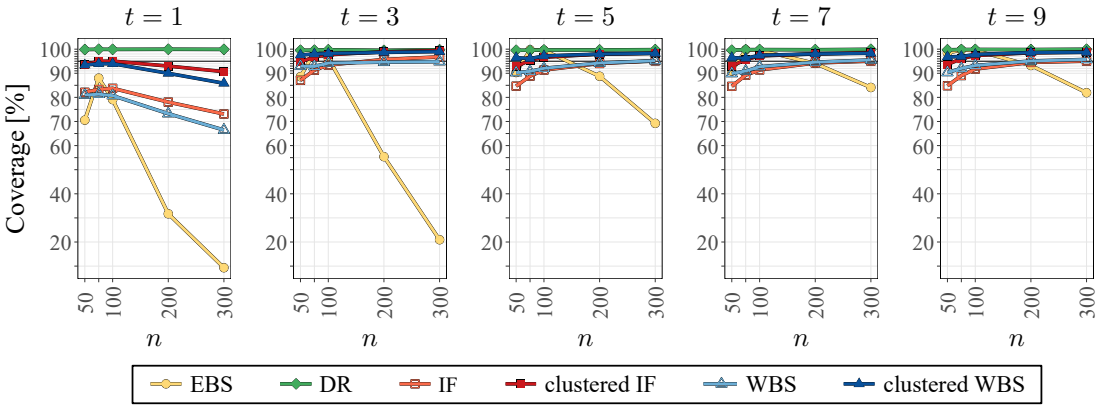


Figure B.79: Coverage of the PS-matched CIs in the scenario with light censoring and $\beta_{01A} = 0$.

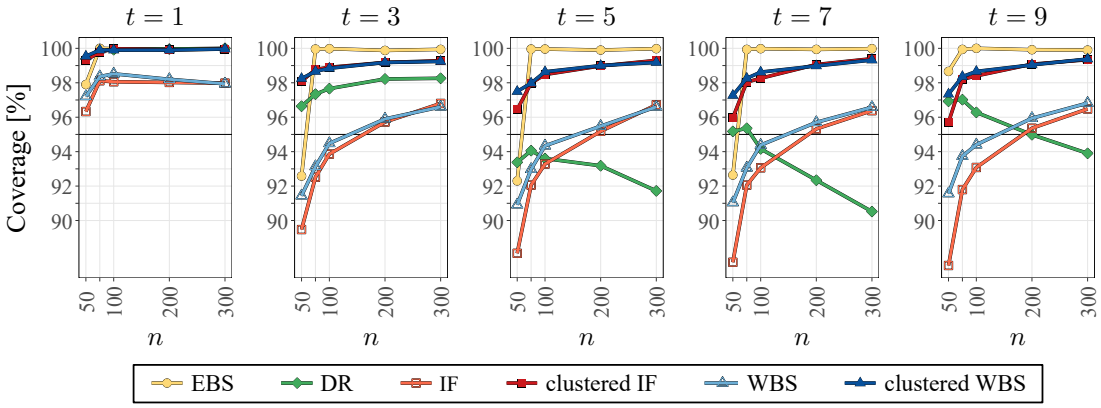


Figure B.80: Coverage of the PS-matched CIs in the scenario with light censoring and $\beta_{01A} = 2$.

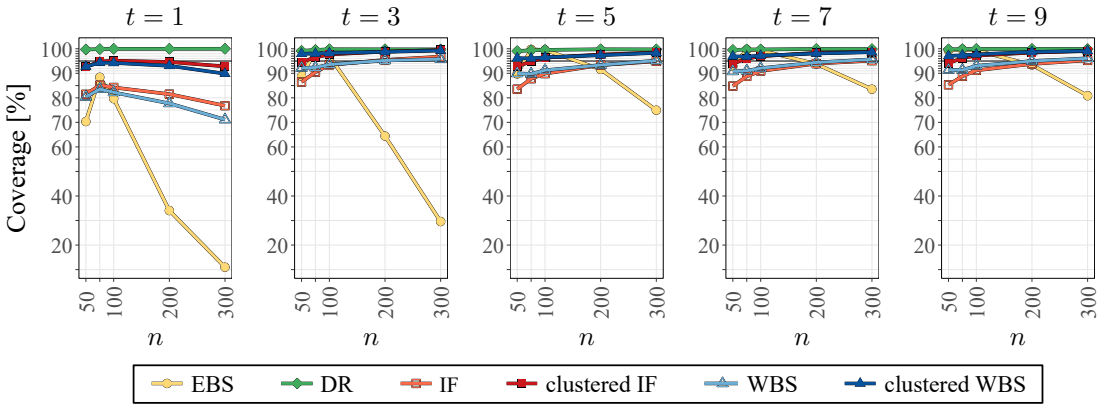


Figure B.81: Coverage of the PS-matched CIs in the scenario with heavy censoring and $\beta_{01A} = -2$.

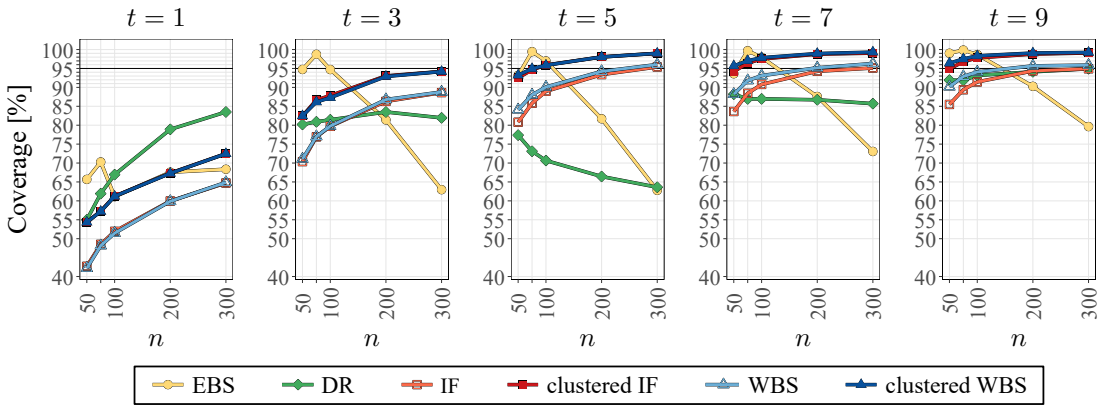


Figure B.82: Coverage of the PS-matched CIs in the scenario with heavy censoring and $\beta_{01A} = 2$.

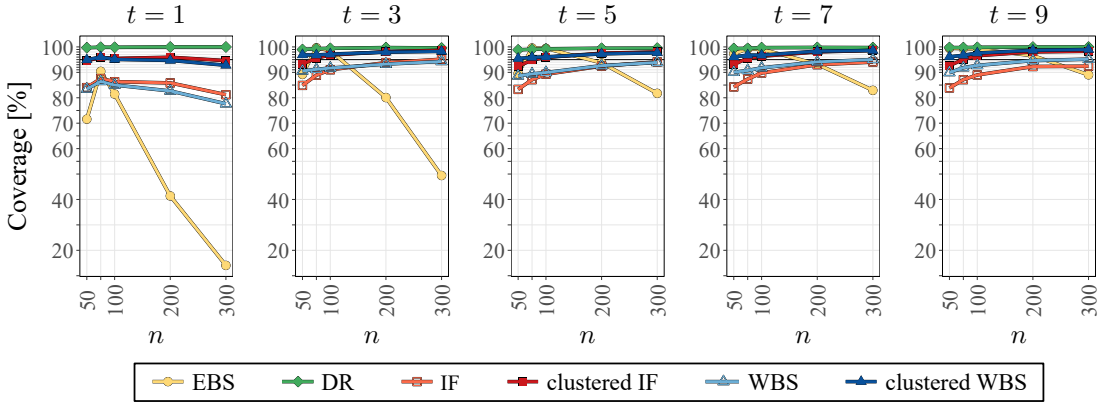


Figure B.83: Coverage of the PS-matched CIs in the scenario with low treatment probability and $\beta_{01A} = -2$.

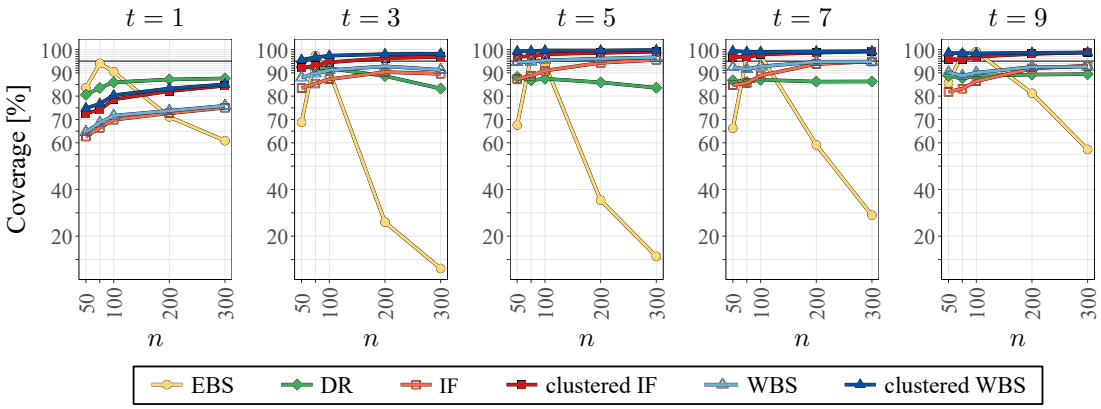


Figure B.84: Coverage of the PS-matched CIs in the scenario with low treatment probability and $\beta_{01A} = 0$.

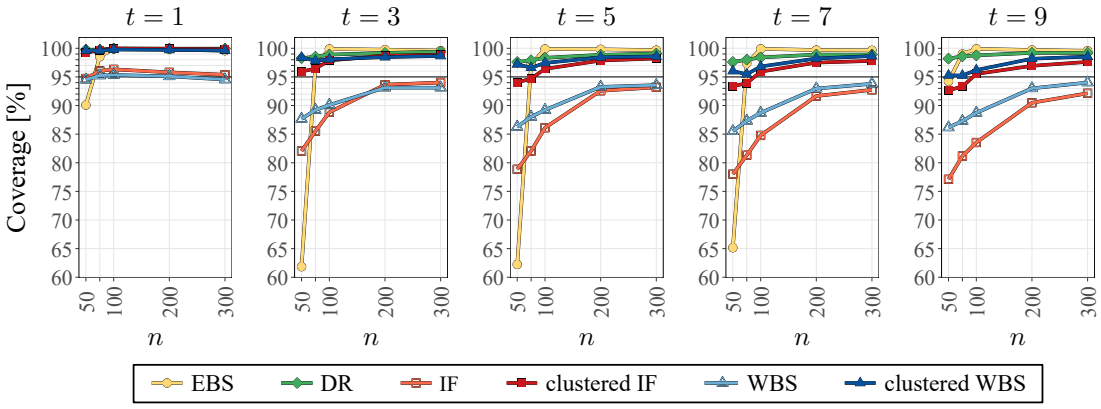


Figure B.85: Coverage of the PS-matched CIs in the scenario with low treatment probability and $\beta_{01A} = 2$.

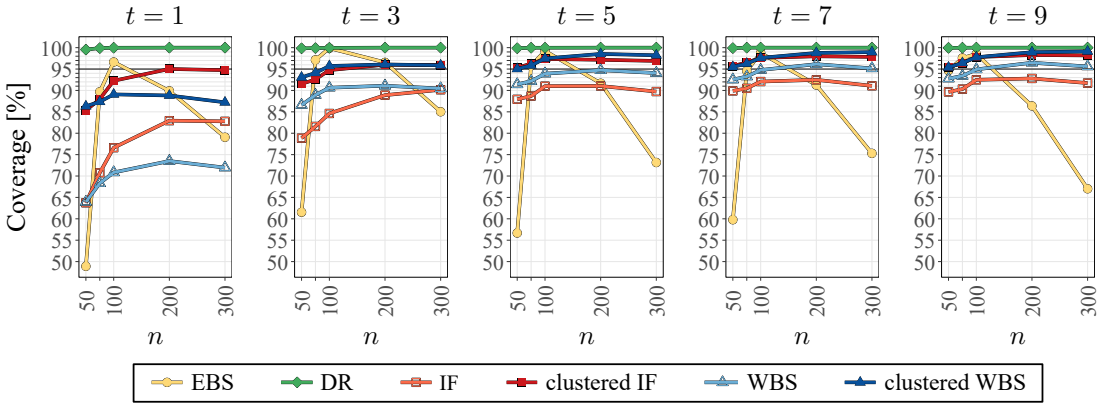


Figure B.86: Coverage of the PS-matched CIs in the scenario with high treatment probability and $\beta_{01A} = -2$.

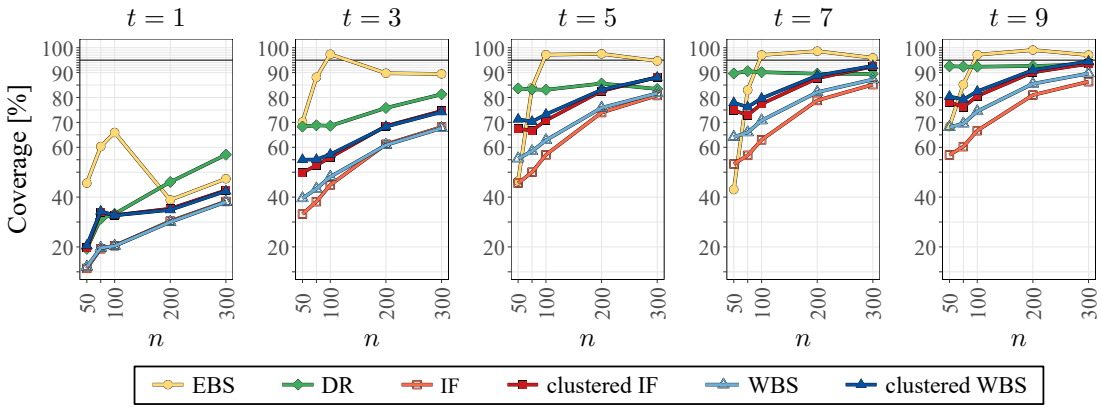


Figure B.87: Coverage of the PS-matched CIs in the scenario with high treatment probability and $\beta_{01A} = 0$.

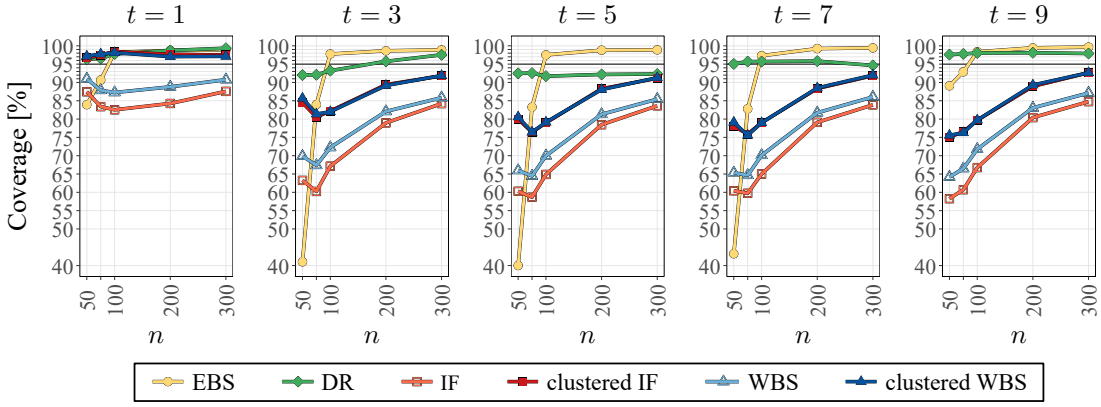


Figure B.88: Coverage of the PS-matched CIs in the scenario with low variance of the covariates and $\beta_{01A} = -2$.

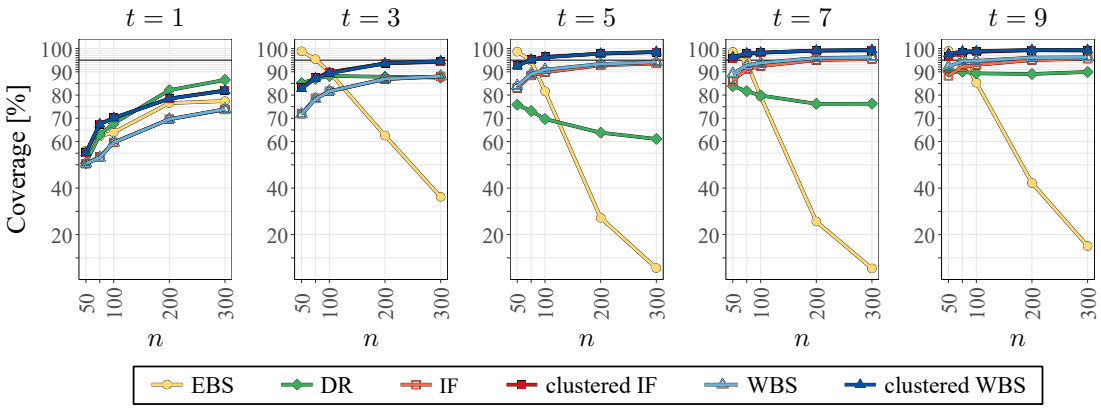


Figure B.89: Coverage of the PS-matched CIs in the scenario with low variance of the covariates and $\beta_{01A} = 0$.

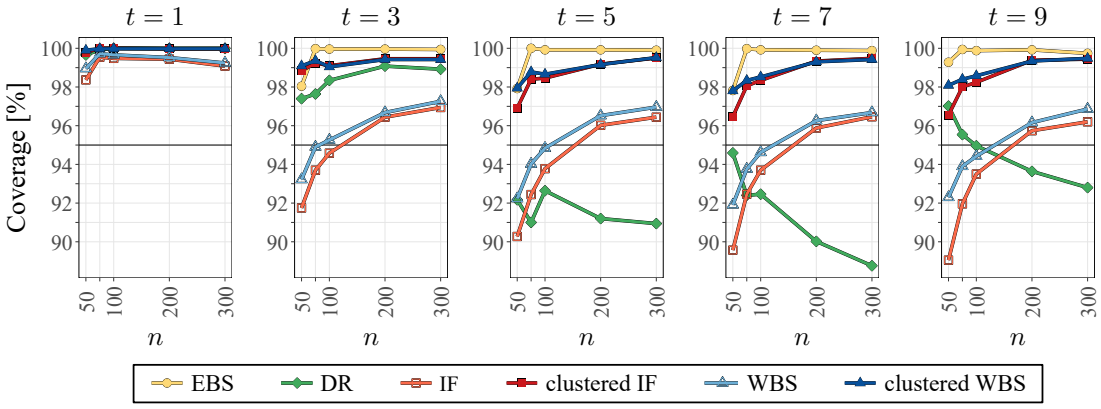


Figure B.90: Coverage of the PS-matched CIs in the scenario with low variance of the covariates and $\beta_{01A} = 2$.

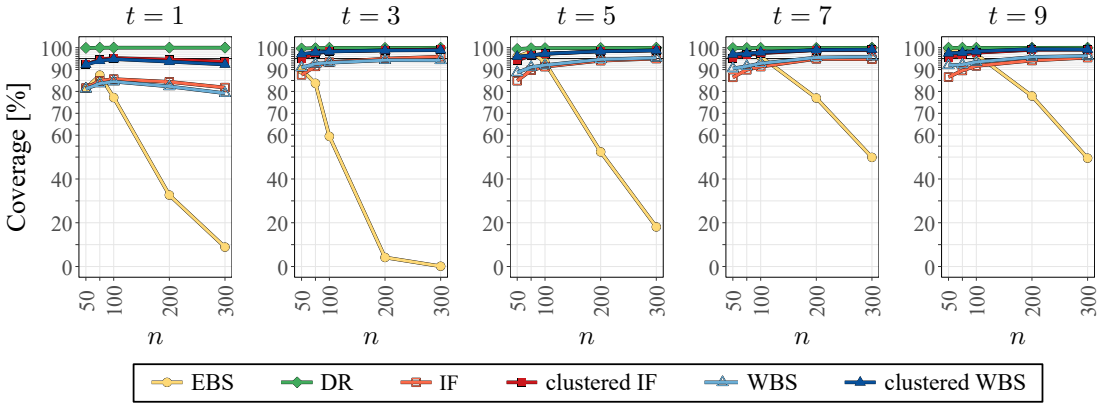


Figure B.91: Coverage of the PS-matched CIs in the scenario with high variance of the covariates and $\beta_{01A} = -2$.

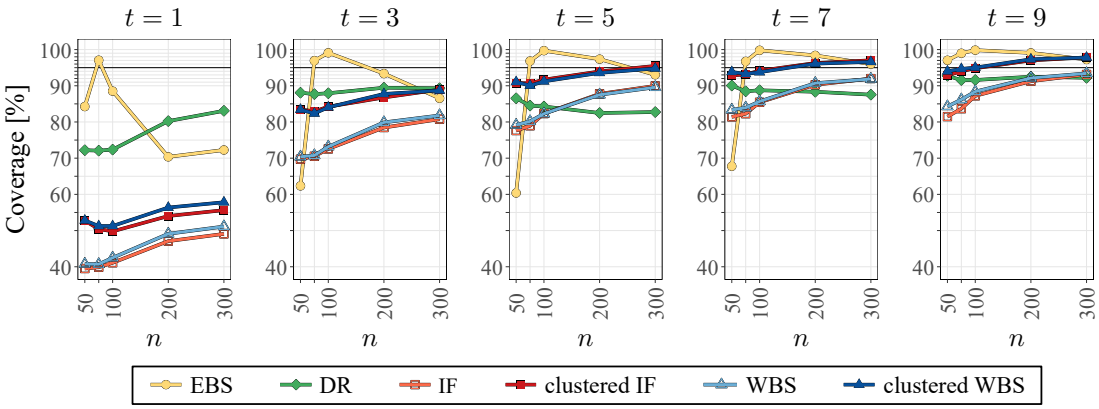


Figure B.92: Coverage of the PS-matched CIs in the scenario with high variance of the covariates and $\beta_{01A} = 0$.

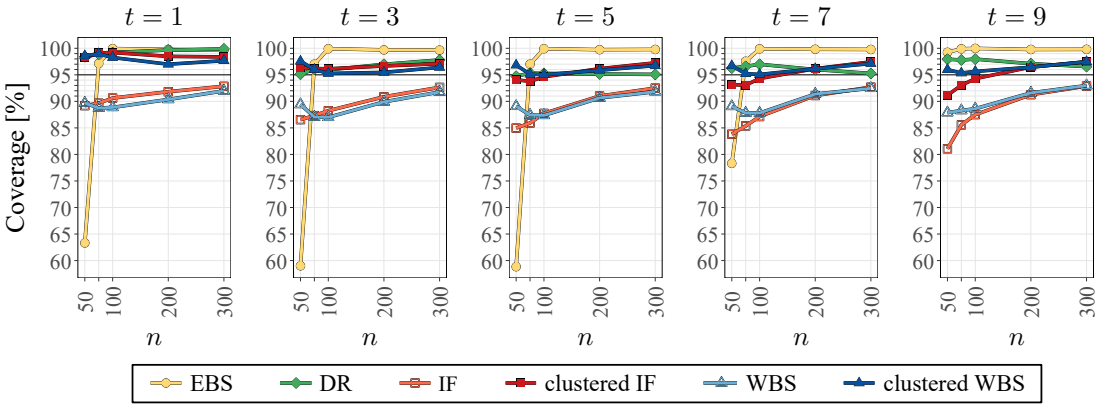


Figure B.93: Coverage of the PS-matched CIs in the scenario with high variance of the covariates and $\beta_{01A} = 2$.

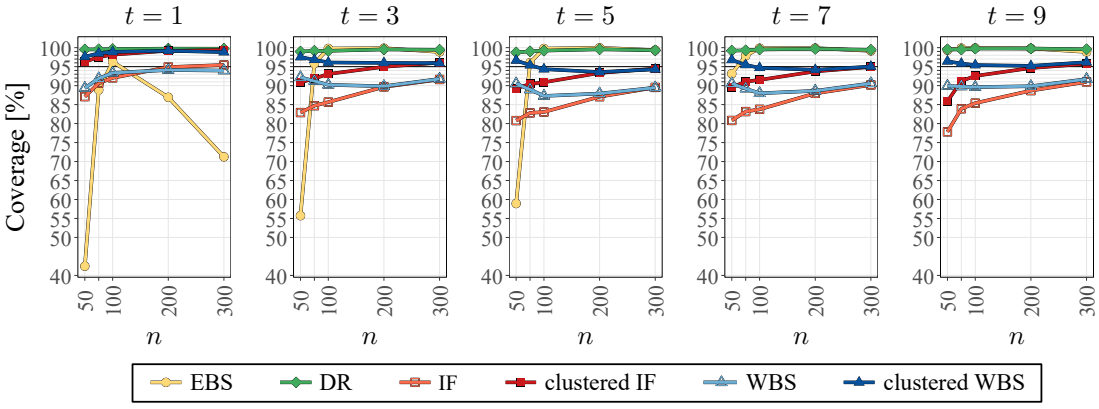


Figure B.94: Coverage of the PS-matched CIs in the scenario with type II censoring and $\beta_{01A} = -2$.

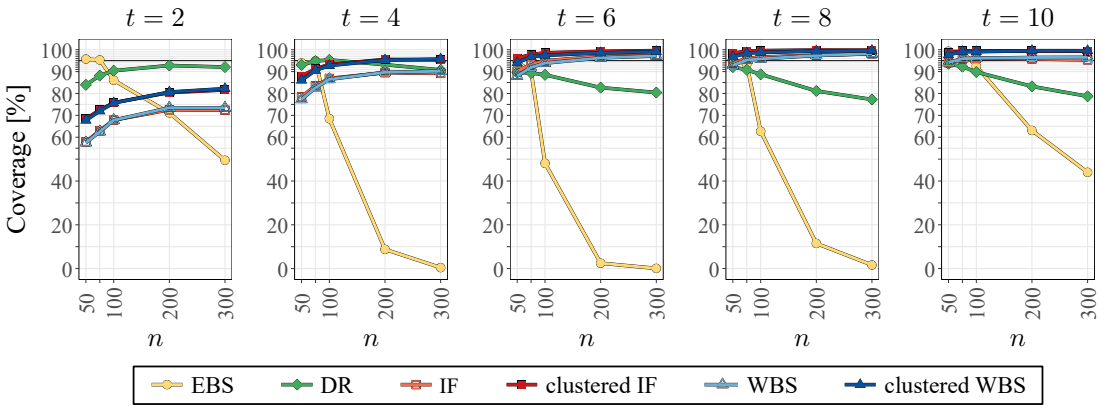


Figure B.95: Coverage of the PS-matched CIs in the scenario with type II censoring and $\beta_{01A} = 0$.

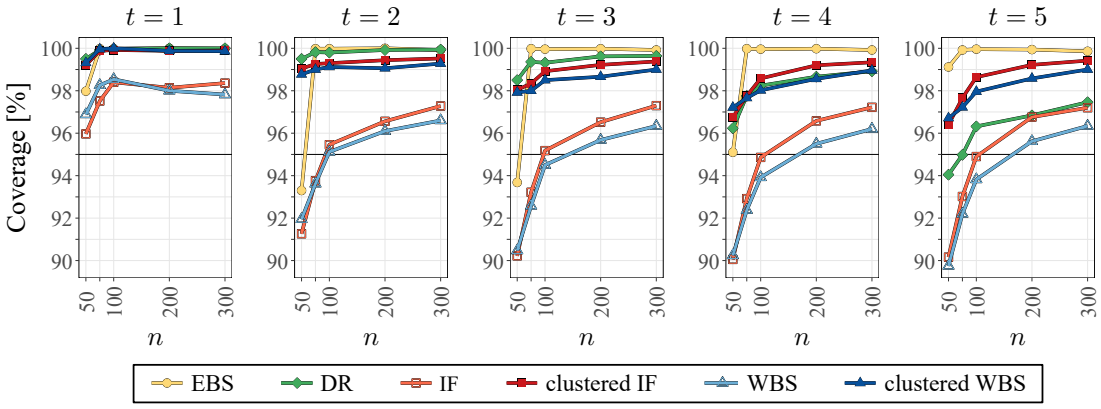


Figure B.96: Coverage of the PS-matched CIs in the scenario with type II censoring and $\beta_{01A} = 2$.

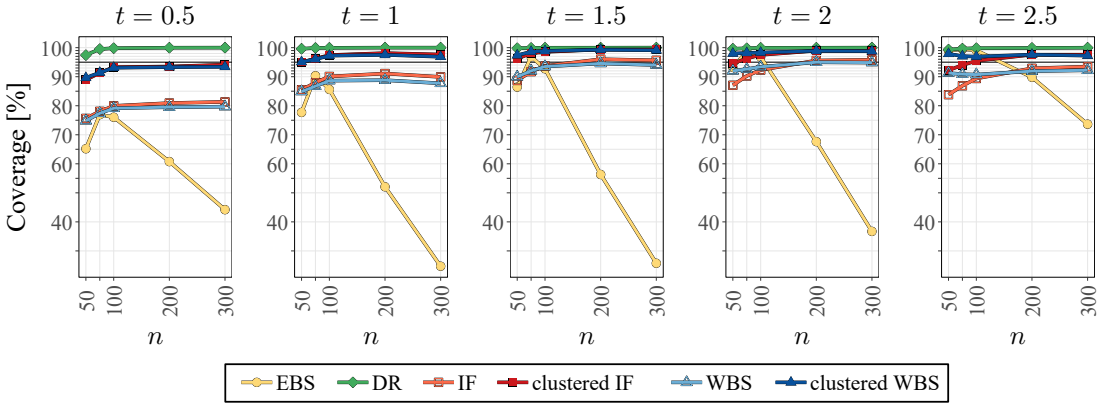


Figure B.97: Coverage of the PS-matched CBs in the scenario with no censoring and $\beta_{01A} = -2$.

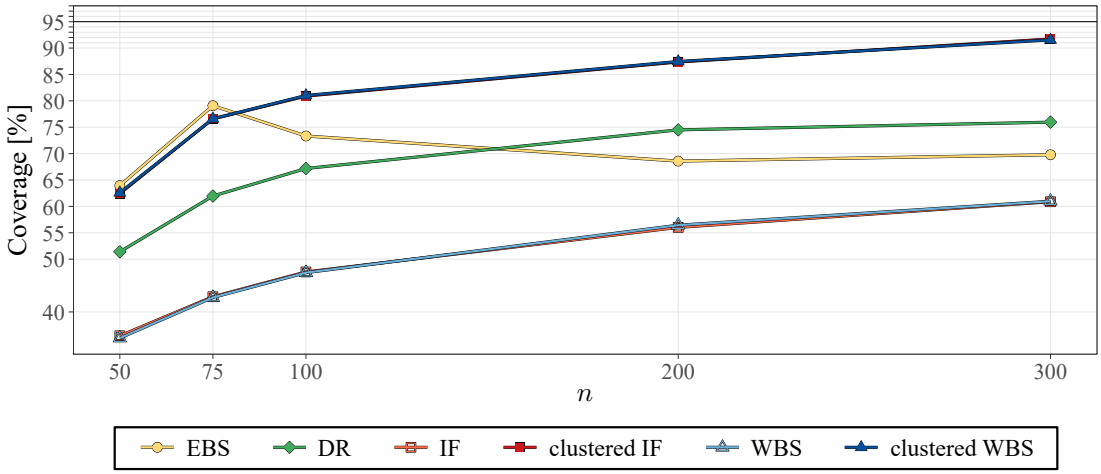


Figure B.98: Coverage of the PS-matched CBs in the scenario with no censoring and $\beta_{01A} = 0$.

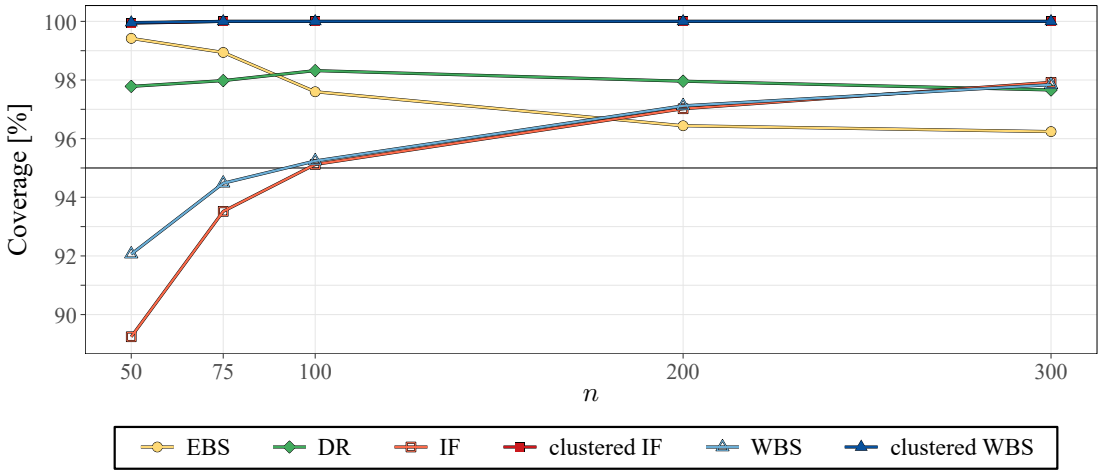


Figure B.99: Coverage of the PS-matched CBs in the scenario with no censoring and $\beta_{01A} = 2$.

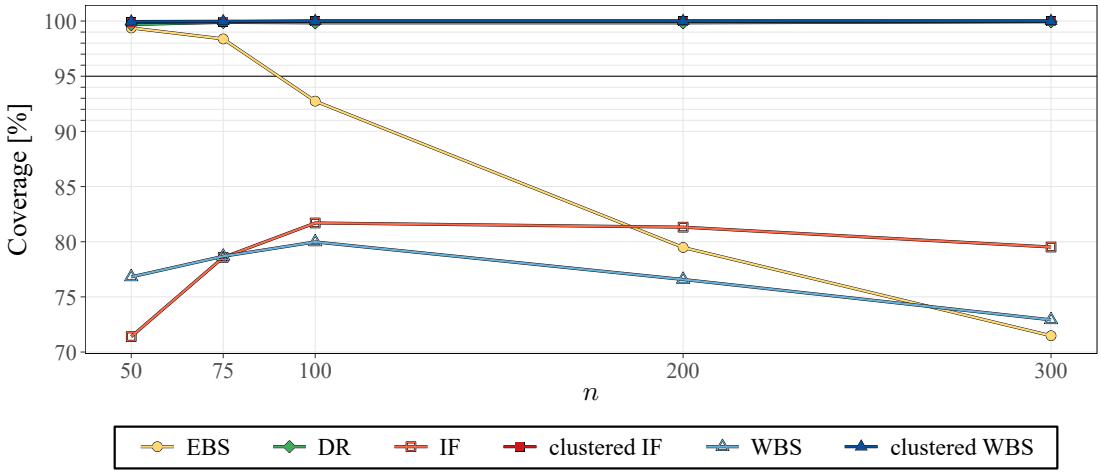


Figure B.100: Coverage of the PS-matched CBs in the scenario with light censoring and $\beta_{01A} = 0$.

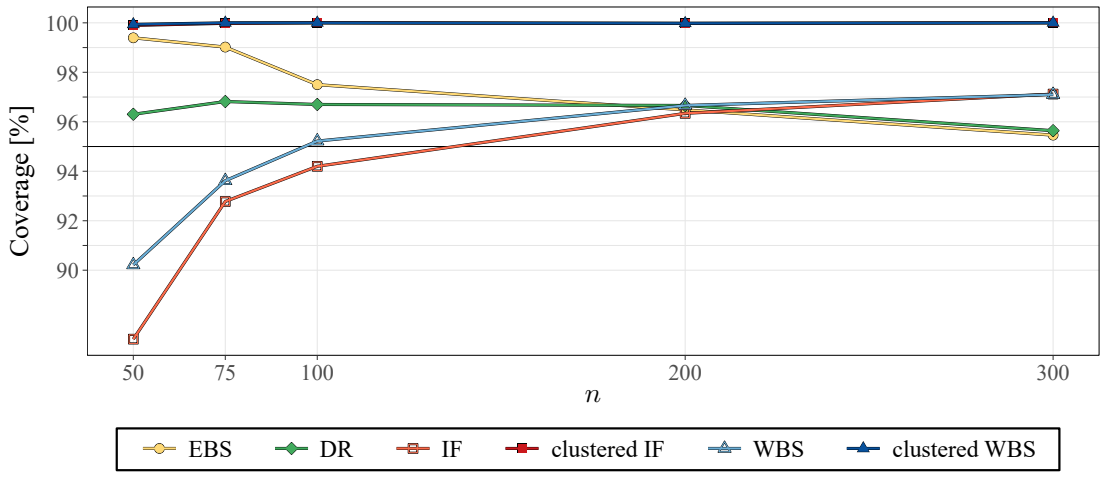


Figure B.101: Coverage of the PS-matched CBs in the scenario with light censoring and $\beta_{01A} = 2$.

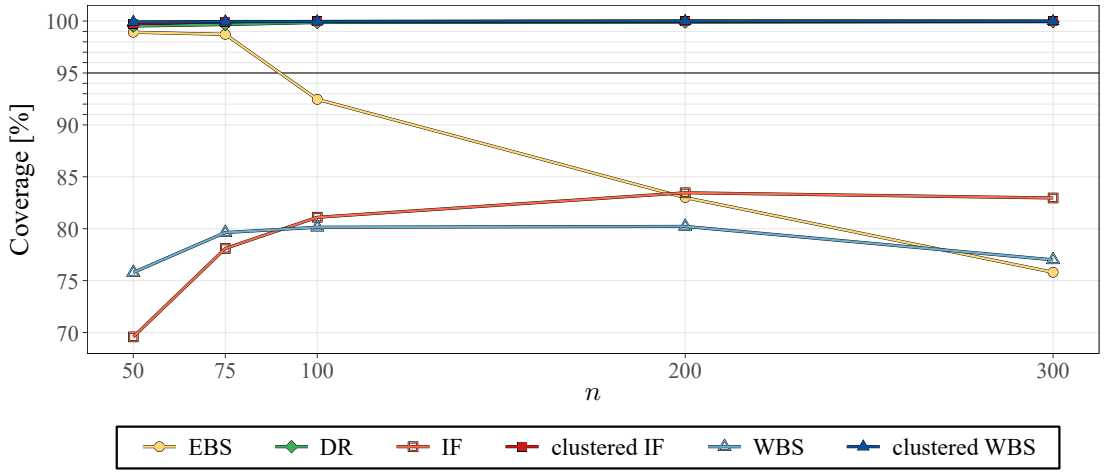


Figure B.102: Coverage of the PS-matched CBs in the scenario with heavy censoring and $\beta_{01A} = -2$.

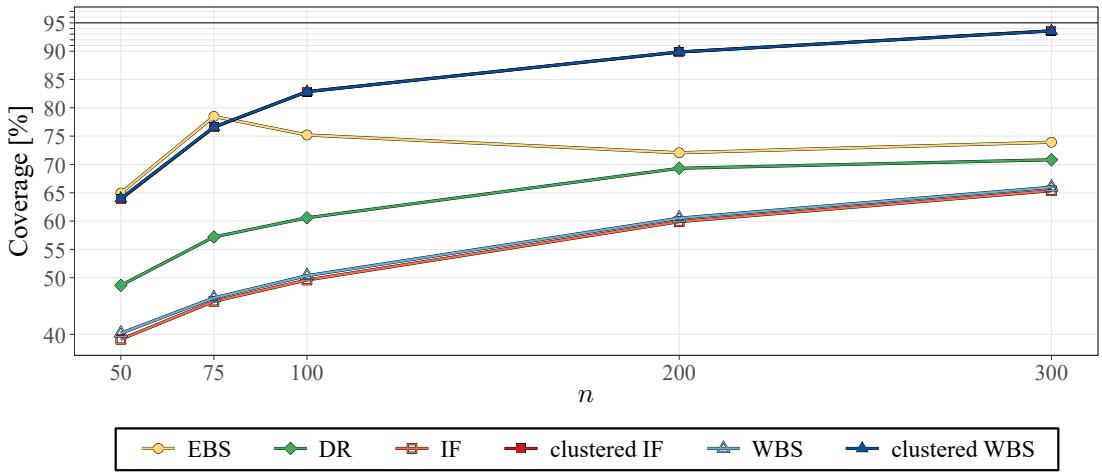


Figure B.103: Coverage of the PS-matched CBs in the scenario with heavy censoring and $\beta_{01A} = 0$.

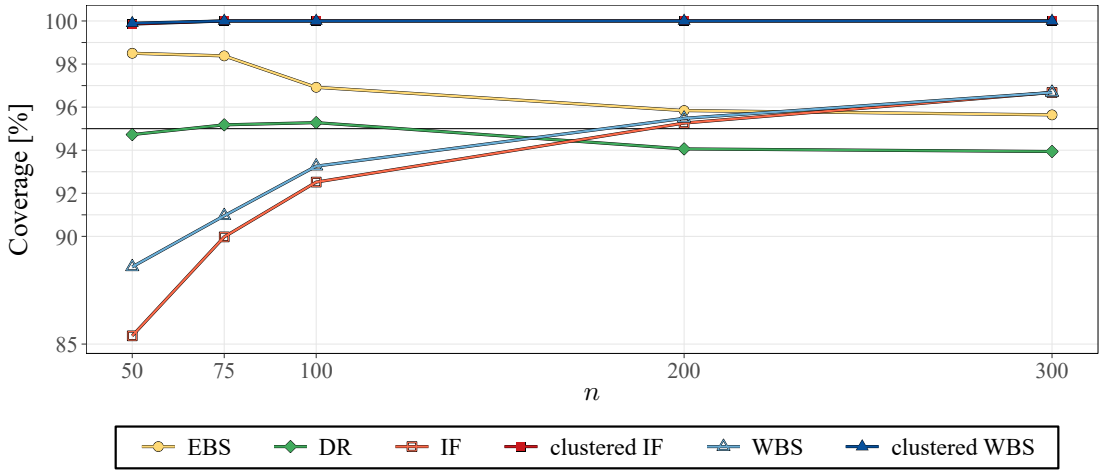


Figure B.104: Coverage of the PS-matched CBs in the scenario with heavy censoring and $\beta_{01A} = 2$.

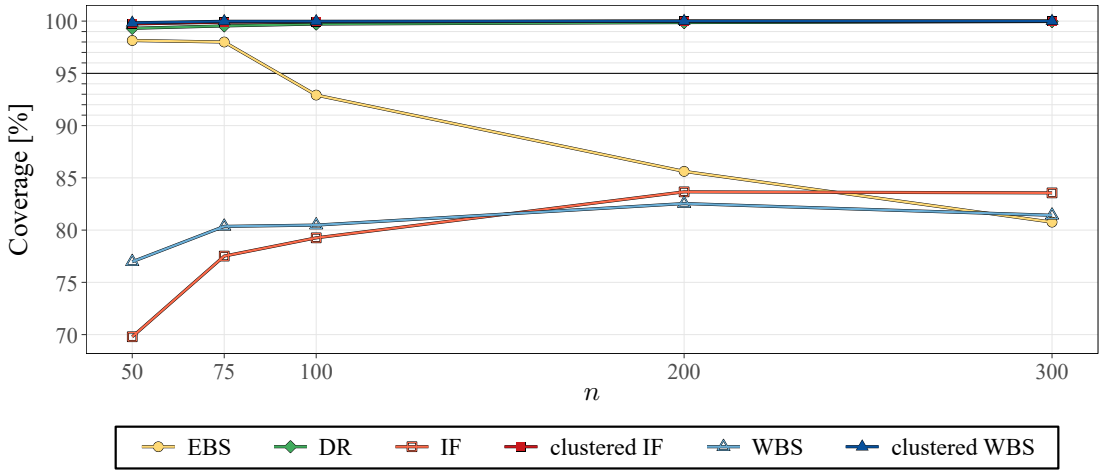


Figure B.105: Coverage of the PS-matched CBs in the scenario with low treatment probability and $\beta_{01A} = -2$.

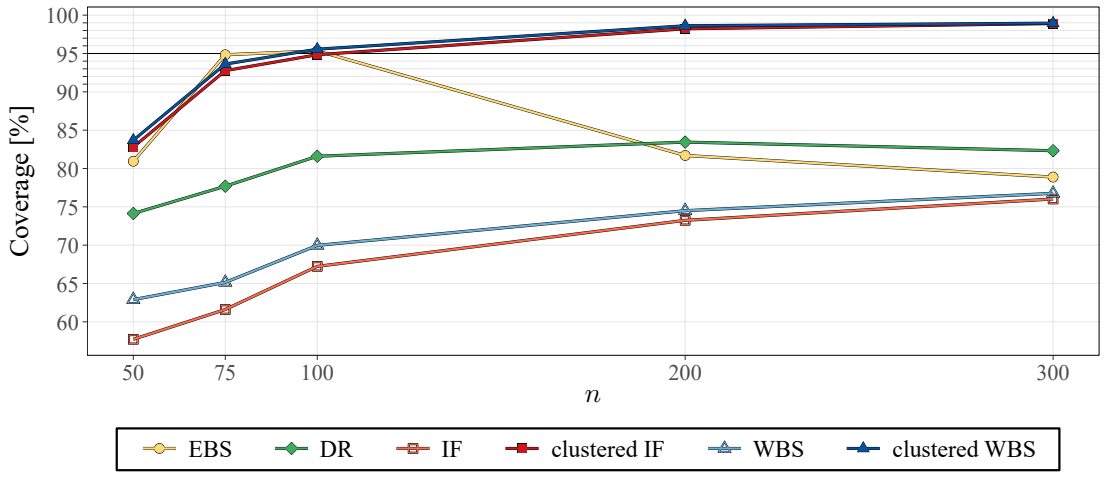


Figure B.106: Coverage of the PS-matched CBs in the scenario with low treatment probability and $\beta_{01A} = 0$.

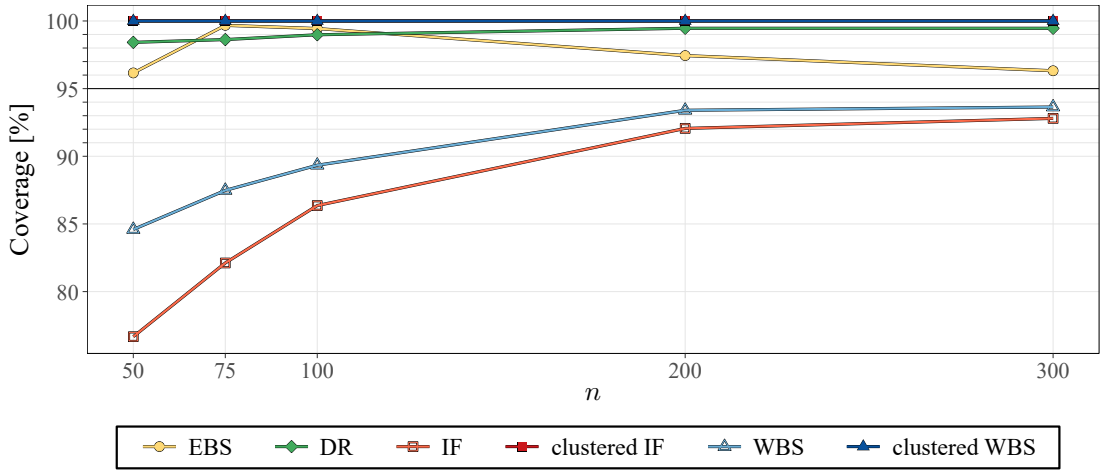


Figure B.107: Coverage of the PS-matched CBs in the scenario with low treatment probability and $\beta_{01A} = 2$.

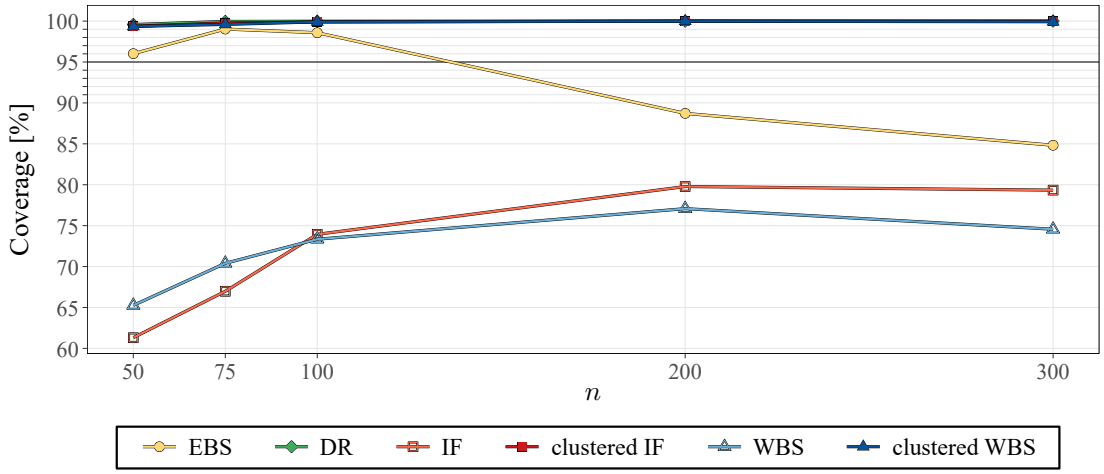


Figure B.108: Coverage of the PS-matched CBs in the scenario with high treatment probability and $\beta_{01A} = -2$.

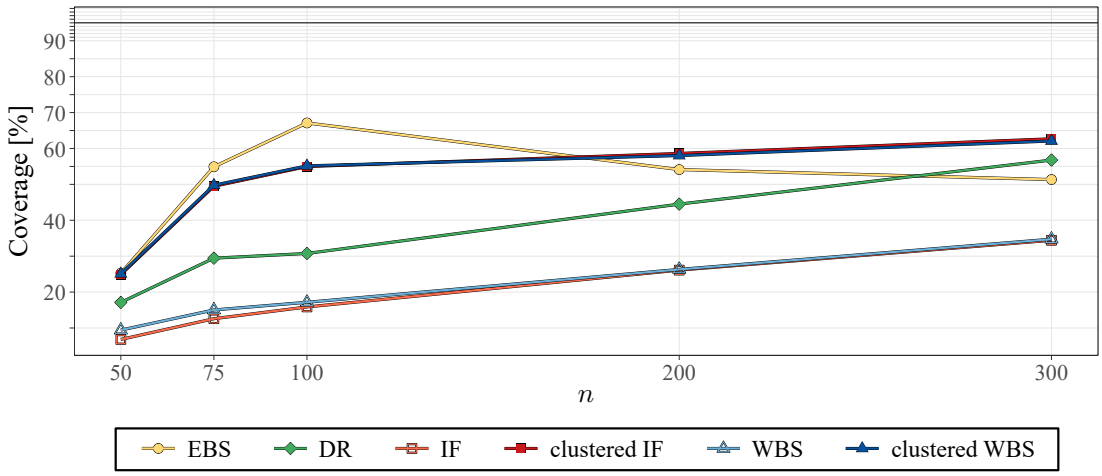


Figure B.109: Coverage of the PS-matched CBs in the scenario with high treatment probability and $\beta_{01A} = 0$.

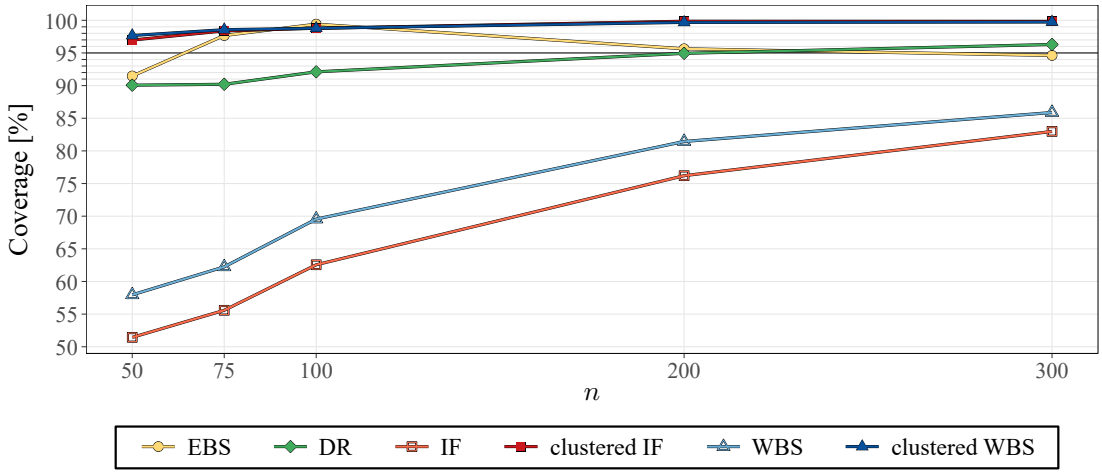


Figure B.110: Coverage of the PS-matched CBs in the scenario with high treatment probability and $\beta_{01A} = 2$.

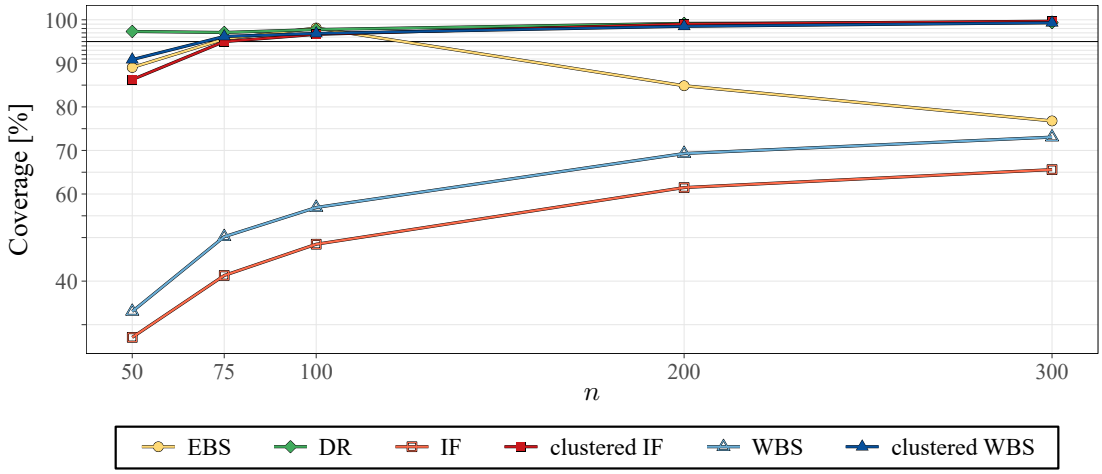


Figure B.111: Coverage of the PS-matched CBs in the scenario with low variance of the covariates and $\beta_{01A} = -2$.

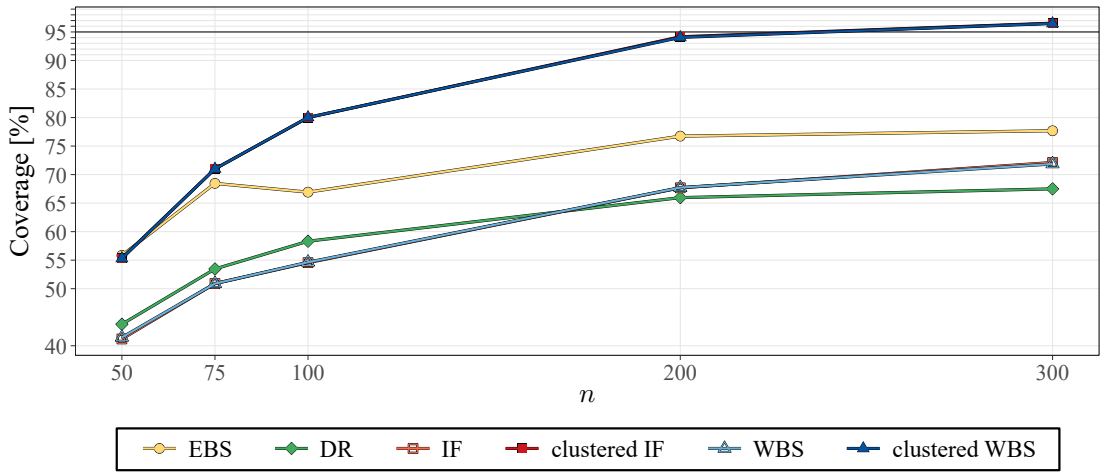


Figure B.112: Coverage of the PS-matched CBs in the scenario with low variance of the covariates and $\beta_{01A} = 0$.

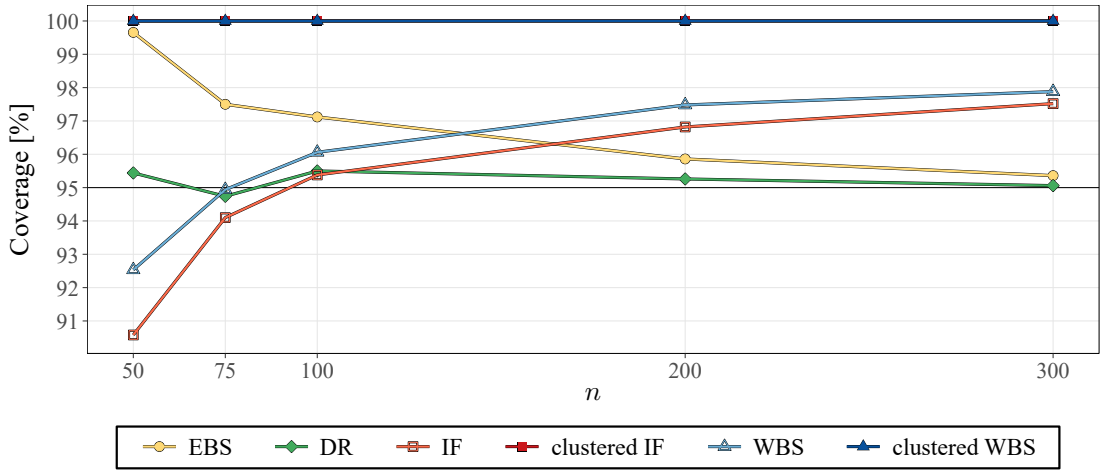


Figure B.113: Coverage of the PS-matched CBs in the scenario with low variance of the covariates and $\beta_{01A} = 2$.

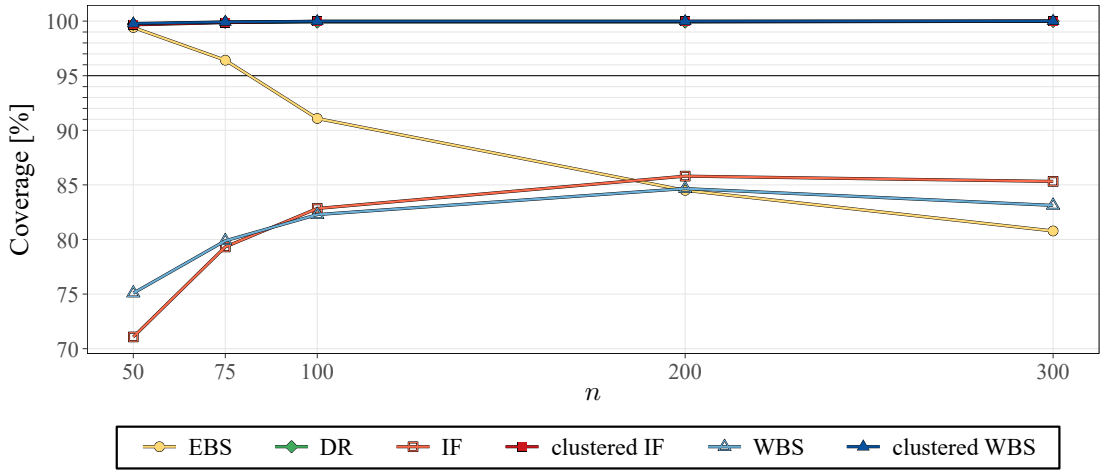


Figure B.114: Coverage of the PS-matched CBs in the scenario with high variance of the covariates and $\beta_{01A} = -2$.

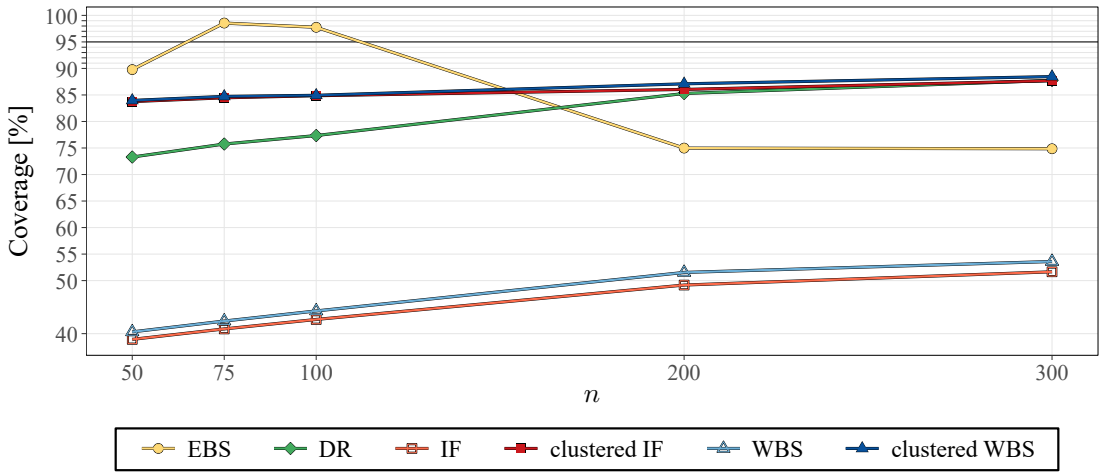


Figure B.115: Coverage of the PS-matched CBs in the scenario with high variance of the covariates and $\beta_{01A} = 0$.

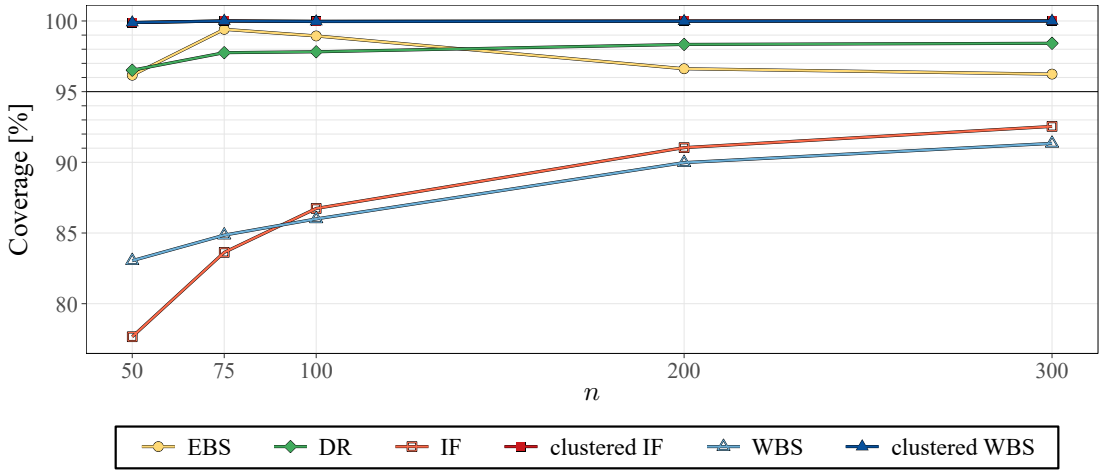


Figure B.116: Coverage of the PS-matched CBs in the scenario with high variance of the covariates and $\beta_{01A} = 2$.

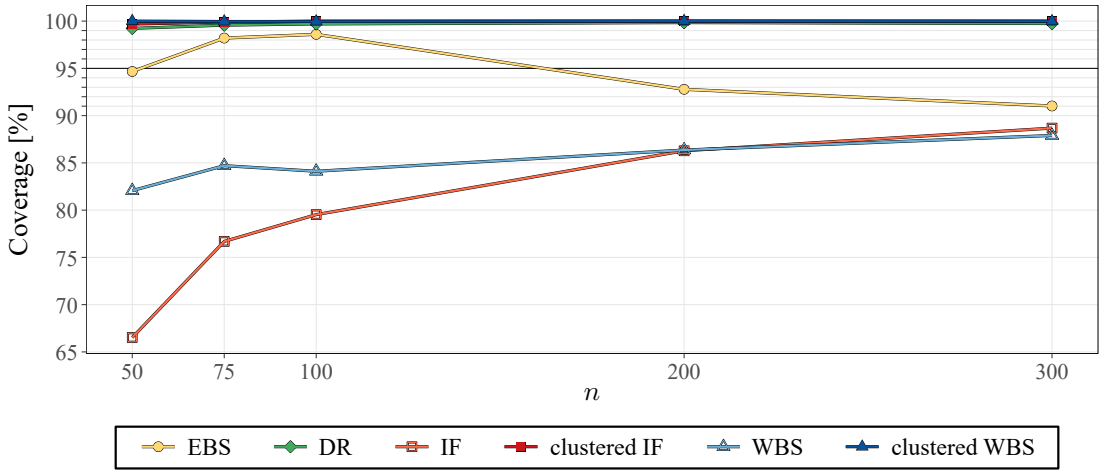


Figure B.117: Coverage of the PS-matched CBs in the scenario with type II censoring and $\beta_{01A} = -2$.

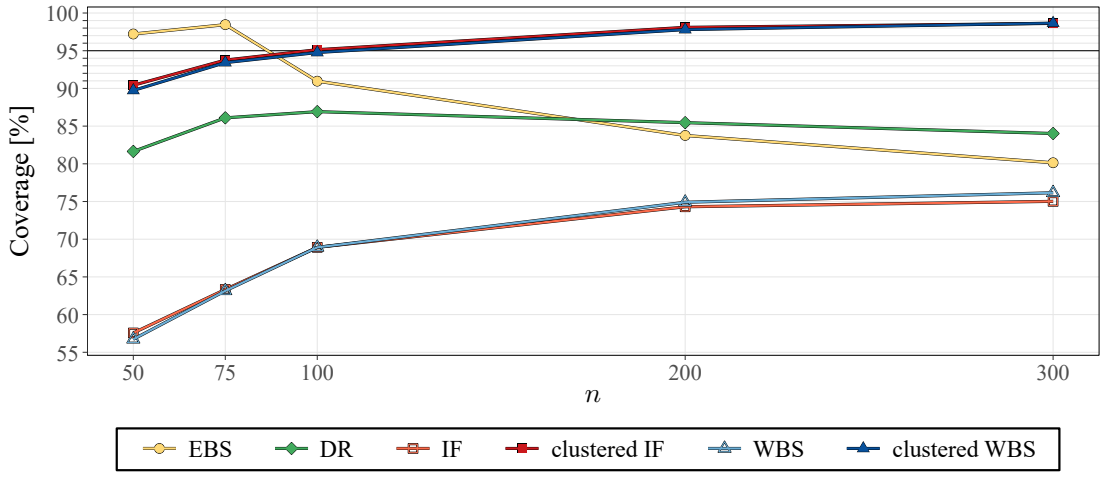
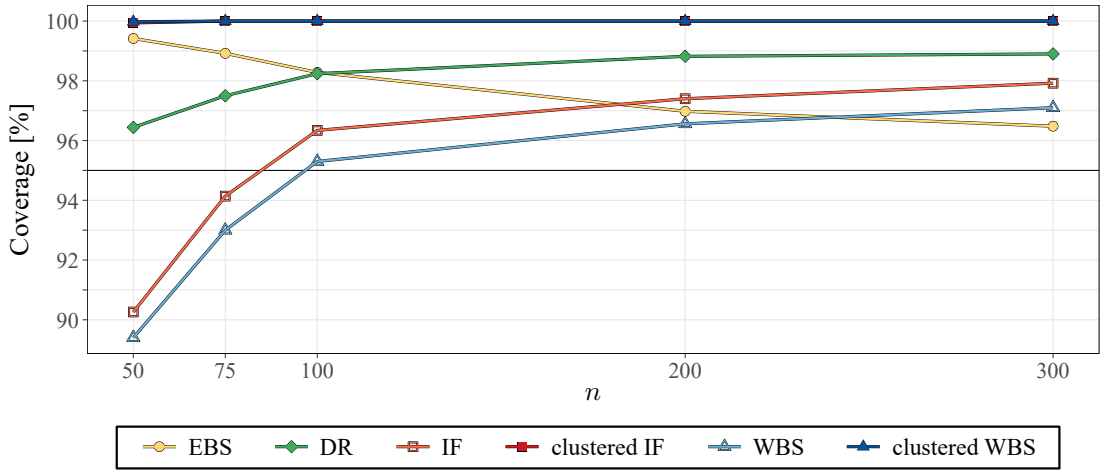


Figure B.118: Coverage of the PS-matched CBs in the scenario with type II censoring and $\beta_{01A} = 0$.



Bibliography

Aalen, O. O. (1978). “Nonparametric inference for a family of counting processes.” In: *The Annals of Statistics* 6.4, pp. 701–726.

Aalen, O. O. (1989). “A linear regression model for the analysis of life times.” In: *Statistics in Medicine* 8.8, pp. 907–925.

Aalen, O. O., Borgan, Ø., and Gjessing, H. K. (2008). *Survival and Event History Analysis. A Process Point of View*. New York: Springer Science & Business Media.

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

Abadie, A. and Imbens, G. W. (2006). “Large sample properties of matching estimators for average treatment effects.” In: *Econometrica* 74.1, pp. 235–267.

Abadie, A. and Imbens, G. W. (2008). “On the failure of the bootstrap for matching estimators.” In: *Econometrica* 76.6, pp. 1537–1557.

Abadie, A. and Imbens, G. W. (2011). “Bias-corrected matching estimators for average treatment effects.” In: *Journal of Business & Economic Statistics* 29.1, pp. 1–11.

Adusumilli, K. (2022). *Bootstrap inference for propensity matching*. Working paper.

Akritas, M. G. (1986). “Bootstrapping the Kaplan-Meier estimator.” In: *Journal of the American Statistical Association* 81.396, pp. 1032–1038.

Andersen, P. K. (2005). “Censored data.” In: *Encyclopedia of Biostatistics*. Ed. by P. Armitage and T. Colton. 2nd ed. John Wiley & Sons.

Andersen, P. K., Borgan, Ø., Gill, R. D., and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. New York: Springer Science & Business Media.

Andersen, P. K., Geskus, R. B., de Witte, T., and Putter, H. (2012). “Competing risks in epidemionlogy: possibilities and pitfalls.” In: *International Journal of Epidemiology* 41.3, pp. 861–870.

Andersen, P. K. and Gill, R. D. (1982). “Cox’s regression model for counting processes: a large sample study.” In: *The Annals of Statistics* 10.4, pp. 1100–1120.

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). “Identification of causal effects using instrumental variables.” In: *Journal of the American Statistical Association* 91.434, pp. 444–472.

Austin, P. C. and Cafri, G. (2020). “Variance estimation when using propensity-score matching with replacement with survival or time-to-event outcomes.” In: *Statistics in Medicine* 39.11, pp. 1623–1640.

Austin, P. C., Steyerberg, E. W., and Putter, H. (2021). “Fine-Gray subdistribution hazard models to simultaneously estimate the absolute risk of different event types: cumulative total failure probability may exceed 1.” In: *Statistics in Medicine* 40.19, pp. 4200–4212.

Baden, L. R. et al. (2021). “Efficacy and safety of the mRNA-1273 SARS-CoV-2 vaccine.” In: *New England Journal of Medicine* 384, pp. 403–416.

Benichou, J. and Gail, M. H. (1990). “Estimates of absolute cause-specific risk in cohort studies.” In: *Biometrics* 46.3, pp. 813–826.

Beyersmann, J., Allignol, A., and Schumacher, M. (2012). *Competing Risks and Multistate Models with R*. New York: Springer Science & Business Media.

Beyersmann, J., Di Termini, S., and Pauly, M. (2013). “Weak convergence of the wild bootstrap for the Aalen-Johansen estimator of the cumulative incidence function of a competing risk.” In: *Scandinavian Journal of Statistics* 40.3, pp. 387–402.

Beyersmann, J., Latouche, A., Buchholz, A., and Schumacher, M. (2009). “Simulating competing risks data in survival analysis.” In: *Statistics in Medicine* 28.6, pp. 956–971.

Billingsley, P. (1999). *Convergence of probability measures*. 2nd ed. New York: John Wiley & Sons.

Bluhmki, T., Dobler, D., Beyersmann, J., and Pauly, M. (2019). “The wild bootstrap for multivariate Nelson-Aalen estimators.” In: *Lifetime Data Analysis* 25, pp. 97–127.

Bodory, H., Camponovo, L., Huber, M., and Lechner, M. (2016). *A wild bootstrap algorithm for propensity score matching estimators*. Working Papers SES 470. Faculty of Economics and Social Sciences, University of Fribourg.

Breslow, N. E. (1972). “Discussion of the paper ‘Regression models and life-tables’ by D. R. Cox.” In: *Journal of the Royal Statistical Society: Series B* 34.2, pp. 216–217.

Breslow, N. E. (1974). “Covariance analysis of of censored survival data.” In: *Biometrics* 30.1, pp. 89–99.

Bühler, A., Cook, R. J., and Lawless, J. F. (2023). “Multistate models as a framwork for estimand specification in clinical trials of complex processes.” In: *Statistics in Medicine* 42.9, pp. 1368–1397.

Butt, J. H. et al. (2021). “Vitamin K antagonists vs. direct oral anticoagulants after transcatheter aortic valve implantation in atrial fibrillation.” In: *European Heart Journal – Cardiovascular Pharmacotherapy* 7.1, pp. 11–19.

Center for Drug Evaluation and Research (CDER) (2016). *Statistical Review and Evaluation: BLA 761,041*. https://www.accessdata.fda.gov/drugsatfda_docs/nda/2016/761041Orig1s000StatR.pdf. (Accessed July 16, 2024).

Chauhan, L. et al. (2022). “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.” In: *Clinical Infectious Diseases* 75.1, e466–e472.

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

Chernick, M. R. (2012). “Resampling methods.” In: *Wiley Interdisciplinary Reviews (WIREs): Data Mining and Knowledge Discovery* 2, pp. 255–262.

Cox, D. R. (1972). “Regression models and life-tables.” In: *Journal of the Royal Statistical Society: Series B* 34.2, pp. 187–202.

Daniel, R., Zhang, J., and Farewell, D. (2021). “Making apples from oranges: comparing noncollapsible effect estimators and their standard errors after adjustment for different covariate sets.” In: *Biometrical Journal* 63.3, pp. 528–557.

Ditzhaus, M. and Pauly, M. (2019). “Wild bootstrap logrank tests with broader power functions for testing superiority.” In: *Computational Statistics and Data Analysis* 136, pp. 1–11.

Dobler, D., Beyersmann, J., and Pauly, M. (2017). “Non-strange weird resampling for complex survival data.” In: *Biometrika* 104.3, pp. 699–711.

Dobler, D. and Pauly, M. (2014). “Bootstrapping Aalen-Johansen processes for competing risks: handicaps, solutions, and limitations.” In: *Electronic Journal of Statistics* 8.2, pp. 2779–2803.

Efron, B. (1979). “Bootstrap methods: another look at the jackknife.” In: *The Annals of Statistics* 7.1, pp. 1–26.

Efron, B. and Stein, C. (1981). “The jackknife estimate of variance.” In: *The Annals of Statistics* 9.3, pp. 586–596.

Elisei, R. et al. (2013). “Cabozantinib in progressive medullary thyroid cancer.” In: *Journal of Clinical Oncology* 31.29, pp. 3639–3646.

European Medicines Agency (EMA) (2020). *Points to Consider on Implications of Coronavirus Disease (COVID-19) on Methodological Aspects of Ongoing Clinical Trials*. https://www.ema.europa.eu/en/documents/scientific-guideline/points-consider-implications-coronavirus-disease-covid-19-methodological-aspects-ongoing-clinical-trials_en.pdf. (Accessed October 11, 2024).

Fine, J. P. and Gray, R. J. (1999). “A proportional hazards model for the subdistribution of a competing risk.” In: *Journal of the American Statistical Association* 94.446, pp. 496–509.

Fleming, T. R. and Harrington, D. P. (2005). *Counting Processes and Survival Analysis*. 2nd ed. John Wiley & Sons.

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

Gandara, D. R. et al. (2018). “Blood-based tumor mutational burden as a predictor of clinical benefit in non-small-cell lung cancer patients treated with atezolizumab.” In: *Nature Medicine* 24, pp. 1441–1448.

Gerds, T. A., Ohlendorff, J. S., and Ozenne, B. M. (2023). *riskRegression: risk regression models and prediction scores for survival analysis with competing risks*. R package (version 2023.03.22). URL: https://cran.r-project.org/package=riskRegression.

Gerds, T. A. and Schumacher, M. (2001). “On functional misspecification of covariates in the Cox regression model.” In: *Biometrika* 88.2, pp. 572–580.

Gill, R. D. (1980). *Censoring and Stochastic Integrals*. Mathematical Centre Tract 124. Amsterdam: Mathematical Centre.

Grambsch, P. M. and Therneau, T. M. (1994). “Proportional hazards tests and diagnostics based on weighted residuals.” In: *Biometrika* 81.3, pp. 515–526.

Gran, J. M., Lie, S. A., Øyeflaten, I., Borgan, Ø., and Aalen, O. O. (2015). “Causal inference in multi-state models–sickness absence and work for 1145 participants after work rehabilitation.” In: *BMC Public Health* 15.1082.

Greenland, S., Pearl, J., and Robins, J. M. (1999). “Causal diagrams for epidemiologic research.” In: *Epidemiology* 10.1, pp. 37–48.

Hampel, F. R. (1974). “The influence curve and its role in robust estimation.” In: *Journal of the American Statistical Association* 69.346, pp. 383–393.

Hernán, M. A. (2010). “The hazards of hazard ratios.” In: *Epidemiology* 21.1, pp. 13–15.

Hernán, M. A. (2015). “Counterpoint: epidemiology to guide decision-making: moving away from practice-free research.” In: *American Journal of Epidemiology* 182.10, pp. 834–839.

Hernán, M. A., Brumback, B., and Robins, J. M. (2000). “Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men.” In: *Epidemiology* 11.5, pp. 561–70.

Hernán, M. A. and Robins, J. M. (2020). *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC.

Hrba, M., Maciak, M., Peštova, B., and Pešta, M. (2022). “Bootstrapping not independent and not identically distributed data.” In: *Mathematics* 10.24, p. 4671.

Husain, M. et al. (2019). “Oral semaglutide and cardiovascular outcomes in patients with type 2 diabetes.” In: *New England Journal of Medicine* 381, pp. 841–851.

Ishwaran, H. and Kogalur, U. B. (2024). *randomForestSRC: fast unified random forests for survival, regression, and classification*. R package (version 3.3.1). URL: https://cran.r-procject.org/package=randomForestSRC.

Kennedy, E. H. (2022). *Semiparametric doubly robust targeted double machine learning: a review*. ”arXiv”: 2203.06469 (stat.ME).

Keogh, R. H., Gran, J. M., Seaman, S. R., Davies, G., and Vansteenlandt, S. (2023). “Causal inference in survival analysis using longitudinal observational data: sequential trials and marginal structural models.” In: *Statistics in Medicine* 42.13, pp. 2191–2225.

Kleinbaum, D. G. and Klein, M. (2012). *Survival Analysis. A self learning text*. New York: Springer Science & Business Media.

Kosorok, M. R. (2008). *Introduction to Empirical Processes and Semiparametric Inference*. New York: Springer Science & Business Media.

Lamberts, M. et al. (2014). “Relation of nonsteroidal anti-inflammatory drugs to serious bleeding and thromboembolism risk in patients with atrial fibrillation receiving antithrombotic therapy: a nationwide cohort study.” In: *Annals of Internal Medicine* 161.10, pp. 690–698.

Lesko, C. R. and Lau, B. (2017). “Bias due to confounders for the exposure-competing risk relationship.” In: *Epidemiology* 28.1, pp. 20–27.

Lin, D. Y. (1997). “Non-parametric inference for cumulative incidence functions in competing risks studies.” In: *Statistics in Medicine* 16.8, pp. 901–910.

Lin, D. Y., Fleming, T. R., and Wei, L. J. (1994). “Confidence bands for survival curves under the proportional hazards model.” In: *Biometrika* 81.1, pp. 73–81.

Lin, D. Y. and Wei, L. J. (1989). “The robust inference for the Cox proportional hazards model.” In: *Journal of the American Statistical Association* 84.408, pp. 1074–1078.

Lin, D. Y., Wei, L. J., and Ying, Z. (1993). “Checking the Cox model with cumulative sums of martingale-based residuals.” In: *Biometrika* 80.3, pp. 557–572.

Lopuhaä, H. P. and Nane, G. F. (2013). “Shape constrained non-parametric estimators of the baseline distribution in Cox proportional hazards model.” In: *Scandinavian Journal of Statistics* 40.3, pp. 619–646.

Martinussen, T. and Scheike, T. H. (2006). *Dynamic Regression Models for Survival Data*. New York: Springer Science & Business Media.

Martinussen, T. and Stensrud, M. J. (2023). “Estimation of separable and indirect effects in continuous time.” In: *Biometrics* 79.1, pp. 127–139.

Martinussen, T. and Vansteelandt, S. (2013). “On collapsibility and confounding bias in Cox and Aalen regression models.” In: *Lifetime Data Analysis* 19.3, pp. 279–296.

McLaughlin, V. et al. (2015). “Bosentan added to sildenafil therapy in patients with pulmonary arterial hypertension.” In: *European Respiratory Journal* 46, pp. 405–413.

Meyer, P.-A. (1962). “A decomposition theorem for supermartingales.” In: *Illinois Journal of Mathematics* 6.2, pp. 193–205.

Meyer, P.-A. (1963). “Decomposition of supermartingales: the uniqueness theorem.” In: *Illinois Journal of Mathematics* 7.1, pp. 1–17.

Meyer, R. D. et al. (2020). “Statistical issues and recommendations for clinical trials conducted during the COVID-19 pandemic.” In: *Statistics in Biopharmaceutical Research* 12.4, pp. 399–411.

Nelson, W. (1969). “Hazard plotting for incomplete failure data.” In: *Journal of Quality Technology* 1.1, pp. 27–52.

Nelson, W. (1972). “Theory and applications of hazard plotting for censored failure data.” In: *Technometrics* 14.4, pp. 945–966.

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

Nießl, A., Allignol, A., Beyersmann, J., and Mueller, C. (2021). “Statistical inference for state occupation and transition probabilities in non-Markov multi-state models subject to both random left-truncation and right-censoring.” In: *Econometrics and Statistics* 25, pp. 110–124.

O’Quigley, J. (2008). *Proportional Hazards Regression*. New York: Springer Science & Business Media.

Otsu, T. and Rai, Y. (2017). “Bootstrap inference of matching estimators for average treatment effects.” In: *Journal of the American Statistical Association* 520, pp. 1720–1732.

Overgaard, M. and Hansen, S. N. (2021). “On the assumption of independent right censoring.” In: *Scandinavian Journal of Statistics* 48, pp. 1234–1255.

Ozenne, B. M., Scheike, T. H., Stærk, L., and Gerds, T. A. (2020). “On the estimation of average treatment effects with right-censored time to event outcome and competing risks.” In: *Biometrical Journal* 62.3, pp. 751–763.

Ozenne, B. M., Sørensen, A. L., Scheike, T., Torp-Pedersen, C., and Gerds, T. A. (2017). “riskRegression: predicting the risk of an event using Cox regression models.” In: *The R Journal* 9.2, pp. 440–460.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems. Networks of Plausible Inference*. San Mateo: Morgan Kaufman.

Pearl, J. (1995). “Causal diagrams for empirical research.” In: *Biometrika* 82.4, pp. 669–710.

Peto, R. (1972). “Discussion of the paper ‘Regression models and life-tables’ by D. R. Cox.” In: *Journal of the Royal Statistical Society: Series B* 34.2, pp. 205–207.

Pintilie, M. (2006). *Competing Risks. A Practical Perspective*. John Wiley & Sons.

Post, R. A., van den Heuvel, E. R., and Putter, H. (2024). “The built-in selection bias of hazard ratios formalized using structural causal models.” In: *Lifetime Data Analysis* 30.2, pp. 404–438.

Prentice, R. L. and Kalbfleisch, J. D. (2003). “Mixed discrete and continuous Cox regression model.” In: *Lifetime Data Analysis* 9.2, pp. 195–210.

Putter, H., Fiocco, M., and Geskus, R. B. (2007). “Tutorial in biostatistics: competing risks and multi-state models.” In: *Statistics in Medicine* 26.11, pp. 2389–2430.

R Core Team (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Rebolledo, R. (1980). "Central limit theorems for local martingales." In: *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* 51, pp. 269–286.

Rittmeyer, A. et al. (2017). "Atezolizumab versus Docetaxel in patients with previously-treated non-small-cell lung cancer (OAK): a phase 3, open-label, multicentre randomised controlled trial." In: *The Lancet* 389.10066, pp. 255–265.

Robins, J. M. (1986). "A new approach to causal inference in mortality studies with a sustained exposure period - application to control of the healthy worker survivor effect." In: *Mathematical Modelling* 7, pp. 1393–1512.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). "Estimation of regression coefficients when some regressors are not always observed." In: *Journal of the American Statistical Association* 89.427, pp. 846–866.

Rosenbaum, P. R. and Rubin, D. B. (1983). "The central role of the propensity score in observational studies for causal effects." In: *Biometrika* 70.1, pp. 41–55.

Rubin, D. B. (1974). "Estimating causal effects of treatments in randomized and nonrandomized studies." In: *Journal of Educational Psychology* 66.5, pp. 688–701.

Rubin, D. B. (2006). "Causal inference through potential outcomes and principal stratification: application to studies with "censoring" due to death." In: *Statistical Science* 21.3, pp. 299–309.

Rühl, J., Beyersmann, J., and Friedrich, S. (2023). "General independent censoring in event-driven trials with staggered entry." In: *Biometrics* 79.3, pp. 1737–1748.

Rühl, J. and Friedrich, S. (2024a). "Asymptotic properties of resampling-based processes for the average treatment effect in observational studies with competing risks." In: *Scandinavian Journal of Statistics*. DOI: 10.1111/sjos.12714, pp. 1–27.

Rühl, J. and Friedrich, S. (2024b). "Resampling-based confidence intervals and bands for the average treatment effect in observational studies with competing risks." In: *Statistics and Computing* 34.101.

Ryalen, P. C., Stensrud, M. J., Fosså, S., and Røysland, K. (2020). "Causal inference in continuous time: an example on prostate cancer therapy." In: *Biostatistics* 21.1, pp. 172–185.

Rytgaard, H. C. and van der Laan, M. J. (2024). “Targeted maximum likelihood estimation for causal inference in survival and competing risks analyis.” In: *Lifetime Data Analysis* 30.1, pp. 4–33.

Scheike, T. H. and Zhang, M.-J. (2002). “An additive-multiplicative Cox-Aalen regression model.” In: *Scandinavian Journal of Statistics* 29.1, pp. 75–88.

Scheike, T. H. and Zhang, M.-J. (2008). “Flexible competing risks regression modeling and goodness-of-fit.” In: *Lifetime Data Analysis* 14.4, pp. 464–483.

Singh, K. (1981). “On the asymptotic accuracy of Efron’s bootstrap.” In: *The Annals of Statistics* 9.6, pp. 1187–1195.

Sitbon, O. et al. (2015). “Selexipag for the treatment of pulmonary arterial hypertension.” In: *New England Journal of Medicine* 373, pp. 2522–2533.

Stærk, L. et al. (2017). “Ischaemic and haemorrhagic stroke associated with non-vitamin K antagonist oral anticoagulants and warfarin use in patients with atrial fibrillation: a nationwide cohort study.” In: *European Heart Journal* 38.12, pp. 907–915.

Stensrud, M. J., Young, J. G., Didelez, V., Robins, J. M., and Hernán, M. A. (2022). “Separable effects for causal inference in the presence of competing events.” In: *Journal of the American Statistical Association* 117.537, pp. 175–183.

Tsiatis, A. A. (1981). “A large sample study of Cox’s regression model.” In: *The Annals of Statistics* 9.1, pp. 93–108.

Van der Laan, M. J. and Rubin, D. B. (2006). “Targeted maximum likelihood learning.” In: *The International Journal of Biostatistics* 2.1.

Van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge: Cambridge University Press.

Van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes. With Applications to Statistics*. New York: Springer Science & Business Media.

Vandenbroucke, J. and Pearce, N. (2015). “Point: incident exposures, prevalent exposures, and causal inference: does limiting studies to persons who are followed from first exposure onward damage epidemiology?” In: *American Journal of Epidemiology* 182.10, pp. 826–833.

Vansteelandt, S., Dukes, O., van Lancker, K., and Martinussen, T. (2024). “Assumption-lean Cox regression.” In: *Journal of the American Statistical Association* 119.545, pp. 475–484.

Wang, T. et al. (2024). “Propensity score matching for estimating a marginal hazard ratio.” In: *Statistics in Medicine* 43, pp. 2783–2810.

Wu, C. F. J. (1986). “Jackknife, bootstrap and other resampling methods in regression analysis.” In: *The Annals of Statistics* 14.4, pp. 1261–1295.

Young, J. G., Stensrud, M. J., Tchetgen Tchetgen, E. J., and Hernán, M. A. (2020). “A causal framework for classical statistical estimands in failure-time settings with competing events.” In: *Statistics in Medicine* 39.8, pp. 1199–1236.

Zepeda-Tello, R. et al. (2022). *The delta-method and influence function in medical statistics: a reproducible tutorial*. ”arXiv”: 2206.15310 (stat.ME).