UNIVERSITY OF COPENHAGEN
FACULTY OF HEALTH AND MEDICAL SCIENCES

PhD thesis

Causal inference in time-to-event analysis

Simon Christoffer Ziersen

Advisors:
Esben Budtz-Jørgensen
Thomas Alexander Gerds
Brice Maxime Ozenne
Lars Vedel Kessing

This thesis has been submitted to the Graduate School of the Faculty of Health and Medical Sciences, University of Copenhagen on July 27, 2024

Causal inference in time-to-event analysis

PhD thesis

Simon Christoffer Ziersen

Section of Biostatistics
Department of Public Health
University of Copenhagen

July 27, 2024

Academic advisors:
Esben Budtz-Jørgensen, University of Copenhagen
Thomas Alexander Gerds, University of Copenhagen
Brice Maxime Ozenne, University of Copenhagen
Lars Vedel Kessing, University of Copenhagen

Preface

The work in this thesis was carried out at the Section of Biostatistics, University of Copenhagen from 2021 to 2024. Four months was spend visiting Torsten Hothorn and the Deparment of Biostatistics, University of Zürich in the spring/summer 2023. During my research stay in Zürich, my new found colleagues welcomed me to a friendly atmosphere and invited me along to social events outside of working hours. I would like to thank the people at the Department of Biostatics, Univervisty of Zürich, for making my stay an excellent experience.

I would also like to thank my supervisors. My collaboration with Lars Vedel Kessing started before my time as a PhD student, and through our work I was introduced to time-to-event analysis, which (as the title of the thesis would suggest) has become a big part of my research. I owe a great thanks to Lars for our collaborations and for motivating the methodological work of the thesis. I would like to thank Brice Maxime Hugues Ozenne for his always open door and feedback on much of my work. I thank Thomas Alexander Gerds and Esben Budtz-Jørgensen for their support through out the entirety of my PhD and for sticking it out with me in the end when time was precious.

Finally, I would like to extend my sincerest gratitude to Torben Martinussen for our collaborations, which eventually made up the two thirds of the thesis.

Summary

The aim of this thesis is to provide statistical methods for assessing treatment effects with registry data, where the outcome of interest is time to an event. It is often the case that only a censored version of the underlying event time is available, as patients may leave the risk set due to emigration or end of follow-up. Furthermore, some competing event, such as death, may prevent observation of the event of interest and methods from survival analysis allowing for competing risks has to be combined with causal inference methodology for inferring the treatment effect.

The thesis is comprised of a synopsis consisting of seven chapters and three manuscripts. Chapter 2 gives an introduction to semiparametric efficiency theory and the use of data-adaptive methods for functional estimation, which form the basis of the methodological work undertaken in the manuscripts. Chapter 3-4 introduce the causal inference methodology considered in the manuscripts, and Chapter 5 discusses the application to a study based on data from the Danish national registers. Chapter 6 gives a summary of the manuscripts and Chapter 7 reflects on the limitations of the work and some potential avenues for future research. The contributions in this thesis can be grouped in two categories.

Average treatment effect estimation with censoring and competing risks
Manuscript I considers estimation of the average treatment effect based on the τ-year absolute risk with a high-dimensional set of potential confounders. We derive an estimator for the target parameter, that allows for penalized regressions for nuisance parameter estimation. The method is applied to a study comparing the response to different antidepressants using data from the Danish national registers.

Manuscript III derives a measure of treatment effect based on the number of life years lost due to a specific event. This definition of treatment effect returns the interpretation to the timescale of the study, which is easier to communicate compared to risks. We derive an estimator that allows for data-adaptive estimation of the nuisance parameters and give high-level assumptions for valid inference of the estimator.

Assessment of heterogeneous treatment effects
Manuscript II extends a *treatment effect variable importance measure* to censored data. The measure is used to assess the amount of treatment effect heterogeneity explained by a given set of covariates. Additionally, a new measure is derived as a *best partially linear projection* of the conditional average treatment effect. The projection measures the heterogeneity explained by a single covariate and it has interpretation as a regression coefficient. Manuscript III extends the projection measure from Manuscript II to the treatment effect based on the number of life years lost due to a specific event. The method is applied to the register study on response to different antidepressants.

Resumé

Målet for denne afhandling er at bidrage med statistiske metoder til vurdering af behandlingseffekter ved hjælp af registerdata, hvor responsen er tid til en begivenhed. Ofte observerer man kun en censureret udgave af den underlæggende begivenhedstids, da patienter kan udgå fra risikomængden som følge af emigration eller studiets afslutning. Ydermere, kan nogle konkurrerende begivenheder, såsom død, forhindre observation af responsbegivenheden, og metoder fra overlevelsesanalyse, der tillader konkurrerende begivenheder, må kombineres med metoder fra kausal inferens for at udlede behandlingseffekten.

Denne afhandling indeholder en synopsis på syv kapitler og tre manuskripter. Kapitel 2 giver en introduktion til semiparametrisk efficiensteori og brugen af data-adaptive metoder til funktionalestimation, der danner basis for det metodiske arbejde i manuskripterne. Kapitel 3 og 4 introducerer den kausal inferens-metodik, der er undersøgt i manuskripterne, og kapitel 5 diskuterer anvendelsen i et studie baseret på data fra de danske registre. Kapitel 6 giver et resumé af manuskripterne, og kapitel 7 reflekterer over begrænsningerne af resultaterne og potentielle emner til fremtidig forskning. Bidragene i denne afhandling kan grupperes i to kategorier.

Estimation af den gennemsnitlig behandlingseffekt med censurering og konkurrerende risici

Manuskript I omhandler estimation af den gennemsnitlige behandlingseffekt baseret på den τ-år absolutte risiko med høj-dimensionale konfoundere. Vi konstruerer en estimator, der tillader estimation af *nuisance*-parametrene vha. penaliserede regressioner. Metoden anvendes på et studie, der sammenligner responsen til forskellige antidepressiva med data fra de danske registre.

Manuskript III udleder et mål for behandlingseffekt baseret på antallet af mistet leveår som følge af en specifik begivenhed. Denne definition af behandlingseffekt bringer fortolkningen tilbage til tidsenheden i studiet, som er nemmere at kommunikerer end risiko. Vi konstruerer en estimator, der tillader data-adaptiv estimation af *nuisance*-parametrene, og giver overordnede betingelser for valid inferens.

Vurdering af heterogene behandlingseffekter

Manuskript II udvider et mål for *variable importance* på behandlingseffekter til censureret data. Målet bruges til at vurdere størrelsen af behandlingseffekt-heterogenitet, der kan tilskrives et givent sæt af kovariater. Ydermere, udvikles et nyt mål som den *bedste partielle lineære projektion* af den betingede behandlingseffekt. Projektionen måler heterogenitet forklaret af en enkelt kovariat og har fortolkning som en regressionskoefficient. Manuskript III udvider projektionsmålet til behandlingseffekter baseret på antallet af mistet leveår som følge af en specifik begivenhed. Metoden anvendes på registerstudiet vedrørende respons til antideppresiva.

Contents

Chapter 1

Introduction

The work carried out in this thesis was undertaken as part of work package one of a larger research project called BrainDrugs.[1] The overall aim of BrainDrugs is to identify patient features that determine drug response in patients with major depressive disorders (MDD) and epilepsy, respectively. The project was divided into seven work packages, each with a different focus and researchers from different scientific areas. The aim of work package one was to study response to antidepressants in patients with MDD using data from the Danish national registers. The overall aim of this thesis is to provide statistically sound methods for estimating drug response using registry data and to provide methods for identification of treatment effect modifiers.

In registers, patients are followed over time and the response to a given antidepressant is defined in terms of the time to a specific event of interest. Statistical methods for time-to-event analysis are complicated by the fact that the data are not fully observed, as some patients may leave the study before an observation with the event of interest, due to censoring. Examples of a censoring mechanism in register data are emigration or end-of-follow-up, after which the only information on the time to the event of interest is that it did not happen in a certain period. Furthermore, some patients may die before the event of interest. This prevents observation of the event of interest, and we say that death constitutes a competing risk.

The question of determining drug response from observational data is causal in nature. When contrasting the response to antidepressant A and B, only one of the two is observed for a given patient and the underlying question becomes ”how would the patient have responded if he or she was given the other antidepressant”. On a population level, the average difference in drug response (or the *average treatment effect* in the causal language) is then interpreted as ”what is the expected difference in drug response if everyone was treated with antidepressant A compared to if everyone was treated with B”. As a patient is only treated with either A or B, methods from the causal inference literature has to be employed for answering such questions using observational data.

The study carried out in Kessing et al. (2024) laid the motivation for the methodological explorations in this thesis. The study uses methods from the causal inference literature to emulate a targeted randomized trial (Hernán and Robins, 2016) when data are subject to censoring and competing risk.

[1] https://braindrugs.nru.dk/

Objectives

The objectives of the thesis can group in two categories

(i) Reliable estimation of the ATE in the presence of censoring and competing risks.

(ii) Identification of treatment effect modifiers in the presence of censoring and competing risks.

Objective (i) addresses some challenges involved with estimation of causal effects using high-dimensional data. One assumption needed for identification of the ATE in observational data, is that of "no unmeasured confounding", which with high-dimensional data amounts to including a large covariate set in certain regressions. Machine-learning methods designed for the purpose of statistical learning using high-dimensional data are attractive in such setting, but a naive implementation of such methods breaks the asymptotic inference of the obtained ATE estimate. Methods from semiparametric efficiency theory are then needed for producing reliable estimators for the ATE in the presence of censored data with competing risks, using data-adaptive nuisance estimators.

Objective (ii) refers directly to the overall aim of BrainDrugs: identifying patient features that determine drug response. Many different machine-learning methods exist for estimating the treatment effect on a patient level with censoring and competing risks, but they are often given as black-box machines and provide little inside into the driving features determining the treatment effect. Thus, objective (ii) deals with the development of methods for detecting treatment effect modifiers, while still having the flexibility of nonparametric machine-learning methods.

Overview

The thesis is comprised of a synopsis and three manuscripts and is organised as follows. Chapter 2 introduces semiparametric efficiency theory and its use in functional estimation with nonparametric models. Chapter 3 presents two different definitions of the ATE in the presence of censoring and competing risks, each with a different interpretation, and constructs estimators with parametric-like behaviour in the presence of data-adaptive nuisance estimators, using the methods from Chapter 2. Chapter 4 defines two different treatment effect variable importance measures for detecting treatment effect modifiers in the presence of censoring and competing risks. Estimators of the measures are constructed using results from Chapter 2. Chapter 5 details the work related to BrainDrugs and drug response to antidepressants using registers data. Chapter 6 gives a summary of the manuscripts and Chapter 7 concludes the synopsis with a some perspectives and ideas for future research.

Notation

The following notation will be used throughout the thesis. We write $Pf = \int f \,\mathrm{d}P$, and when $\hat{f}$ is estimated from data, $P\hat{f}$ considers $\hat{f}$ fixed, i.e., not averaging over the randomness in $\hat{f}$. For observations $X_1, \ldots, X_n$ we denote $\mathbb{P}_n f = \sum_{i=1}^{n} f(X_i)$ the empirical measure of f. The expression E_P is used to denote the expectation with respect to the measure P. We let $\|\cdot\|$ be the $L_2(P)$-norm such that $\|f\| = \sqrt{\int f^2 \,\mathrm{d}P}$, where the dependence on P is implicitly understood, unless otherwise specified. We let $\mathcal{O}$ denote the sample space and $\mathcal{O}_n$ the observed data. Finally, for some process X_n, we take $X_n = o_p(n^{-\epsilon})$ to mean $n^{\epsilon} X_n \overset{P}{\to} 0$, and $X_n = o_p(1)$ to mean $X_n \overset{P}{\to} 0$.

Chapter 2

Functional estimation

Many statistical problems can be formulated as estimation of a functional ψ defined on a family of probability measures $\mathcal{M}$. When the model is indexed by some euclidean parameter, it can be written as $\mathcal{M}_\theta = \{P_\theta : \ \theta \in \mathbb{R}^k\}$ and the statistical estimation problem is often defined as estimation of $\psi(P_\theta) = \theta$. When the model is indexed by a combination of a finite and infinite dimensional parameter it is referred to as a semiparametric model and we write $\mathcal{M}_{\theta,\nu} = \{P_{\theta,\nu} : \theta \in \mathbb{R}^k, \nu \in \mathcal{F}\}$ for some suitable function class $\mathcal{F}$. Here, ν is a nuisance parameter and the estimation problem is again defined as estimation of $\psi(P_{\theta,\nu}) = \theta$. There exist a vast literature on semiparametric efficiency theory, which deals with estimation of functional parameters in semiparametric models (Bickel et al., 1993, van der Vaart, 2000, Tsiatis, 2006). When the model is parametrized by only an infinite dimensional nuisance parameter ν, the model is said to be nonparametric, which can be analysed as a special case using semiparametric efficiency theory. Going forward, we denote $\mathcal{M}$ the nonparametric model.

In recent years, functional estimation in nonparametric models using data-adaptive nuisance parameter estimators has received a lot of attention (van der Laan and Rose, 2011, Kennedy, 2016, Chernozhukov et al., 2018) and nice reviews of the theory can be found e.g., in Kennedy (2022) and Hines, Dukes, et al. (2022). The nonparametric model allows one to define the target parameter of estimation as a map $\psi : \mathcal{M} \rightarrow \mathbb{R}$ (for the sake of illustration, we only consider one-dimensional target parameters) without relying on restrictive parametric model assumptions. This makes the nonparametric model attractive for estimation using observational data, where little in known on the data generating mechanism. As such, many of the recent advances have been related to or motivated by problems from the causal inference literature, with the aim of estimating a causally related target parameter, also known as an estimand (Petersen and van der Laan, 2014). As an example of a commonly targeted estimand, consider the average treatment effect (ATE)

$$\psi(P) = E_P\{E_P(Y \mid A = 1, X) - E_P(Y \mid A = 0, X)\}, \tag{2.1}$$

where $Y \in \mathbb{R}$ is an outcome of interest, $A \in \{0, 1\}$ is a binary treatment indicator, X is a covariate vector and E_P is the expectation under P. A natural approach for estimating $\psi(P)$ with flexible nuisance parameter estimation is with the plug-in estimator $\psi(\hat{P})$, where $\hat{P}$ is some estimate of P (or the relevant parts of it) using appropriate machine-learning tools. For the ATE, the plug-in estimator based on n

i.i.d. observations of (Y, A, X) is given by

$$\psi(\hat{P}) = \frac{1}{n}\sum_{i=1}^{n} \hat{\mu}_1(X_i) - \hat{\mu}_0(X_i),$$

where $\hat{\mu}_a$ is an estimate of the regression function $E(Y \mid A = a, X = x)$ and the distribution of X is estimated by the empirical measure. In Chapter 3, we give a detailed discussion of the causal interpretation of the ATE in an incomplete data setting.

Plug-in estimation of the target parameter does not generally provide valid inference, as the estimation of $\hat{P}$ is targeted towards P rather than $\psi(P)$, which leaves a bias term that converges on a rate slower than $n^{-1/2}$, thus preventing parametric-like inference of the target parameter estimate (Kennedy, 2022). The bias term is related to the so-called *efficient influence function* (EIF) corresponding to the map ψ, which plays a crucial role in the semiparametric inference literature. In this chapter we give a brief introduction to EIF-based estimation of functionals defined on nonparametric models, enabling parametric-like inference while using data-adaptive nuisance estimators.

2.1 Efficient influence function

We consider the problem of estimating a map $\psi : \mathcal{M} \to \mathbb{R}$ at a given $P_0 \in \mathcal{M}$. Here, we think of an experiment based on n observations, say $X_1, \dots, X_n$ where $X_i \sim P_0$, and the aim is to estimate the target parameter $\psi_0 = \psi(P_0)$. Consider an estimator ψ_n of ψ_0. The estimator is called *asymptotically linear* if it admits the representation

$$\psi_n - \psi_0 = \mathbb{P}_n \mathbb{IF}_{P_0} + o_p(n^{-1/2}) \tag{2.2}$$

where $\mathbb{IF}_P : \mathcal{O} \to \mathbb{R}$ is defined as a map on the sample space $\mathcal{O}$ for a given P with $P\mathbb{IF}_P = 0$ and $P\mathbb{IF}^2 < \infty$. The function $\mathbb{IF}$ is called the *influence function* of ψ_n and it characterizes the asymptotic distribution of the estimator, since an application of Slutsky's theorem together with the central limit theorem gives that

$$\sqrt{n}(\psi_n - \psi_0) \longrightarrow \mathcal{N}(0, P_0\mathbb{IF}^2_{P_0}).$$

The question is then, how to construct estimators that are asymptotically linear and how to identify their influence functions using data-adaptive nuisance estimators? It turns out, that, when the map ψ is smooth, it is possible to calculate the so-called *efficient influence function* (EIF), which characterizes the information bound among all regular estimators. Intuitively, if one considers all estimators on the form (2.2), the EIF is the influence function which minimizes $P\mathbb{IF}^2_P$. Once the EIF is known, several methods exist for constructing an estimator that is asymptotically linear with the EIF as its influence function. The EIF is related to the map ψ and the model $\mathcal{M}$ and it is defined without reference to any specific estimator.

The definition of the EIF is quite involved, and we will sketch the definition here but refer to van der Vaart (2000) ch. 25.3 for a rigorous definition. Let $\mathcal{O}$ be a sample space and let $\mathcal{H} = \{g : \mathcal{O} \to \mathbb{R} \mid P_0 g = 0,\ P_0 g^2 < \infty\}$ be the Hilbert space of measurable functions with mean zero and finite variance equipped with the inner product $\langle g_1, g_2 \rangle = P_0 g_1 g_2$ and norm $\|g\| = \sqrt{Pg^2}$. Define a one-dimensional smooth parametric submodel $\mathcal{M}_\epsilon \subset \mathcal{M}$ by $\mathcal{M}_\epsilon = \{P_\epsilon \in \mathcal{M} : \epsilon \in \mathbb{R}\}$ which is differentiable at $\epsilon = 0$, and note that it passes through P_0 at $\epsilon = 0$. Since the submodel is differentiable at $\epsilon = 0$, we can define its score function by $s_\epsilon(o) = \frac{d}{d\epsilon}|_{\epsilon=0} \log dP_\epsilon(o) \in \mathcal{H}$. Define $T(P_0)$ by the collection of all score functions defined according to a given parametric submodel and denote the *tangent space* by $\overline{T(P_0)}$, the closure of the linear span of $T(P_0)$. The map ψ is said to be *pathwise differentiable* at P_0 if, for all parametric submodel scores $s_\epsilon \in \overline{T(P_0)}$, it holds that

$$\left.\frac{d}{d\epsilon}\right|_{\epsilon=0} \psi(P_\epsilon) = \langle s_\epsilon, \tilde{\psi}_{P_0} \rangle, \tag{2.3}$$

for a unique function $\tilde{\psi}_{P_0} \in \overline{T(P_0)}$ called the *efficient influence function.* In the proper semiparametric case, multiple functions fulfil (2.3) and $\tilde{\psi}_P$ is not uniquely defined by the equation, but it can be obtained by projecting any functions satisfying (2.3) onto the tangent space. Here, we only consider the nonparametric case, where the tangent space equals the entire Hilbert space $\mathcal{H}$ (Tsiatis, 2006, Theorem 4.4). Hence, any function satisfying (2.3) is automatically in the tangent space and, thus, is the EIF.

Earlier, we hinted at the EIF as the influence function with the lowest variance. With the definition in place, this can now be formalised. For any parametric submodel $\mathcal{M}_\epsilon$, estimation of $\psi(P_0)$ should be "easier" for $P_0 \in \mathcal{M}_\epsilon$ as opposed to $P_0 \in \mathcal{M}$. The Cramér-Rao bound gives that the optimal asymptotic variance for estimation of $\psi(P_\epsilon)$ at $\epsilon = 0$ with score s_ϵ is given by

$$\frac{\left(\frac{d}{d\epsilon}\big|_{\epsilon=0} \psi(P_\epsilon)\right)^2}{P_0 s_\epsilon^2} = \frac{(P_0 \tilde{\psi}_{P_0} s_\epsilon)^2}{P_0 s_\epsilon^2} \leq P_0 \tilde{\psi}_{P_0}^2$$

where $P_0 s_\epsilon^2$ is the Fisher information. The equality follows from the definition of the EIF and the inequality follows from Cauchy-Schwarz. Taking the supremum over all scores $s_\epsilon \in \overline{T(P_0)}$ shows that the variance of the EIF provides the smallest upper bound on the optimal asymptotic variance for estimating $\psi(P_0)$ over all parametric submodels, since the EIF is itself an element of the tangent space. Hence, the EIF defines the lower bound on the attainable asymptotic variance for estimating $\psi(P_0)$ in the nonparametric model $\mathcal{M}$.

The above derivations are summarized in van der Vaart (2000) Theorem 25.19 and we refer to the surrounding chapter (25.3) for a detailed discussion. In particular, any regular estimator (see van der Vaart, 1991 for a definition) of a pathwise differentiable target parameter is asymptotically efficient if and only if it is asymptotically linear with the EIF as its influence function (van der Vaart, 2000 Theorem 25.23).

Even though equation (2.3) provides a direct approach for finding the EIF of a given target parameter, it is often complicated, as it involves solving an integral equation. An alternative approach using Gateaux derivatives is discussed in Hines, Dukes, et al. (2022), Kennedy (2022) and Ichimura and Newey (2022). The Gateaux derivative at P_0 in the direction of $Q \in \mathcal{M}$ is defined as the ordinary derivative

$$\frac{\psi(\epsilon Q + (1-\epsilon)P_0)) - \psi(P_0)}{\epsilon}, \quad \epsilon \to 0.$$

Assume the data to be discrete and let Q be the Dirac measure at a single observation O. Define the parametric submodel $P_\epsilon = \epsilon Q + (1-\epsilon)P_0$ and note that is has score function $s_\epsilon(o) = \frac{dQ-dP}{dP}$. If the map ψ is pathwise differentiable at P_0, equation (2.3) gives that

$$\frac{d}{d\epsilon}\bigg|_{\epsilon=0} \psi(P_\epsilon) = \int \tilde{\psi}_{P_0}(o)\,\mathrm{d}(Q-P)(o) = \tilde{\psi}_{P_0}(O).$$

Thus, when the target parameter is pathwise differentiable, the Gateaux derivative provides a direct approach for calculating the corresponding EIF. The above approach can be made rigorous for continuous data by approximating Q by a kernel (Ichimura and Newey, 2022), but the result remain the same. In manuscript II and III, we take this approach for deriving the EIF.

With the EIF at hand, several approaches exist for constructing estimators that are asymptotically linear with the EIF as their influence function (van der Vaart, 2000, van der Laan and Rose, 2011, Kennedy, 2022). In the following, we describe the one-step estimator, which adds the bias obtained from data-adaptive nuisance estimation in the plug-in estimator to the plug-in estimate itself.

2.2 One-step estimator

For a pathwise differentiable target parameter $\psi(P_0)$ and plug-in estimator $\psi(\hat{P})$, the one-step estimator is defined as

$$\hat{\psi}^{OS} = \psi(\hat{P}) + \mathbb{P}_n\tilde{\psi}_{\hat{P}},$$

where $\tilde{\psi}_{\hat{P}}$ is the EIF with the estimated $\hat{P}$ in place of P_0 (Kennedy, 2022). We will restrict our attention to the setting where the EIF is linear in the target parameter, i.e. $\tilde{\psi}_{P_0} = \varphi_{P_0} - \psi(P_0)$, for some measurable function φ_{P_0} defined on the sample space with finite variance, which we denote the uncentered EIF. This restriction plays no role in the analysis of the one-step estimator, but since the EIF's corresponding to the target parameters considered in manuscript I, II and III are all on this form, we adopt this setting, as to not confuse notation. The one-step estimator becomes $\hat{\psi}^{OS} = \mathbb{P}_n\varphi_{\hat{P}}$, which we note is also equal to the estimating equation based estimator defined as the solution to $\mathbb{P}_n\tilde{\psi}_{\hat{P}} = 0$ when the EIF is linear in the target parameter. To analyse the asymptotic properties of the one-step estimator, consider the expansion

$$\begin{aligned}
\hat{\psi}^{OS} - \psi(P_0) &= \mathbb{P}_n\varphi_{\hat{P}} - \psi(P_0)\\
&= \mathbb{P}_n\varphi_{\hat{P}} - \psi(P_0) + \mathbb{P}_n\tilde{\psi}_{P_0} - \mathbb{P}_n\tilde{\psi}_{P_0}\\
&= \mathbb{P}_n\tilde{\psi}_{P_0} + \mathbb{P}_n(\varphi_{\hat{P}} - \varphi_{P_0})\\
&= \mathbb{P}_n\tilde{\psi}_{P_0} + \mathbb{P}_n(\varphi_{\hat{P}} - \varphi_{P_0}) + P_0(\varphi_{\hat{P}} - \varphi_{P_0}) - P_0(\varphi_{\hat{P}} - \varphi_{P_0})\\
&= \mathbb{P}_n\tilde{\psi}_{P_0} + \underbrace{(\mathbb{P}_n - P_0)(\varphi_{\hat{P}} - \varphi_{P_0})}_{\text{empirical process term}} + \underbrace{P_0\varphi_{\hat{P}} - \psi(P_0)}_{\text{remainder term}}
\end{aligned}$$

where the last equality follows by $P_0\varphi_{P_0} = \psi(P_0)$. If the empirical process term and the remainder term in the above display are both $o_p(n^{-1/2})$, the one-step estimator is seen to be asymptotically linear with $\tilde{\psi}_{P_0}$ as its influence function and hence $\sqrt{n}(\hat{\psi}^{OS} - \psi(P_0)) \overset{d}{\to} \mathcal{N}(0, P\tilde{\psi}^2_{P_0})$. Accordingly, we define an estimator for the variance $P_0\tilde{\psi}^2_{P_0}$ by

$$\hat{\sigma}^2 = \mathbb{P}_n(\varphi_{\hat{P}} - \hat{\psi}^{OS})^2.$$

The standard error of $\hat{\psi}^{OS}$ is given as $\sqrt{\hat{\sigma}^2/n}$, which can be used for inference.

The remainder term can in many cases be shown to admit a double robust structure, whereby $\hat{P}$ (or parts of it) is only required to be estimated at $n^{-1/4}$-rate in order to obtain $n^{-1/2}$-convergences of the remainder term. In the ATE-example from equation (2.1), the nuisance parameters are $\pi(a \mid x) = P_0(A = a \mid X = x)$ and $\mu_a(x)$, and defining their corresponding estimates by $\hat{\pi}$ and $\hat{\mu}_a$, the corresponding remainder term is bounded by $\sum_{a=0,1} \|\hat{\pi} - \pi_0\| \, \|\hat{\mu}_a - \mu_a\|$ (Kennedy, 2016). Hence, $n^{-1/2}$-convergence of the remainder term in the ATE example is obtained if the product of the convergence rates of the nuisance parameter estimators is $n^{-1/2}$. This requirement is fulfilled for many machine-learning estimators and as an example with neural networks for nuisance parameter estimation, see Farrell et al. (2021).

By lemma 19.24 in van der Vaart (2000), the empirical process is $o_p(n^{-1/2})$ if the estimator $\hat{P}$ is assumed to belong to a Donsker class (see van der Vaart, 2000 ch. 19 for definition), which amounts to nuisance estimators that are not too complex. Furthermore, the estimates of the uncentered EIF is required to be consistent as $\left\|\varphi_{\hat{P}} - \varphi_{P_0}\right\| = o_p(1)$. The Donsker class requirement has been shown to be too restrictive for some data-adaptive estimators (Chernozhukov et al., 2018), and a type of sample splitting (termed *cross-fitting*) is used to alleviate the Donsker class condition.

2.3 Cross-fitting

We give a construction of the sample splitting used to the define the cross-fitted one-step estimator as it is given in Manuscript I, II and II, where the EIF is linear in the target parameter. For a construction of a general one-step estimator see Kennedy (2022) and Chernozhukov et al. (2018). Denote the observed data $\mathcal{O}_n$, which consists of n i.i.d. observations from the sample space. Split the data into K approximately equally large folds. It is important that K does not depend on the sample size (which would be the case for leave-one-out cross-fitting). Formally, define the index vector $(i_1, \ldots, i_n)$ by n draws from a multinomial distribution with K events and event probabilities $p_k = \frac{1}{K}, k = 1, \ldots, K$. Define the k'th index set $\mathcal{T}_k = \{i_j : j = k\}$ and let $\mathcal{V}_k = \{o_{i_j} \in \mathcal{O}_n, j = 1, \ldots, n : i_j \in \mathcal{T}_k\}$ be the corresponding k'th fold. Define $\mathcal{V}_{-k} = \cup_{l \neq k} \mathcal{V}_l$ as the set of the observations not included in the k'th fold and let $\hat{P}_{-k}$ be the estimate of P_0 obtained from $\mathcal{V}_{-k}$. Furthermore, let $\mathbb{P}_n^k$ be the empirical measure of the observations in $\mathcal{V}_k$, and let n_k be the number of observations in the k'th fold. The cross-fitted one-step estimator is then defined as

$$\hat{\psi}^{CF} = \sum_{k=1}^{K} \frac{n_k}{n} \mathbb{P}_n^k \varphi_{\hat{P}_{-k}}.$$

Consider the decomposition

$$\begin{aligned}
&\mathbb{P}_n^k \varphi_{\hat{P},-k} \\
=&(\mathbb{P}_n^k - P_0)\varphi_{\hat{P},-k} + P_0 \varphi_{\hat{P},-k} \\
=&(\mathbb{P}_n^k - P_0)(\varphi_{\hat{P},-k} - \varphi_{P_0}) + (\mathbb{P}_n^k - P_0)\varphi_{P_0} + P_0 \varphi_{\hat{P}_{-k}} \\
=&\mathbb{P}_n^k \tilde{\psi}_{P_0} + (\mathbb{P}_n^k - P_0)(\varphi_{\hat{P}_{-k}} - \varphi_{P_0}) + P_0 \varphi_{\hat{P}_{-k}},
\end{aligned}$$

from which we have the following decomposition of the difference $\hat{\psi}^{CF} - \psi(P_0)$, analogous to the one-step estimator.

$$\begin{aligned}&\hat{\psi}^{CF} - \psi(P_0)\\ =&\mathbb{P}_n\tilde{\psi}_{P_0} + \sum_{k=1}^{K}\frac{n_k}{n}\underbrace{(\mathbb{P}_n^k - P_0)(\varphi_{\hat{P}_{-k}} - \varphi_{P_0})}_{k\text{'th empirical process term}} + \sum_{k=1}^{K}\frac{n_k}{n}\underbrace{(P_0\varphi_{\hat{P}_{-k}} - \psi(P_0))}_{k'\text{'th remainder term}}.\end{aligned}$$

By Lemma 2 in Kennedy et al. (2020), the k'th empirical process term is $o_p(n^{-1/2})$ if $\left\|\varphi_{\hat{P}_{-k}} - \varphi_{P_0}\right\| = o_p(1)$. Since $\frac{n_k}{n} \overset{P}{\to} 1/K$, it follows by the continuous mapping theorem, that $\hat{\psi}^{CF}$ is asymptotically linear with the EIF as its influence function if $\varphi_{\hat{P}_{-k}}$ is consistent in $L_2(P_0)$-norm for each k and if the $k'th$ remainder term is $o_p(n^{-1/2})$ for each k.

We define the cross-fitted variance estimator as

$$\hat{\sigma}^{2,CF} = \sum_{k=1}^{K}\frac{n_k}{n}\mathbb{P}_n^k(\varphi_{\hat{P}_{-k}} - \hat{\psi}^{CF})^2.$$

Lemma 1 in Manuscript II shows that the cross-fitted variance estimator is consistent under the same assumptions as required for asymptotic linearity of the cross-fitted one-step estimator. The standard error $\sqrt{\hat{\sigma}^{2,CF}/n}$ of $\hat{\psi}^{CF}$ can thus be used for the construction of confidence intervals and statistical tests.

Figure 2.1 shows the estimated bias of one-step estimators with and without cross-fitting resulting from a simulation study in Manuscript III (corresponding to the upper left panel in Figure 1 in the manuscript). The simulation study is based on 1000 simulations for sample sizes $n = 250, 500, 750, 1000$, respectively, and it relates to the estimation of a best linear projection, to be discussed in Chapter 4. Two different nuisance parameter estimators were considered, where **cor** corresponds to correctly specified parametric models and **RF** corresponds to random forest. The suffix **CF** indicates that the cross-fitted was used in combination with the given nuisance parameter estimator. From the figure it is clear that cross-fitting is needed when considering complex data-adaptive estimators, such as random forest.

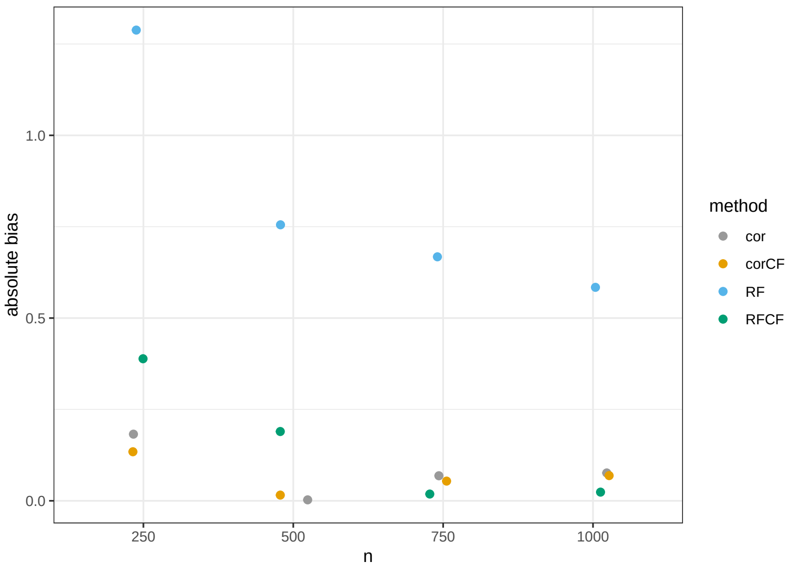


Figure 2.1: Absolute bias from the simulation study in Ziersen and Martinussen (2024a). The cor and corCF correspond to one-step estimators with correctly specified (semi)parametric nuisance parameter estimators and with and without cross-fitting, respectively. RF and RFCF correspond to one-step estimators with nuisance parameters estimated by random forest and with and without cross-fitting, respectively.

Chapter 3

Average treatment effect with censoring and competing risk

In chapter 2, the average treatment effect (ATE) was defined as

$$\psi(P_0) = E_{P_0}\{E_{P_0}(Y \mid A = 1, X) - E_{P_0}(Y \mid A = 0, X)\},$$

where (Y, A, X) denotes the outcome, treatment and covariate, respectively. The target parameter $\psi(P_0)$ is defined on the observed data, and from the definition above, it is not obvious how $\psi(P_0)$ defines a causal parameter, or which assumptions (if any) are needed in order to give it a causal interpretation.

This chapter defines the ATE as a causal parameter in a specific data setting, where the outcome is not fully observed due to censoring. For clarity, the ATE is first introduced in the simple data structure from the example above.

We introduce the counterfactual framework (sometimes also referred to as potential outcomes) (Neyman, 1923, Rubin, 1974, Robins, 1986 and Hernán and Robins, 2010). For an outcome $Y \in \mathbb{R}$ and a treatment variable $A \in \{0, 1\}$, let Y^a denote the outcome Y under treatment a. In a given study (whether randomized or observational), only the pair (Y, A) is observed for each patient (we will tend to switch between the terminology *patient* and *observation*), and Y^a can be read as "the outcome one would have observed, had the patient, possibly contrary to the fact, received treatment a", thus giving rise to the name *counterfactual outcome*. Let X be a covariate vector and assume that the *full* data is given by independent replications of $(Y^0, Y^1, A, X) \sim \tilde{P}$, where $\tilde{P}$ is assumed to belong to some appropriate model. The ATE is defined as

$$E_{\tilde{P}}\{Y^1\} - E_{\tilde{P}}\{Y^0\}.$$

Assume that the observed data $\mathcal{O}_n$ are given by independent replications of $(Y, A, X) \sim P_0$. We make the following assumptions for identifying the ATE based on the observed data.

Assumption A (Identification)**.**

A1 (Consistency) $Y = AY^1 + (1 - A)Y^0$ *conditional on* A.

A2 (Exchangeability) $Y^a \perp\!\!\!\perp A \mid X, \ A = 0, 1.$

A3 (Positivity) $P_0(A = a \mid X = x) > \eta > 0, \ \forall x, \ a = 0, 1.$

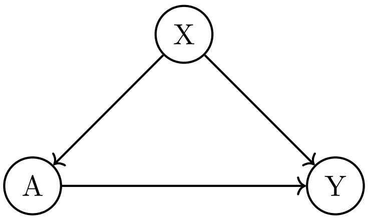


Figure 3.1: Example of a confounder.

The consistency assumptions states that the observed outcome for a given patient is indeed the counterfactual outcome under the observed treatment assignment for that patient.

The exchangeability assumption states that the counterfactual outcome is independent of the treatment assignment given the covariate X. The assumption is sometimes also referred to as the assumptions of *no unmeasured confounders*, meaning that we have observed all common causes of the treatment assignment and the outcome. Here, common cause refers to the graphical representation of causal effects as directed acyclic graphs (Pearl, 2009), see Figure 3.1. The graphical representation does not rely on counterfactuals and it is included here as an intuitive tool for explaining the exchangeability assumption, only. The correct formulation in the counterfactual framework is indeed assumption A2, but some authors have connected the counterfactuals with the graphical framework through *single world intervention graphs* (Richardson and Robins, 2013).

Finally, the positivity assumptions states that all patients must be susceptible to either treatment.

Under assumption A, the ATE is identified from the observed data by the g-formula (Robins, 1986) as

$$\psi(P_0) = \mathrm{E}_{P_0}\{E_{P_0}(Y \mid A = 1, X) - E_{P_0}(Y \mid A = 0, X)\} = E_{\tilde{P}}\{Y^1\} - E_{\tilde{P}}\{Y^0\}.$$

Given the identification formula, estimation and inference of the ATE can be carried out in the observed data, using the methods from chapter 2 to estimate $\psi(P_0)$. The identification problem can be seen as a type of incomplete data problem, where the outcome of interest (the counterfactuals Y^a) is not fully observed. Another type of incomplete data is known from survival analysis, where the time to death is not fully observed due to some censoring mechanism. One common example of censoring in time-to-event studies is the end of follow-up. When the study terminates, the only information available for patients still alive is that their survival time is at least greater than the length of study (or some other time-scale). The following sections defines the average treatment in a time-to-event setting with censored data. The setting is further complicated by competing risks, which may prevent observation of the event of interest, even when data are uncensored.

3.1 Absolute risk with censored data

Let $T \in \mathbb{R}_+$ be the time to event and let $\Delta \in \{1, 2\}$ denote the event indicator. Throughout, $\Delta = 1$ denotes the event of interest and $\Delta = 2$ denotes the competing event. Let $C \in \mathbb{R}_+$ be the time to censoring, $A \in \{0, 1\}$ a baseline treatment, and $X \in \mathcal{X}$ a d-dimensional covariate vector in some sample space $\mathcal{X}$. The observed event time is $\tilde{T} = T \wedge C$ and the observed event indicator is $\tilde{\Delta} = \mathbb{1}(T \leq C)\Delta$. Let

$Y_j(t) = \mathbb{1}(T \leq t, \Delta = j)$, $j = 1, 2$, denote the outcome defined in the uncensored observed event time and event indicator for event j. That is, whether or not event j has occurred in the time-horizon $[0, t]$. Let T^a and Δ^a denote the counterfactual event time and event indicator, respectively, and define the counterfactual outcomes $Y_j^a(t)$ as $Y_j(t)$ under treament a. The full data are represented by independent draws of $(T^0, T^1, \Delta^0, \Delta^1, C, A, X) \sim \tilde{P}$ and the observed data is given by replications of $(\tilde{T}, \tilde{\Delta}, A, X) \sim P_0$. The ATE is now defined for the event of interest in a specific time-horizon $[0, t]$ as

$$E_{\tilde{P}}\{Y_1^1(t)\} - E_{\tilde{P}}\{Y_1^0(t)\} = \tilde{P}(T^1 \leq t, \Delta^1 = 1) - \tilde{P}(T^0 \leq t, \Delta^0 = 1), \tag{3.1}$$

i.e., the difference in the absolute risk of event $\Delta = 1$ in the time-horizon $[0, t]$. Note that the ATE is only defined in terms of the effect of a baseline treatment assignment. In time-to-event studies, patients may switch treatment before an observation of the event time T or they might terminate treatment all together due to non-compliance or side-effects. The ATE defined above only measures the effect of treatment assignment at baseline and it can be interpreted as an *intention to treat*-effect (Hernán and Robins, 2010).

With censored data, methods from survival analysis that allow for competing risks (Andersen et al., 1993, Martinussen and Scheike, 2006) are combined with the g-formula to identify the ATE from the observed data (Gill et al., 1997, van der Laan and Robins, 2003, Ozenne et al., 2020, Rytgaard et al., 2023). We take a step back from the causal framework and give a short introduction to some results from survival analysis. For a rigorous treatment we refer to Andersen et al. (1993) and Martinussen and Scheike (2006). It turns out that a counting process framework is convenient for describing time-to-event data and connecting the censored event times with the unobserved data. Let $N_j(t) = \mathbb{1}(T \leq t, \Delta = j)$ be the counting process for the j'th event and let $\lambda_j(t \mid a, x)$ denote the conditional cause-specific hazard for event j defined by

$$\lambda_j(t \mid a, x) = \lim_{h \to 0} \frac{P(T \in [t, t+h), \Delta = j \mid A = a, X = x)}{h}.$$

Let $\Lambda_j(t \mid a, x) = \int_0^t \lambda_j(s \mid a, x)\, \mathrm{d}s$ be the conditional cumulative hazard for event j. By the Doob-Meyer decomposition there exist a martingale $M_j(t \mid a, x)$ such that

$$M_j(t \mid a, x) = N_j(t) - \mathbb{1}(T \geq t)\Lambda_j(t \mid a, x)$$

and we say that $N_j(t)$ has compensator $\mathbb{1}(T \geq t)\Lambda_j(t \mid a, x)$ (the martingale is defined w.r.t. the history generated by N_1, N_2 but it is left out for notational convenience). Letting $S(t \mid a, x) = P(T > t \mid A = a, X = x) = \exp(-\Lambda_1(t \mid a, x) - \Lambda_2(t \mid a, x))$ denote the conditional survival function, it follows that the conditional cumulative incidence function for the j'th event is given by

$$F_j(t \mid a, x) \equiv P(T \leq t, \Delta = j \mid A = a, X = x) = \int_0^t S(s \mid a, x)\Lambda_j(ds \mid a, x).$$

To see how this connects to the censoring, let $\tilde{N}_j(t) = \mathbb{1}(\tilde{T} \leq t, \tilde{\Delta} = j)$ be the j'th counting process in the censored data and let $\tilde{\Lambda}_j(t \mid a, x)$ be the associated conditional cumulative hazard function in the censored data. Assuming conditional independent censoring, there exist a martingale $\tilde{M}(t \mid a, x)$ such that

$$\tilde{M}(t \mid a, x) = \tilde{N}_j(t) - \mathbb{1}(\tilde{T} \geq t)\Lambda(t, \mid a, x).$$

Hence, the compensator of the observed counting process $\tilde{N}_j(t)$ only differs from the compensator of $N_j(t)$ by the at-risk process $\mathbb{1}(\tilde{T} \geq t)$ and it follows that the absolute risk of event j in time-horizon $[0, t]$ in the uncensored data is identified from the observed data by

$$P(\tilde{T} \leq t, \tilde{\Delta} = j \mid a, x) = \int_0^t S(t \mid a, x)\Lambda_j(ds \mid a, x) = F_j(t \mid a, x).$$

Returning to the causal framework, we can now update the identifiability assumptions for the censored data setting.

Assumption B (Identification)**.**

A1 (Consistency) $Y_1(t) = AY_1^1(t) + (1 - A)Y_1^0(t)$ *conditional on* A*.*

A2 (Exchangeability) $Y_1^a(t) \perp\!\!\!\perp A \mid X,\ A = 0, 1.$

A3 (Positivity) $P_0(A = a \mid X = x)P(C > s \mid a, x) > \eta > 0,\ \forall x, s \in \mathcal{X} \times [0, t]\ a = 0, 1.$

A4 (Independent censoring) $T \perp\!\!\!\perp C \mid A, X.$

Under assumption B, the ATE in (3.1) is identified in the observed data as

$$\psi(P_0) = E_{P_0}\{F_1(t \mid 1, X) - F_1(t \mid 0, X)\} = E_{\tilde{P}}\{Y_1^1(t)\} - E_{\tilde{P}}\{Y_1^0(t)\}, \tag{3.2}$$

where the cumulative hazards in F_1 are defined w.r.t. P_0 (see supplementary material of Rytgaard et al., 2023).

3.2 Number of life-years lost due to a specific event

In the survival setting, i.e. without competing risks, an analogous ATE can be defined as

$$\begin{aligned} E_{\tilde{P}}\{\mathbb{1}(T^1 > t)\} - E_{\tilde{P}}\{\mathbb{1}(T^0 > t)\} &= \tilde{P}(T^1 > t) - \tilde{P}(T^0 > t) \\ &= E_{P_0}\{S(t \mid A = 1, X) - S(t \mid A = 0, X)\}, \end{aligned}$$

which is the difference in survival probabilities at time t, where the identification in the observed data is given by assumption B without competing risk (Westling et al., 2023). Another popular estimand is the difference in *restricted mean survival time* (RMST) in the time horizon [0,t] defined as

$$E_{\tilde{P}}\{T^1 \wedge t\} - E_{\tilde{P}}\{T^0 \wedge t\} = E_{P_0}\left\{\int_0^t S(s \mid A = 1, X)\,\mathrm{d}s - \int_0^t S(s \mid A = 0, X)\,\mathrm{d}s\right\}.$$

See e.g. Cui et al. (2023) for an example of a generalized random forest approach for causal inference based on RMST. The RMST has an interpretation of the treatment effect in terms of time-scale on which the data is observed. For example, if the study measures the time in days from inclusion until death or censoring, the ATE based on the RMST has the interpretation ”in the time-span [0,t], patients receiving treatment $A = 1$ lived x days longer compared to patients receiving treatment $A = 0$”. The interpretation makes the RMST based ATE attractive, and in Manuscript III, we derive an analogous target parameter in the competing risk setting based on Andersen (2013). We give a summary of derivation from Manuscript III.

Returning to the competing risk setting, Andersen (2013) shows that

$$t - \int_0^t S(s)\,\mathrm{d}s.$$

can be interpreted as the expected number of life years lost before time t. Hence

$$L(0,t \mid a,x) = t - \int_0^t S(s \mid a,x)\,\mathrm{d}s.$$

has interpretation as the expected number of life years lost before time t in strata (a,x). The function can be decomposed into

$$L(0,t \mid a,x) = L_1(0,t \mid a,x) + L_2(0,t \mid a,x)$$

where

$$L_j(0,t \mid a,x) = \int_0^t F_j(s \mid a,x)\,\mathrm{d}s$$

has the interpretation as the *number of life years lost before time t due to event j* (Andersen, 2013). To define the ATE with the same interpretation, we introduce T_j as the time to the j'th event. As noted in Andersen (2013), T_j is improper as $P(T_j = \infty) > 0$, but $T_j \wedge t$ is proper with conditional expectation

$$E(T_j \wedge t \mid A = a, X = x) = t - \int_0^t F_j(s \mid a,x)\,\mathrm{d}s.$$

Define $Y_j(t) = t - T_j \wedge t$ and let the counterfactual outcome be given by $Y_j^a(t) = t - T_j^a \wedge t$ for $a = 0,1$. As is shown in Manuscript III, the ATE defined by $Y_j^a(t)$ can be identified from the observed data under assumption B by

$$\psi(P_0) = E_{P_0}\{L_1(0,t \mid 1,X) - L_1(0,t \mid 0,X)\}. \tag{3.3}$$

The ATE has the interpretation as the difference in the expected number of life years lost due to event 1 before time t. As with the ATE defined by the RMST in the survival setting, the ATE defined above returns the interpretation to the time-scale of the study, which may be easier to communicate than probabilities. As the ATE defined here is stated in terms of differences of the area under the cumulative incidence functions, it is also not as sensitive to the choice of time-horizon for detecting potentially early effects of treatment. As a crude example, imagine that the difference in absolute risk is large for some period in $[0,t]$ but 0 at time t. The ATE defined on absolute risk is 0, but the ATE defined on the number of life years lost before time t due to event j will detect the early differences in absolute risk.

3.3 Estimation

To distinguish the ATE's defined in (3.2) and (3.3), let $\psi_{AR}(P_0)$ be the target parameter given in (3.2) and let $\psi_{LYL}(P_0)$ be given by (3.3). The cross-fitted one-step estimator from chapter 2 is defined by the EIF corresponding to the target parameter in question. The EIF's corresponding to the two target parameters are similar, and for the sake of illustration, we will only include the EIF corresponding to $\psi_{LYL}(P_0)$ (the EIF corresponding to $\psi_{AR}(P_0)$ in a discrete time setting is given in e.g. Moore and van der Laan, 2009, van der Laan and Rose, 2011, and in a continuous time setting in Rytgaard et al., 2023). Specifically, they can both be parametrized be the same nuisance parameter, and the general construction of the cross-fitted estimator is analogous.

The EIF corresponding to $\psi_{LYL}(P_0)$ is derived in Manuscript III and it is parametrized by the nuisance parameter $\nu = (\Lambda_1, \Lambda_2, \Lambda_c, \pi)$, where Λ_c is the conditional cumulative hazard function for the censoring time and $\pi(a \mid X = x) = P(A = a \mid X = x)$. Define $\tau_1(x) = L_1(0, t \mid 1, x) - L_1(0, t \mid 0, x)$ for a given time-horizon $[0, t]$. The EIF is given by

$$\tilde{\psi}_{P_0} = \varphi_\nu - \psi^{LYL}(P_0),$$

where φ_ν is a real-valued function defined on the sample space $\mathcal{O}$ of $(\tilde{T}, \tilde{\Delta}, A, X)$ at a given value of ν with

$$\begin{aligned} &\varphi_\nu(O) \\ =&\tau_1(X) + \left(\frac{\mathbb{1}(A=1)}{\pi(1 \mid X)} - \frac{\mathbb{1}(A=0)}{\pi(0 \mid X)}\right)\left\{\sum_{i=1,2}\int_0^{t^*} \frac{H_{i1}(s, t^* \mid A, X)}{S_C(s \mid A, X)}\,\mathrm{d}M_i(s \mid A, X)\right\}\end{aligned}$$

where

$$H_{ij}(s, t \mid a, x) = \int_s^t \mathbb{1}(i = j) + \frac{F_j(s \mid a, x) - F_j(u \mid a, x)}{S(s \mid a, x)}\,\mathrm{d}u.$$

Now, the cross-fitted one-step estimator is defined through $\varphi_{\hat{\nu}}$ for some chosen nuisance parameter estimator $\hat{\nu}$. In Manuscript I, we consider estimation of $\psi_{AR}(P_0)$ in a high-dimensional covariate setting, and $\hat{\nu}$ is based on penalized regression models in order to alleviate the high-dimensional setting. For the cumulative hazard functions, Cox regressions with elastic net penalization (Zou and Hastie, 2005, Wu, 2012) were used, which combines lasso and ridge regression. In related studies with high-dimensional confounders and censored data, lasso penalization were shown to work well for nuisance parameter estimation, as long as the underlying regression functions are moderately sparse (Hou et al., 2021). This was confirmed through a simulation study in Manuscript III, but similar theoretical results are still unknown for estimation of $\psi_{AR}(P_0)$.

Chapter 4

Heterogeneity

The ATE defines the effect of a treatment on population level, but some patient may react differently to the treatment. To define the treatment effect for a given patient, we introduce the *conditional average treatment effect* (CATE). In chapter 3, we saw different definitions of the ATE according to different data structures and different desired interpretations of the treatment effect. In the simple data setting, (Y, A, X), the ATE was given by $E\{E(Y \mid A = 1, X) - E(Y \mid A = 0, X)\}$ and the corresponding CATE is defined as

$$\tau(x) = E(Y \mid A = 1, X = x) - E(Y \mid A = 0, X = x).$$

In time-to-event data with competing risks, the CATE corresponding to the absolute risk ATE, $\psi_{AR}(P_0)$, is

$$\tau(x) = F_1(t \mid 1, X = x) - F_1(t \mid 0, X = x)$$

and for $\psi_{LYL}(P_0)$, the CATE is

$$\tau(x) = L_1(0, t \mid 1, X = x) - L_1(0, t \mid 0, X = x).$$

For each definition of $\tau(x)$, the corresponding ATE is given by $E\{\tau(X)\}$, and for the rest of this chapter we work with a general $\tau(x)$, where its definition is taken implicitly from the context unless otherwise specified.

Estimation of the CATE using data-adaptive methods have received much attention in recent years. Wager and Athey (2018) develops an estimator for the CATE function using random forests in the simple setting without censoring, and their approach is extended to survival data in Cui et al. (2023). Hu et al. (2021) compares different machine-learning methods for CATE estimation with survival data. van der Laan (2006) and Semenova and Chernozhukov (2021) use a certain type of linear projection of a pseudo-outcome, related to the EIF, to obtain inference on a target a target function, where $\tau(x)$ is a special case. Kennedy (2023) develops certain meta-learners akin to van der Laan (2006), and the optimality of CATE estimation in terms of minimax convergence rates for certain function classes are derived in Kennedy et al. (2024), and Xu et al. (2023) compares different meta-learnes for CATE estimation with survival data.

An estimate $\hat{\tau}$ can be used to predict the expected treatment effect for a single individual, which can be used to guide treatment decisions for a given patient. But, when the estimation procedure used to obtain $\hat{\tau}$ involves complex meta-learners and machine-learning methods, the estimate itself gives little inside into the driving features of the potential treatment effect heterogeneity. By understanding which subgroups of patients respond differently to the treatment, one gains inside that can be used in treatment plans or to inform the underlying pharmacology of the treatment and its interaction with certain patient features.

This chapter introduces two different approaches for identifying patient features that drives the underlying treatment effect heterogeneity. The two approaches are developed in Manuscript II in the survival setting and the second approach is extended to the CATE defined by the number of life years lost due to a specific event in Manuscript III.

4.1 Treatment effect variable importance measure

Treatment effect variable importance is defined in a number of different ways in the causal inference literature. van der Laan (2006) and Semenova and Chernozhukov (2021) defines variable importance through a target function $g(v) = E(\tau(X) \mid V = v)$, where V is a subset of the covariate vector X. As such, their approach can be viewed as a certain type of subgroup analysis, where the ATE is estimated for a certain patient group with $V = v$. Levy et al. (2021) defines a measure of treatment effect heterogeneity by $\mathrm{var}\{\tau(X)\}$, where the intuition is that $\mathrm{var}\{\tau(X)\}$ should be small for low levels of heterogeneity and big for high levels. As an example, if $\tau(x)$ is constant then $\mathrm{var}(\tau(X)) = 0$. Hines, Dukes, et al. (2022) builds on the this idea and defines a treatment effect variable importance measure by a nonparametric ANOVA/R^2 analog of the CATE, and in Manuscript II, we extend the work of Hines, Dukes, et al. (2022) to a survival setting, considering the CATE defined by the survival function and RMST, repsectively. We give a short introduction to the variable importance measure in the general definition, and in Section 4.3 we discuss the extension to survival data in terms of estimation.

We consider a d-dimensional covariate X, and for a given subset $l \in \{1, \dots, d\}$ define $\tau_l(x_{-l}) = E(\tau(x) \mid X_{-l} = x_{-l})$, where X_{-l} are the covariates with an index not in l. A measure of variable importance for X_l can be defined as

$$\Theta_l(P_0) = \mathrm{var}(\tau(X)) - \mathrm{var}(\tau_l(X_{-l})) \geq 0.$$

The parameter $\Theta_l(P_0)$ can be interpreted as the amount of heterogeneity not already explained by X_{-l}, and it is large when a large amount of the total heterogeneity, $\mathrm{var}(\tau(X))$ is explained by X_l. The target parameter considered in Hines, Dukes, et al. (2022) and Manuscript II is a scaled version of Θ_l given by

$$\Psi_l(P_0) = \frac{\Theta_l}{\mathrm{var}(\tau(X))} = 1 - \frac{\mathrm{var}(\tau_l(X_{-l}))}{\mathrm{var}(\tau(X))} \in [0, 1).$$

The parameter $\Psi_l(P_0)$ is interpreted as the proportion of the total heterogeneity explained by X_l. Considering different groups of covariates, their importance in terms of explaining the potential treatment effect heterogeneity can be ranked by estimates of $\Psi_l(P_0)$ for l varying across the corresponding covariate groups, where the highest importance is given to groups with the largest estimate of $\Psi_l(P_0)$.

4.2 Best partially linear projection

Another example of variable importance measures (or treatment effect modifiers) are given by the *best linear projection* of the CATE. The resulting coefficients are then given a regression type interpretation as an association between the CATE and a given covariate. As already mentioned, van der Laan (2006) and Semenova and Chernozhukov (2021) use best linear projections of the CATE function to approximate a target function, whereas the Boileau et al. (2023) and Cui et al. (2023) use best linear projections of the CATE functions in order to give regression-like interpretation of the CATE function - even when it is estimated using data-adaptive methods. Manuscript II contributes to the latter approach by defining the *best partially linear projection* of the CATE function in the spirit of the assumption-lean inference approach by Vansteelandt and Dukes (2022), developed for survival data. The approach is extended in Manuscript III to cover the competing risk setting defined by the number of life years lost due to a specific event.

We outline the concept from Manuscript II for a general $\tau(x)$ and discuss its extension to censored data in Section 4.3. For a single covariate X_j, define

$$\Omega_j(P_0) = \frac{E\{\operatorname{cov}(X_j, \tau(X) \mid X_{-j})\}}{E\{\operatorname{var}(X_j \mid X_{-j})\}},$$

where X_{-j} are the covariates different from X_j. The parameter $\Omega_j(P_0)$ can be interpreted as a weighted average of the conditional association of $\tau(X)$ and X_j given X_{-j}. As the parameter is scale sensitive, ranking of variable importance of different covariates are based on the p-value associated with the test of the hypothesis $H : \Omega_j(P_0) = 0$. To further motivate the parameter, define (without loss of generality) the CATE function as

$$\tau(x) = \beta x_j + w(x_j) + R_1^{\beta,w}(x_j, x_{-j}),$$

where β is a real valued coefficient and w and $R_1^{\beta,w}$ are a measurable functions with finite variance. Define the best partially linear projection by

$$(\beta^*, w^*) = \operatorname*{arg\,min}_{\beta,w} E\{R_1^{\beta,w}(X_j, X_{-j})^2\} = \operatorname*{arg\,min}_{\beta,w} E\{(\tau(X) - \beta X_j - w(X_{-j}))^2\}.$$

Then $\beta^* = \Omega_j(P_0)$ (see Appendix A in Manuscript II). Hence, the parameter $\Omega_j(P_0)$ can be interpreted as a measure of association between $\tau(X)$ and X_j that minimizes the heterogeneity otherwise explained by interactions of X_j and X_{-j}. If the partially linear model holds for the CATE function, the parameter $\Omega_j(P_0)$ has an exact interpretation as a regression coefficient. When the model does not hold, it is seen from the definition, that $\Omega_j(P_0)$ still has interpretation as measure of treatment effect variable importance.

A comparison of the best partially linear projection and the best linear projection can be given in terms of their respective remainder terms. Define the CATE function as

$$\tau(x) = \gamma + \alpha^T x + R_2^{\gamma,\alpha}(x)$$

for $(\gamma, \alpha) \in \mathbb{R} \times \mathbb{R}^d$, and define the beast linear projection by

$$(\gamma^*, \alpha^*) = \operatorname*{arg\,min}_{\gamma,\alpha} E\{R_2^{\gamma,\alpha}(X)^2\} = \operatorname*{arg\,min}_{\gamma,\alpha} E\{(\tau(X) - \gamma - \alpha^T X)^2\}.$$

Since the linear model is a subspace of the partially linear model, standard Hilbert space geometry gives that $\left\|R_1^{\beta^*,w^*}\right\| \leq \left\|R_2^{\gamma^*,\alpha^*}\right\|$ (Tsiatis, 2006, ch. 2, and Manuscript II, Appendix A). Heuristically, $\left\|R_1^{\beta,w}\right\|$ and $\|R_2^{\gamma,\alpha}\|$ define the distance from τ to functions in the partially linear and linear model, respectively. By definition of projections, the function $x \mapsto x_j\beta^* + w^*(x_{-j})$ minimises the distance from τ to the partially linear model, and since all linear functions are special cases of partially linear functions, the distance $\left\|R_1^{\beta^*,w^*}\right\|$ is smaller than the distance $\left\|R_2^{\gamma^*,\alpha^*}\right\|$. Hence, the best partially linear projection gives a smaller error in terms of measuring the treatment effect variable importance of X_j compared to the best linear projection.

4.3 Estimation

Estimation of the target parameters $\Psi_l(P_0)$ and $\Omega_j(P_0)$ follow the cross-fitted one-step estimator approach. The difference from earlier is that the parameters are defined as ratios, and the corresponding EIF's are no longer linear in the target parameter. But, if the estimators of two target parameters are asymptotically linear with their corresponding EIF's as their influence functions, then the ratio of the estimators is also asymptotically linear with its influence function given by the EIF corresponding to the ratio of the target parameters (van der Vaart, 2000, ch. 25.7). Hence, estimation of $\Psi_l(P_0)$ and $\Omega_j(P_0)$ will follow by the ratio of cross-fitted one-step estimators of the involved parameters.

We make a shift in notation and denote the EIF corresponding to a parameter $\psi(P_0)$ by $\tilde{\psi}_\psi$, where the dependence of P_0 is implicitly understood. We proceed with the construction of an estimator for $\Psi_l(P_0)$. Let φ_{P_0} denote the uncentered EIF of $E_{P_0}\{\tau(X)\}$. The EIF of $\Theta_l(P_0)$ is given by

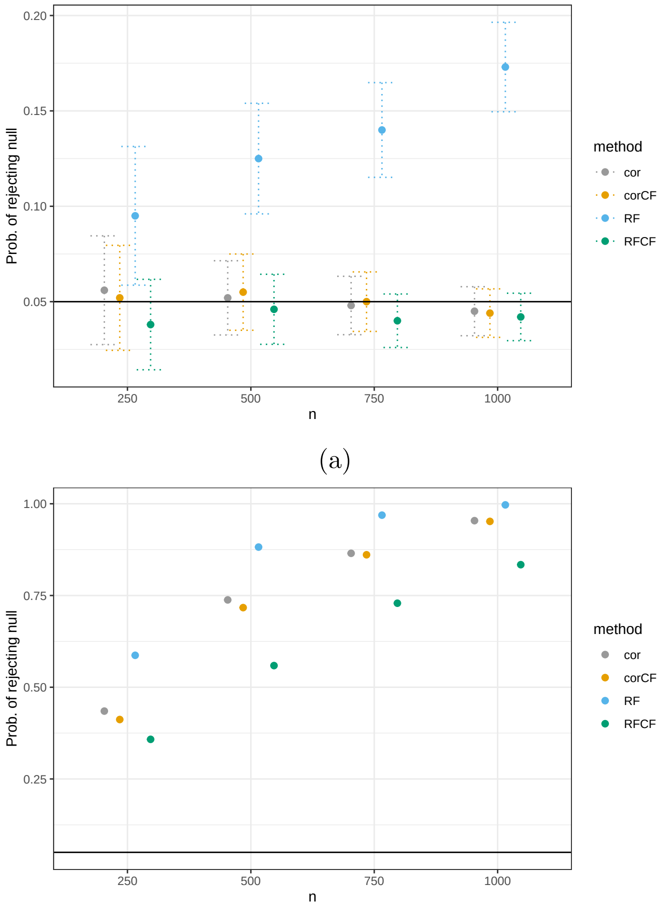
$$\tilde{\psi}_{\Theta_l} = (\varphi_{P_0} - \tau_l)^2 - (\varphi_{P_0} - \tau)^2 - \Theta_l(P_0).$$

With a slight abuse of notation, denote $\Theta_d(P_0) = \text{var}(\tau(X))$ and $\tau_d(P_0) = E\{\tau(X)\}$. The EIF of $\Theta_d(P_0)$ is given by

$$\tilde{\psi}_{\Theta_d} = (\varphi_{P_0} - \tau_d(P_0))^2 - (\varphi_{P_0} - \tau)^2 - \Theta_d(P_0).$$

The EIF's above are defined for a general $\tau(x)$, and they are stated in terms of the uncentered EIF for the ATE associated with $\tau(x)$. Hence, once the EIF for the ATE is known for a given $\tau(x)$, the EIF's of Θ_l and Θ_d are immediately given. Since both EIF's are linear in the corresponding target parameter, the cross-fitted one-step approach as defined in chapter 2 can be employed. The nuisance parameters are now comprised of the nuisance parameters associated with φ and τ_l. In the survival setting, the nuisance parameters are given by $\nu = (\Lambda, \Lambda_c, \pi, \tau_l)$. It is difficult to use an off-the-shelf machine-learning method for estimating τ_l, and in Manuscript II, we obtain an estimate by regressing an predictions from an estimate of the CATE function $\hat{\tau}(X)$ onto X_{-l}, using some data-adaptive method, in line with the recommendations in Hines, Dukes, et al. (2022). Finally, an estimate of $\Psi_l(P_0)$ is obtained by a ratio of the cross-fitted one-step estimators of $\Theta_l(P_0)$ and $\Theta_d(P_0)$. The standard error is calculated by the cross-fitted variance estimator from chapter 2.

Estimation of $\Omega_j(P_0)$ follows the exact same steps and we leave out the details and refer to Manuscript II for a detailed derivation. The only difference is that the nuisance parameters are comprised of the nuisance parameters associated with φ, τ_j and $E(X_j \mid X_{-j} = x_{-j})$. For estimation of the latter, many machine-learning methods exist, as it is just an ordinary regression function. An interesting thing to note regarding estimation of $\Omega_j(P_0)$, is the performance of the test statistic for the test $H : \Omega_j(P_0) = 0$, defined by the cross-fitted one-step estimator and the associated cross-fitted estimator of the standard error. By corollary 4 in Manuscript II, the test statistic is asymptotically standard normal under the same assumptions required for estimation of $\Omega_j(P_0)$, and in Figure 4.1 we see the results of a simulation study regarding the power and type-1 error of the test statistic in the competing risk setting from Manuscript III. When using data-adaptive nuisance estimators in form of random forests, the test statistic associated with the cross-fitted estimator is seen to perform in line with the test statistic defined by correctly specified parametric nuisance estimators in terms of type-1 error. But the power of the test statistic associated with data-adaptive nuisance estimators is smaller compared to the parametric counterparts. We expect this to be a general phenomenon for test statistics estimated with data-adaptive nuisance estimators, since the standard error is estimated by a cross-fitted plug-in estimator, and thus inherits the possibly slow convergence rates coming from data-adaptive nuisance estimating. An interesting avenue for future research would be to derive cross-fitted one-step estimators for standard error estimation, with the aim of regaining some of the lost power.



(a)

(b)

Figure 4.1: Type-1 error (panel a) and power (panel b) of estimators of Ω_j^l from the simulation study in Ziersen and Martinussen (2024a) based on 1000 simulations. The cor and corCF correspond to one-step estimators with correctly specified (semi)parametric nuisance parameter estimators and with and without cross-fitting, respectively. RF and RFCF correspond to one-step estimators with nuisance parameters estimated by random forest and with and without cross-fitting, respectively.

Chapter 5

BrainDrugs - application to Danish registers

The inspiration for the methodological work carried out in the Manuscripts and described in the preceding chapters was heavily laid by research conducted in BrainDrugs and in particular the study presented in Kessing et al. (2024). In this chapter, we outline this study and discuss the application of the methods derived in Manuscripts I-III and outlined in the previous chapters.

The aim of the study in Kessing et al. (2024) is to compare the response to treatment with different antidepressants in patients with MDD, using data from the Danish national registers. The registers hold information on patients medical histories in the form of hospital admissions with a given diagnosis as well as prescription based drug purchases at Danish pharmacies. Additionally, the registers hold information on age and sex. The study includes patients between 1995 and 2018 with their first diagnosis with MDD at a Danish psychiatric hospital. After discharge, the first purchase with an antidepressant defines the patients baseline treatment and the index date is defined at the date of purchase.

As the aim of the study is to compare the response to treatment with different antidepressants, an outcome is needed to reflect the response. Unfortunately, the registers do not hold direct information on response to antidepressants in terms of, e.g., HAM-D scores. Instead an event of non-response was defined as a switch to or add on of another antidepressant, antipsychotic medication or lithium, or readmission at a Danish psychiatric ward. The outcome of interest is then given as the time from the index date to non-response. Competing risks were defined as a diagnosis with bipolar disorder, schizophrenia or organic mental disorder, or death. Censoring was defined at emigration or end of follow-up 2017-12-31. Patients were followed from inclusion until an event of non-response, competing event, or censoring.

The target parameter for comparing non-response between two antidepressants was chosen as the ATE based on absolute risk of non-response (Chapter 3.1) at two years after index date. Hence, when comparing two drugs, the one with the lowest risk of non-response is the advantageous one. In order to control for confounding by indication, antidepressants were grouped into five classes based on their shared pharmacology and only antidepressants in the same group were compared. Additionally, patients with prior purchases within a given time-horizon prior to inclusion were excluded from the study in a sensitivity analysis to further account for confounding by indication.

To estimate the ATE, Kessing et al. (2024) used the g-formula based on plug-in estimation with cause-specific hazards for non-response and competing risks estimated by Cox regressions. To account for confounding, the following covariates were included in the Cox regressions: age, sex, psychiatric comorbidity (defined as secondary psychiatric diagnoses at inclusion) and somatic disease histories. The disease histories were defined by a diagnosis within one of nine disease chapters defined by the ICD-10 classifications.

5.1 High-dimensional confounding

The somatic disease histories were included as confounders, but the grouping of specific disease into nine groups was more practical than medically informed. When using an ungrouped definition of the disease histories based on the ICD-10 classification, we observed ~ 1000 different somatic diseases in the study group. For some of the treatment groups in the aforementioned sensitivity analysis, the number of different somatic histories exceeds the number of observations, and a naive inclusion of the ~ 1000 covariates in the Cox regressions is infeasible.

The high-dimensional confounding setting motivated the work in Manuscript I. Here, we examined the finite sample performance of the cross-fitted one-step estimator (Chapter 2.3) when using penalized regressions to handle potentially high-dimensional covariates. The estimator was implemented to recreate the sensitivity analysis from Kessing et al. (2024), taking into account the high-dimensionality of the covariates, including all ungrouped somatic histories. The results were largely unchanged, but the confidence intervals corresponding to the estimated ATE's were generally wider. This is a price to pay, when inference is based on the EIF from a nonparametric model as compared to a smaller parametric model, that assumes the structure of the underlying nuisance parameters.

The work in Manuscript III developed an ATE based on the number of life years lost due to a specific event as an alternative to the absolute risk. As an illustration, the new ATE was used to compare the non-response to Escitalopram and Setraline, respectively. Here, the interpretation of the new estimand was ”the difference in the number of healthy days lost due to non-response”, where ”healthy days” was defined as days without non-response. Accordingly, the estimated ATE based on absolute risk in Kessing et al. (2024) was 0.10 (95% CI: 0.09; 0.12) in favor of Setraline, i.e., the the risk non-response was 10% larger in Escitalopram compared to Setraline. The corresponding estimate based on the number of healthy days lost due to non-response was 49 (95% CI: 40; 58), with the interpretation that patients on Escitalopram lost 49 healthy days more due to non-response before two years compared to patients on Setraline.

5.2 Treatment effect modifiers

One of the key aims of BrainDrugs was to identify patient features determining drug response to antidepressants. This task motivated the work carried out in Manuscript II and III, where Manuscript II developed methods for assessing variable importance in treatment effects with censored data, and Manuscript III extended one of the methods to competing risk data with treatment effect defined based on the number of life years lost due to a specific event. The method was applied to the comparison of Escitalopram and Setraline above, and indicated that age and sex could be treatment modifiers. The parameter estimate for sex was 18.7 (95% CI: 1.3; 36.1), and since the derived variable importance measure has interpretation as a regression coefficient, this translates to the treatment effect being larger among women compared to men. Specifically, the difference in the number of healthy days lost due to non-response between Escitalopram and Setraline was 19 days larger among women compared to men. The interpretation of the estimate as a regression coefficient relies on a partially linear model to hold for the CATE, but as discussed in Chapter 4.2, the estimate continues to provide a measure of the association between sex and treatment effect when the model does not hold.

Chapter 6

Summary of manuscripts

In this chapter, we give a summary of each of Manuscript I, II and III.

Manuscript I *On estimation of the average treatment effect with register data: competing risk and high-dimensional covariates - a case study*

In this Manuscript, the outset is the study in Kessing et al. (2024) and the challenges involved with estimating the ATE based on register data. In register studies, the outcome of interest is often defined as a time-to-event in the presence of competing risks. One only observes a censored version of the outcome and competing risks processes, and the aim of the Manuscript is to provide a methodology applicable to ATE estimation in register studies. We consider the problem of including a potentially high-dimensional covariate to control for confounding in the ATE estimation. Specifically, we use a semiparametric approach and develop an estimator based on the EIF corresponding to the ATE based on the absolute risk at a specific point in time (Chapter 3.1). The EIF is parametrised in the nuisance parameters corresponding to cause-specific hazard functions and a logistic regression, and we proposed penalized methods for nuisance parameter estimation to handle the high-dimensional setting. As advocated in Chernozhukov et al. (2018), nuisance estimators for high-dimensional data, such as lasso, may not belong to a Donsker class and cross-fitting is necessary to provide valid inference. Accordingly, we derive a cross-fitted one-step estimator for the ATE, which takes in penalized regressions for nuisance parameter estimation.

The finite sample performance of the derived estimator is analysed in a simulation study under various settings. The performance of the estimator with and without cross-fitting is contrasted for varying covariate dimensions and in varying covariate sparsity settings. Additionally, varying penalization parameters are chosen for the nuisance estimators according to lasso, ridge and elastic net. The simulations show that cross-fitting enhances the performance of the one-step estimator, in terms of providing valid inference, when lasso or elastic net penalization is used to combat the high-dimensionality of the covariates.

Lastly, the cross-fitted one-step estimator with lasso penalization is used to analyse the data in Kessing et al. (2024), including a high-dimensional covariate to control for confounding.

Manuscript II *Variable importance measures for heterogeneous treatment effects with survival outcome*

This manuscript considers the problem of detecting modifiers of the treatment effect. Hines, Dukes, et al. (2022) develop a variable importance measure for ranking the association of different covariate groups and the treatment effect. The measure and the subsequent estimation procedure are developed for a continuous outcome in fully observed data, and we extend their methods to survival outcomes with censored data. We derive an estimator based on the EIF of the corresponding variable importance measure and we provide high-level assumptions needed for valid asymptotic inference.

Additionally, we derive a new measure of variable importance based on the assumption-lean inference approach (Vansteelandt and Dukes, 2022) in the survival setting, and we remark on its extension to other data settings based on a general structure of the EIF. The new measure is given in terms of a best partially linear projection of the CATE function. We derive an estimator based on the EIF and give high-level assumptions required for valid asymptotic inference.

The assumptions required for inference for both of the derived estimators are seen to be fairly mild, and the estimators allow for the use of machine-learning approaches to estimate the involved nuisance parameters.

The finite sample performance of the estimators of the two variable importance measures was investigated in a simulation study. Estimation of the measure derived in Hines, Dukes, et al. (2022) requires a large number of observations for valid inference, even in the case of correctly specified parametric nuisance parameter estimators. A similar result was also found in Levy et al. (2021), who considered a parameter closely related to the one in Hines, Dukes, et al. (2022). In contrast, estimation of the best partially linear projection parameter required much fewer observations for valid inference, also when considering flexible nuisance estimation via random forests.

Manuscript III *Causal effect on the number of life years lost due to a specific event*

This manuscript derives a measure of treatment effect based on the number of life years lost due to a specific event (Andersen, 2013). The measure of treatment effect serves as an alternative to the absolute risk definition (Chapter 3.1) in time-to-event settings with competing risks and censored data. The definition can be seen as a competing risks analogue for the restricted mean survival time (RMST) in a survival setting. As in the survival setting, the new measure defines treatment effect directly on the time scale of the study and it provides results that are easier to communicate than absolute risks. To the best of our knowledge, this is the first extension of an RMST analogue for competing risk to causal effects.

We derive an estimator for the ATE based on the number of life years lost due to a specific event, and give high-level assumptions on the nuisance parameter estimators required for valid inference. Furthermore, we extended the best partially linear projection parameter from Manuscript II. We derived an estimator an gave high-level assumptions required for inference. In both settings, the estimators were derived as cross-fitted one-step estimators based on the EIF corresponding to each target parameter, respectively.

The finite sample performance of the derived estimators was investigated in a simulation study. There was not much difference between the cross-fitted and un-cross-fitted one-step estimators when parametric nuisance parameter estimators were used, but when random forests were used for nuisance parameter estimation, the one-step estimators was not able to provide valid inference, whereas the cross-fitted one-step estimator performed in line with the parametric nuisance estimators.

Lastly, the derived estimators were applied to a part of the study in Kessing et al. (2024).

Chapter 7

Perspectives and future research

In this chapter, we reflect on the work carried out in the Manuscripts and discuss its limitations. The limitations themselves open the door for future research projects, and we highlight some potentially interesting avenues for future work.

Inference with high-dimensional covariates and censored data. The simulation study in Manuscript I suggests that penalized regressions can be used for handling high-dimensional covariates in ATE estimation based on censored data. The sparsity settings analysed in the simulations were chosen based on results from Hou et al. (2021). They analysed a different, but related, target parameter and gave theoretical results for the target parameter estimator under a specific choice of penalisation parameter, and gave assumptions on the underlying sparsity of nuisance parameters required for valid inference. In contrast, we only rely on simulation studies for justifying the validity of our proposed estimator, which may not give a full picture of the performance across different setting. A limitation of our approach is that a parametric structure is assumed for essentially all nuisance parameters. We addressed this issue in a simulation study, where one of the nuisance parameter estimators was misspecified. The results related to bias and coverage were promising, but the underlying nuisance functions were in all cases assumed to be ultra sparse. Further investigation is needed to see how the estimator performs in different settings combining misspecification and moderate sparsity.

An interesting topic for future research is to derive theoretical properties of our derived estimator akin to Hou et al. (2021). The difficulty lies in structure of the remainder term associated with the one-step estimator for the ATE with censored data (both with and without competing risks), where the challenge is to bound a certain integral difference (see e.g. assumption B3 in Manuscript III). In ATE estimation without censoring, the Cauchy-Schwarz inequality is used to bound the remainder term, and with censored data, the same technique can be used when the hazard estimators are assumed to have density w.r.t. the Lebesgue measure (Rytgaard et al., 2023). Essentially, if we assume absolute continuous hazard estimators, the remainder term for the ATE given in Manuscript III Assumption B is bounded as

$$\begin{aligned}
&\left| \mathrm{E}\left\{ \sum_{i=1,2} \int_0^{t^*} S(s \mid a, X) \hat{H}_{ij}(s, t^* \mid a, X) \right. \right. \\
&\qquad \left. \left. \times \left(1 - \frac{\pi(a \mid X) S_C(s \mid a, X)}{\hat{\pi}(a \mid X) \hat{S}_C(s \mid a, X)}\right) \mathrm{d}\left[\hat{\Lambda}_i(s \mid a, X) - \Lambda_i(s \mid a, X)\right] \right\} \right| \\
&\leq \sum_{i=1,2} \mathrm{E} \left| \int_0^{t^*} S(s \mid a, X) \hat{H}_{ij}(s, t^* \mid a, X) \right. \\
&\qquad \left. \times \left(1 - \frac{\pi(a \mid X) S_C(s \mid a, X)}{\hat{\pi}(a \mid X) \hat{S}_C(s \mid a, X)}\right) \left[\hat{\lambda}_i(s \mid a, X) - \lambda_i(s \mid a, X)\right] \mathrm{d}s \right| \\
&\leq K \sum_{i=1,2} \left\| \hat{\pi} \hat{S}_C - \pi S_C \right\|_{L_2(\mu \otimes m)} \left\| \hat{\lambda}_i - \lambda_i \right\|_{L_2(\mu \otimes m)},
\end{aligned}$$

for some $K > 0$, where the expectation considers the estimated nuisance functions fixed. Here, μ is distribution of X and m denotes the Lebesgue measure. The inequality follows the Cauchy-Schwarz inequality together with some positivity assumptions on $\hat{H}$, S, $\hat{S}_C$ and $\hat{\pi}$. Hence, $n^{-1/2}$-convergence of the remainder term boils down to $n^{-1/4}$-convergence of each of the nuisance functions.

This approach fails when the cumulative hazard estimators are defined as jump processes (as is the case in our setting, where Breslow estimators based on penalised Cox regression are used). We note that this is a general challenge for EIF-based estimation of the ATE (and other functionals) with censored data, and it may be insurmountable for general cumulative hazard estimators. But, since the Breslow estimator based on Cox regression provides some structure of the cumulative hazards, it may be possible to leverage this structure in bounding the remainder term.

Power optimal tests in nonparametric models. Analysis of treatment effect variable importance based on the best partially linear projection parameter defined in Manuscript II and extended in Manuscript III relies on a p-value associated with the test $H : \Omega_j(P_0) = 0$ (i.e., the test of no importance for variable j). The estimator for the target parameter itself was shown to provide valid inference when using flexible machine-learning for nuisance parameter estimation, and the associated test was seen (in our simulation study) to achieve a Type I error rate in finite samples of approximately 5%. But, when using machine-learning for nuisance estimation, simulations also showed that the test statistic was underpowered compared to its counterpart using correctly specified parametric models. A limitation of our test statistic is that the standard error is estimated with a cross-fitted plug-in estimator, and when using machine-learning for nuisance estimation, the standard error inherits the possible slow convergence rates of the nuisance estimator.

The variance of the EIF can be defined as yet another target parameter $\psi(P_0) = P_0 \tilde{\psi}^2_{P_0}$, and one can calculate the associated EIF (if it exists) to produce an EIF-based estimator for the standard error of the $\hat{\Omega}_j^{CF}$. The estimator for the standard error would then follow the parametric rate of $\hat{\Omega}_j^{CF}$, which would possibly regain some of the lost power coming from flexible nuisance estimation. An interesting avenue for future research would be to develop power-optimal tests in nonparametric models based on EIF-estimation of the standard errors.

Another limitation of our variable importance measure Ω_j is that it is defined as the projection onto the space of functions $x \mapsto \beta x_j + w(x_{-j})$, regardless of whether x_j is binary or continuous. A more reasonable space could be the space of functions $x \mapsto m(\beta; x_j) + w(x_{-j})$ for some working model m indexed by a euclidean parameter $\beta \in \mathbb{R}^k$ for some k. The target parameter is still defined as β^* corresponding to the best partially linear projection of τ, but now allowing for e.g. splines in order to capture non-linearities in x_j. This approach is taken in Semenova and Chernozhukov (2021) for the best linear projection, and an interesting topic for future research is to extend this idea to the best partially linear projection.

Treatment effect modifiers of non-response to antidepressants. Manuscript III concluded with an application of the best partially linear projection parameter to a part of the data from Kessing et al. (2024). We only considered analysis of treatment effect modifiers for a single comparison of two antidepressants, but preferably, the analysis should be extended to other comparisons as well. A challenge arises from the variable importance being based on an associated test, and multiple testing becomes a concern with many comparisons resulting in many tests. One can employ standard methods for multiple testing, but a naive approach may result in overconservative correction for the p-value. With an estimate of the EIF for the target parameter associated with each covariate, the covariance matrix of the multi-dimensional target parameter estimator can be calculated. The dependence of the individual parameter estimates can then be used in a multiple testing correction, that is not overconservative (Hothorn et al., 2008).

A practical challenge for future research is to extend the software implementation from Manuscript II and III to a multi-dimensional target parameter, with the aim of providing multiple test corrections of the p-values.

Bibliography

Andersen, P. K., Borgan, O., Gill, R. D., & Keiding, N. (1993). *Statistical models based on counting processes*. Springer Science & Business Media.

Andersen, P. K. (2013). Decomposition of number of life years lost according to causes of death. *Statistics in medicine*, *32*(30), 5278–5285.

Bickel, P. J., Klaassen, C. A. J., Ritov, Y., & Wellner, J. A. (1993). *Efficient and adaptive estimation for semiparametric models* (Vol. 4). Springer.

Boileau, P., Leng, N., Hejazi, N. S., van der Laan, M., & Dudoit, S. (2023). A nonparametric framework for treatment effect modifier discovery in high dimensions. *arXiv preprint arXiv:2304.05323*.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters.

Cui, Y., Kosorok, M. R., Sverdrup, E., Wager, S., & Zhu, R. (2023). Estimating heterogeneous treatment effects with right-censored data via causal survival forests. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *85*(2), 179–211.

Cui, Y., Zhu, R., Zhou, M., & Kosorok, M. (2022). Consistency of survival tree and forest models: Splitting bias and correction. *Statistica Sinica*, *32*(3), 1245–1267.

Farrell, M. H., Liang, T., & Misra, S. (2021). Deep neural networks for estimation and inference. *Econometrica*, *89*(1), 181–213.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, *33*(1), 1.

Gill, R. D., & Johansen, S. (1990). A survey of product-integration with a view toward application in survival analysis. *The annals of statistics*, *18*(4), 1501–1555.

Gill, R. D., Van Der Laan, M. J., & Robins, J. M. (1997). Coarsening at random: Characterizations, conjectures, counter-examples. *Proceedings of the First Seattle Symposium in Biostatistics: Survival Analysis*, 255–294.

Hammer, S. M., Katzenstein, D. A., Hughes, M. D., Gundacker, H., Schooley, R. T., Haubrich, R. H., Henry, W. K., Lederman, M. M., Phair, J. P., Niu, M., et al. (1996). A trial comparing nucleoside monotherapy with combination therapy in hiv-infected adults with cd4 cell counts from 200 to 500 per cubic millimeter. *New England Journal of Medicine*, *335*(15), 1081–1090.

Hernán, M. A., & Robins, J. M. (2010). Causal inference.

Hernán, M. A., & Robins, J. M. (2016). Using big data to emulate a target trial when a randomized trial is not available. *American journal of epidemiology*, *183*(8), 758–764.

Hines, O., Diaz-Ordaz, K., & Vansteelandt, S. (2022). Variable importance measures for heterogeneous causal effects. *arXiv preprint arXiv:2204.06030*.

Hines, O., Dukes, O., Diaz-Ordaz, K., & Vansteelandt, S. (2022). Demystifying statistical learning based on efficient influence functions. *The American Statistician*, *76*(3), 292–304.

Hothorn, T., Bretz, F., & Westfall, P. (2008). Simultaneous inference in general parametric models. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, *50*(3), 346–363.

Hou, J., Bradic, J., & Xu, R. (2021). Treatment effect estimation under additive hazards models with high-dimensional confounding. *Journal of the American Statistical Association*, 1–16.

Hu, L., Ji, J., & Li, F. (2021). Estimating heterogeneous survival treatment effect in observational data using machine learning. *Statistics in medicine*, *40*(21), 4691–4713.

Huang, J., Sun, T., Ying, Z., Yu, Y., & Zhang, C.-H. (2013). Oracle inequalities for the lasso in the cox model. *Annals of statistics*, *41*(3), 1142.

Ichimura, H., & Newey, W. K. (2022). The influence function of semiparametric estimators. *Quantitative Economics*, *13*(1), 29–61.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., & Lauer, M. S. (2008). Random survival forests.

Ishwaran, H., Kogalur, U. B., & Kogalur, M. U. B. (2023). Package 'randomforestsrc'. *breast*, *6*(1), 854.

Kennedy, E. H. (2016). Semiparametric theory and empirical processes in causal inference. *Statistical causal inferences and their applications in public health research*, 141–167.

Kennedy, E. H. (2022). Semiparametric doubly robust targeted double machine learning: A review. *arXiv preprint arXiv:2203.06469*.

Kennedy, E. H. (2023). Towards optimal doubly robust estimation of heterogeneous causal effects. *Electronic Journal of Statistics*, *17*(2), 3008–3049.

Kennedy, E. H., Balakrishnan, S., & G'Sell, M. (2020). Sharp instruments for classifying compliers and generalizing causal effects.

Kennedy, E. H., Balakrishnan, S., Robins, J. M., & Wasserman, L. (2024). Minimax rates for heterogeneous causal effect estimation. *The Annals of Statistics*, *52*(2), 793–816.

Kessing, L. V., Ziersen, S. C., Andersen, F. M., Gerds, T., & Budtz-Jørgensen, E. (2024). Comparative responses to 17 different antidepressants in major depressive disorder: Results from a 2-year long-term nation-wide population-based study emulating a randomized trial. *Acta Psychiatrica Scandinavica.*

Levy, J., van der Laan, M., Hubbard, A., & Pirracchio, R. (2021). A fundamental measure of treatment effect heterogeneity. *Journal of Causal Inference*, *9*(1), 83–108.

Li, H., Hubbard, A., & van der Laan, M. J. (2023). Targeted learning on variable importance measure for heterogeneous treatment effect. *arXiv preprint arXiv:2309.13324.*

Martinussen, T., & Scheike, T. H. (2006). *Dynamic regression models for survival data* (Vol. 1). Springer.

Martinussen, T., & Stensrud, M. J. (2023). Estimation of separable direct and indirect effects in continuous time. *Biometrics*, *79*(1), 127–139.

Moore, K. L., & van der Laan, M. J. (2009). Application of time-to-event methods in the assessment of safety in clinical trials. In *Design and analysis of clinical trials with time-to-event endpoints* (pp. 473–500). Chapman; Hall/CRC.

Munch, A., Gerds, T. A., van der Laan, M. J., & Rytgaard, H. C. W. (2024). Estimating conditional hazard functions and densities with the highly-adaptive lasso. *arXiv preprint arXiv:2404.11083.*

Neyman, J. (1923). On the application of probability theory to agricultural experiments. essay on principles. *Ann. Agricultural Sciences*, 1–51.

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

Pearl, J. (2009). *Causality.* Cambridge university press.

Petersen, M. L., & van der Laan, M. J. (2014). Causal models and learning from data: Integrating causal modeling and statistical estimation. *Epidemiology*, *25*(3), 418–426.

Richardson, T. S., & Robins, J. M. (2013). Single world intervention graphs (swigs): A unification of the counterfactual and graphical approaches to causality. *Center for the Statistics and the Social Sciences, University of Washington Series. Working Paper*, *128*(30), 2013.

Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical modelling*, *7*(9-12), 1393–1512.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, *66*(5), 688.

Rytgaard, H. C. W., Eriksson, F., & van der Laan, M. (2023). Estimation of time-specific intervention effects on continuously distributed time-to-event outcomes by targeted maximum likelihood estimation. *Biometrics.*

Rytgaard, H. C. W., Gerds, T. A., & van der Laan, M. J. (2022). Continuous-time targeted minimum loss-based estimation of intervention-specific mean outcomes. *The Annals of Statistics*, *50*(5), 2469–2491.

Rytgaard, H. C. W., & van der Laan, M. J. (2022). Targeted maximum likelihood estimation for causal inference in survival and competing risks analysis. *Lifetime Data Analysis*, 1–30.

Semenova, V., & Chernozhukov, V. (2021). Debiased machine learning of conditional average treatment effects and other causal functions. *The Econometrics Journal*, *24*(2), 264–289.

Simon, N., Friedman, J., Hastie, T., & Tibshirani, R. (2011). Regularization paths for cox's proportional hazards model via coordinate descent. *Journal of statistical software*, *39*(5), 1.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *58*(1), 267–288.

Tsiatis, A. A. (2006). *Semiparametric theory and missing data* (Vol. 4). Springer.

van der Laan, M. J. (2006). Statistical inference for variable importance. *The International Journal of Biostatistics*, *2*(1).

van der Laan, M. J., & Robins, J. M. (2003). *Unified methods for censored longitudinal data and causality.* Springer.

van der Laan, M. J., & Rose, S. (2011). *Targeted learning.* Springer.

van der Vaart, A. (1991). On differentiable functionals. *The Annals of Statistics*, 178–204.

van der Vaart, A. (2000). *Asymptotic statistics* (Vol. 3). Cambridge university press.

Van Houwelingen, H. C., Bruinsma, T., Hart, A. A., Van't Veer, L. J., & Wessels, L. F. (2006). Cross-validated cox regression on microarray gene expression data. *Statistics in medicine*, *25*(18), 3201–3216.

Vansteelandt, S., & Dukes, O. (2022). Assumption-lean inference for generalised linear model parameters. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *84*(3), 657–685.

Verweij, P. J., & van Houwelingen, H. C. (1993). Cross-validation in survival analysis. *Statistics in medicine*, *12*(24), 2305–2314.

Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, *113*(523), 1228–1242.

Wei, W., Petersen, M., van der Laan, M. J., Zheng, Z., Wu, C., & Wang, J. (2023). Efficient targeted learning of heterogeneous treatment effects for multiple subgroups. *Biometrics*, *79*(3), 1934–1946.

Westling, T., Luedtke, A., Gilbert, P. B., & Carone, M. (2023). Inference for treatment-specific survival curves using machine learning. *Journal of the American Statistical Association*, (just-accepted), 1–26.

Wu, Y. (2012). Elastic net for cox's proportional hazards model with a solution path algorithm. *Statistica Sinica*, *22*, 27.

Xu, Y., Ignatiadis, N., Sverdrup, E., Fleming, S., Wager, S., & Shah, N. (2023). Treatment heterogeneity with survival outcomes. In *Handbook of matching and weighting adjustments for causal inference* (pp. 445–482). Chapman; Hall/CRC.

Ziersen, S. C., & Martinussen, T. (2024a). Causal effect on the number of life years lost due to a specific event: Average treatment effect and variable importance. *unpublished manuscript.*

Ziersen, S. C., & Martinussen, T. (2024b). Variable importance measure for heterogeneous treatment effect with survival outcome. *unpublished manuscript.*

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, *67*(2), 301–320.

Manuscripts

Manuscript I

On estimation of the average treatment effect with register data: competing risk and high-dimensional covariates - a case study

Simon Christoffer Ziersen , Esben Budtz-Jørgensen & Thomas Alexander Gerds

Details: In preparation

On estimation of the average treatment effect with registry data: competing risk and high-dimensional covariates - a case study

Simon Christoffer Ziersen , Esben Budtz-Jørgensen & Thomas Alexander Gerds

Section of Biostatistics, University of Copenhagen

July 2024

Abstract

We consider estimation of the average treatment effect, defined as the mean difference in the τ-year risk of an event of interest, using registry data. In observational studies, all potential confounders must be included for identifying the average treatment effect, and for some register studies, this amounts to including a high-dimensional covariate in certain outcome regressions. We construct an estimator for the average treatment effect that allows for penalized regression of certain nuisance functions, alleviating the high-dimensional covariate problem, while still providing valid inference. The estimator is based on semiparametric efficiency theory and its statistical properties are investigated in simulation studies. The estimator is used to compare the response to different antidepressants in a study based on data from the Danish national registers.

1 Introduction

The average treatment effect (ATE) plays an important role in the causal inference literature, and it is the primary target of estimation in many applications, including both randomized trials and observational studies. When data are observational, the ATE can be identified according to the G-formula (Robins, 1986) under some structural assumptions on the treatment mechanism and the outcome, one of which is the assumption of no unmeasured confounding. The assumption asserts that the observed treatment assignment is independent of the outcome, conditional on a set of covariates (termed confounders), which in applications amounts to including the set of potential confounders in an outcome regression from which predictions are used to produce an estimate of the ATE. In recent years, many methodological advances have been made in double robust estimation of the ATE for different data structures (e.g., van der Laan and Robins, 2003, van der Laan and Rose, 2011, Chernozhukov et al., 2018, Ozenne et al., 2020, Rytgaard et al., 2023), where predictions from the outcome regression along with predictions from a model of the treatment propensity are combined in an estimate of the ATE. The double robust estimators are consistent when either the model for the treatment propensity or the model for the outcome regression is correctly specified. This is convenient in observational studies, where the underlying data generating mechanism is rarely known. Furthermore, the estimators facilitate the use of data-adaptive estimation of the nuisance parameters (outcome regression and treatment propensity) and when the set of potential confounders is large relative to the number of observations, this calls for machine-learning methods that are designed for high-dimensional data. An example of such a method is the lasso (Tibshirani, 1996) which is consistent under some assumptions on the sparsity of the outcome regression model and the treatment propensity model, respectively (Chernozhukov et al., 2018).

We review some of the challenges involved in estimating the ATE based on registry data. Here, patients are followed over time and the outcome of interest is the time to a given event, such as a hospital admission which defines the onset of a disease. Typically, one only observes a censored version of the underlying event time, as patients can leave the risk set without an event due to emigration or the end of follow-up. Furthermore, when the event of interest is not all-cause death, one has to account for competing risks in the estimation procedure. Ozenne et al. (2020) develops a double robust estimator for the ATE, defined as the mean risk difference of the event of interest in a given time horizon, in the presence of censoring and competing risks. Their estimator relies on working Cox regression models for the estimation of the cause-specific hazard functions and a working logistic regression model for the treatment propensity. Rytgaard et al. (2023) develops an estimator for the ATE defined as in Ozenne et al. (2020) but with nonparametric estimation of the nuisance parameters. Specifically, they show that when using a highly adaptive lasso (Munch et al., 2024) for estimation of the cause-specific hazard functions and treatment propensity function, the obtained ATE estimator is asymptotically locally efficient. Westling et al. (2023) considers estimation of the ATE in the survival setting (i.e., without competing risk) for general machine-learning methods used for nuisance parameter estimation, and they give high-level assumptions on the estimators needed for asymptotic efficiency of the ATE estimator. Hou et al. (2021) consider estimation of the treatment effect in a high-dimensional setting (without competing risk), where the number of covariates is assumed to be large relative to the number of observations. They define the treatment effect as a parameter in the additive hazard model (Martinussen and Scheike, 2006) and they give sparsity assumptions analogous to Chernozhukov et al. (2018) for asymptotically efficient estimation of this parameter.

We consider data from the Danish national registers based on the study described in Kessing et al. (2023). The aim is to compare the response to different antidepressants treatments available on the Danish market in the period 1995-2018. Patients are enrolled in the study after their first diagnosis with depression at a psychiatric hospital, and a baseline treatment is defined by the first purchase of an antidepressant after discharge from the hospital. The outcome of interest is time to non-response, defined by a switch of treatment or readmission to a psychiatric hospital. The competing risk is comprised of a diagnosis with bipolar disorder, organic mental disorder, schizophrenia or death. In Kessing et al. (2023), the authors define and estimate the ATE in accordance with Ozenne et al. (2020) using working models for the cause-specific hazard functions and the propensity of treatment. In order to control for confounding, disease histories in form of hospital admissions with a given ICD-10 code were included in the Cox regression models. The ICD-10 codes were grouped into nine disease categories corresponding to disease chapters defined for the ICD-10 classification. The grouping of the diseases into 9 groups is coarse, because the ungrouped version contains ~ 1000 different disease-codes which all could be included in the estimation as potential confounders. For the sole purpose of illustrating our methods, we include the ~ 1000 disease codes as covariates into the learning of nuisance parameter models and hence employ methods that allow for high-dimensional covariates.

In this article, we implement a cross-fitted one-step estimator for the ATE (Chernozhukov et al., 2018, Kennedy, 2022), defined as in Ozenne et al. (2020), based on the efficient influence function corresponding to the ATE given in Rytgaard et al. (2023). In order to handle the high-dimensional covariate setting, we employ penalized Cox regression for the nuisance parameters related to non-response, competing risks, and censoring, respectively. The double robustness properties of the estimator are tested in a simulation study, where they are related to the sparsity of each of the nuisance parameters. The estimator is used to re-analyze the comparisons of antidepressants in Kessing et al. (2023), taking the inherent high-dimensionality of the covariate into account.

2 Register study comparing the response to different antidepressants

The methods considered in this paper were inspired by the study presented in Kessing et al. (2023). It is based on data from the Danish national registers, where patient information on diagnoses given at a Danish hospital is available together with information on drug purchases at the Danish pharmacies. Furthermore, the data includes information on age and sex. The aim of the study is to compare the response to treatment with different antidepressants in patients with a first diagnosis of major depressive disorder. The study includes all patients with a first diagnosis of major depressive disorder at a psychiatric hospital between 1995-2018, and their baseline treatment is defined as the first purchase with an antidepressant after discharge of the hospital, which also defines the index date of the study for a given patient. An event of interest, termed non-response, is defined as a switch in treatment (determined from a purchase with a different antidepressant, antipsychotic or lithium) or readmission to a psychiatric hospital. A competing event is defined as admission to a psychiatric hospital with a diagnosis of bipolar disorder, schizophrenia or organic mental disorder, or death. Patients are followed until an event of non-response, competing event or censoring, defined at emigration or end of follow-up in 2018. The antidepressants considered in the comparison are grouped according to their pharmacological profile, and within each group, a reference is chosen as the most widely prescribed antidepressant, and comparisons are made between each of the other antidepressants within the group and the reference drug.

The outcome of interest is defined as time to non-response and the target parameter for comparison of two drugs is defined as the mean absolute risk difference of non-response at two years after index date. To estimate the target parameter, Kessing et al. (2023) fits cause-specific Cox regressions for the time to non-response and competing event, respectively, which are combined into an estimator for the conditional cumulative function. Predictions from the cumulative incidence function are then used to obtain an estimate of the target parameter based on the G-formula, as described in Ozenne et al. (2020), equation (3). To control for confounding, the Cox regressions are adjusted for age, sex, psychiatric and somatic disease history. The somatic disease history is defined as a diagnosis within one of nine disease groups based on the ICD-8 and ICD-10 classification. The collapsing of somatic diseases into nine groups is essentially made in order to achieve parsimonious models, as the number of individual somatic diseases observed in the data is 1170, which is of the order of the number of observations for some of the treatment groups. As such, the cause-specific Cox regression are unfit for including disease histories on individual disease level.

In the next sections we will detail the construction of an estimator of the target parameter, which allows for penalized regression estimation of the nuisance parameters, enabling the inclusion of the somatic histories on individual disease level, while still providing valid inference. The results regarding non-response to treatment obtained by re-analyzing the data of Kessing et al., 2023 with the here derived estimator are presented in Section 6.

3 Setup and notation

Let T and C be the time to event and censoring respectively, and let $\Delta \in \{1, 2\}$ be the event indicator. The observed time to event is $\tilde{T} = T \wedge C$ and the observed event indicator is $\tilde{\Delta} = \mathbb{1}\{T \geq C\}\Delta$. Let $A \in \{0, 1\}$ denote the baseline treatment and let $W = (W_1, ..., W_d) \in \mathbb{R}^d$ denote baseline covariates. The observed data are $\mathcal{O} = (\tilde{T}_i, \tilde{\Delta}_i, A_i, W_i)_{i=1,..,n}$ where $O_1, ..., O_n$ are assumed to be i.i.d. with distribution P_0 which belongs to a suitable nonparametric family of probability distributions $\mathcal{M}$.
Let $N_j(t) = \mathbb{1}\{\tilde{T} \leq t, \tilde{\Delta} = j\}$ be the observed counting process for the j'th event and let $\lambda_j(t|a, w)$, $\lambda_c(t|a, w)$ denote the conditional hazard functions for the j'th event and censoring distribution, respectively, and $\Lambda_j(t|a, w)$, $\Lambda_c(t|a, w)$ are their associated cumulative hazard functions. Furthermore, we denote with $S(t|a, w) = \exp(-\Lambda_1(t|a, w) - \Lambda_2(t|a, w))$ the conditional event-free survival function, with $F_1(t \mid a, w) = \int_0^t S(u \mid a, w)\Lambda_1(\mathrm{d}u \mid a, w)$ the cumulative incidence function of the event of interest, with $S_c(t|a, w) = \exp(-\Lambda_c(t|a, w))$ the survival function of the censoring distribution, with $\pi(a|w) = P(A = a|W = w)$ the propensity score, and with μ the marginal distribution of W.
To define our causal parameter, we introduce the variable $Y_j(\tau) = \mathbb{1}\{T \leq \tau, \Delta = j\}$ and define $Y_j^a(\tau)$ as the potential outcome, that is, the outcome of a person if they, possibly contrary to the fact, had received treatment a.

The target parameter, defined as the average treatment effect, is given by

$$\psi_\tau = \mathrm{E}\, Y_1^1(\tau) - \mathrm{E}\, Y_1^0(\tau)$$

for a given time-horizon $[0, \tau]$. To identify the target parameter from the observed data, we make the following assumptions

(i) (consistency) $Y_1(t) = AY_1^1(t) + (1 - A)Y_1^0(t)$ conditional on A.

(ii) (no unmeasured confounding) $Y_1^a(t) \perp\!\!\!\perp A \mid W, \quad a = 0, 1.$

(iii) (positivity) $\pi(a \mid w)S_c(s \mid a, w)S(s \mid a, w) > \eta > 0, \quad \forall(s, a, w).$

(iv) (independent censoring) $T \perp\!\!\!\perp C \mid A, W.$

Under assumptions (i)-(iv), the target parameter is identifiable from the observed data (Rytgaard et al., 2023) and the target parameter can be expressed as a mapping from $\mathcal{M}$ to the reals as

$$\psi_\tau(P) = \mathrm{E}(F_1(\tau|A = 1, W)) - \mathrm{E}(F_1(\tau|A = 0, W)).$$

4 Estimation of the target parameter

There exist several articles concerning estimation of the target parameter; Ozenne et al., 2020 considers a double robust estimator with working Cox models and logistic regression for nuisance estimation, and Rytgaard et al., 2023 considers a targeted-minimum-likelihood (TMLE) approach with nuisance parameters estimator based on the highly adaptive lasso. We will adopt an influence function based approach similar to the TMLE, but with more restrictive estimators of the nuisance parameters based on penalized Cox regression and logistic regression, mimicking the approach of Ozenne et al., 2020, but with a flexible covariate selection. Both our estimating equation approach and the TMLE rely on semiparametric efficiency theory (Bickel et al., 1993, Van der Vaart, 2000 ch. 25, van der Laan and Rose, 2011) which revolves around the so-called efficient influence function (EIF) of the target parameter. We will give a brief introduction to the general idea, and refer to Kennedy (2022) and Hines et al. (2022) for nice reviews of EIF-based estimation of functional parameters.

For a general parameter ψ, an estimator $\hat{\psi}$ is said to asymptotically linear if admits the expansion $\hat{\psi} - \psi = \mathbb{P}_n \mathbb{IF} + o_p(n^{-1/2})$, where $\mathbb{P}_n$ denotes the empirical measure over $\mathcal{O}$ with $\mathbb{P}_n f = \frac{1}{n}\sum_{i=1}^n f(O_i)$. The function $\mathbb{IF}$ is the influence function corresponding to the estimator $\hat{\psi}$ and it characterises the asymptotic distribution of the estimator since $\sqrt{n}(\hat{\psi} - \psi) \xrightarrow{D} \mathcal{N}(0, \mathrm{E}(\mathbb{IF}(O)^2))$. If the target parameter is smooth as a map on $\mathcal{M}$ (see Van Der Vaart, 1991 for a rigorous definition), there exist a unique function $\tilde{\psi}$ called the efficient influence function, which determines the lower variance bound of any regular estimator of ψ. The EIF can be calculated without reference to any specific estimator, and once it is known, different techniques exist for constructing an estimator of ψ that is asymptotically linear with the EIF as its influence function. Here, we focus on the so-called one-step estimator.

Define $\psi_\tau^a(P) = \mathrm{E}(F_1(\tau|A = a, W))$ for $a = 0, 1$, such that $\psi_\tau(P) = \psi_\tau^1(P) - \psi_\tau^0(P)$. The EIF for our target parameter $\psi_\tau(P)$ is known from the literature, see e.g. Rytgaard and van der Laan, 2022, and is given by

$$\tilde{\psi}_\tau = \tilde{\psi}_\tau^1 - \tilde{\psi}_\tau^0$$

where

$$\tilde{\psi}_\tau^a(P)(O) = \varphi_\tau^a(\nu)(O) - \psi_\tau^a(P)$$

with

$$\varphi_\tau^a(\nu)(O) = \frac{\mathbb{1}\{A=a\}}{\pi(a|w)}\left(\int_0^\tau \frac{S(t-|a,w) - F_1(\tau|a,w) + F_1(t|a,w)}{S(t-|a,w)G(t-|a,w)} dM_1(t|a,w) - \int_0^\tau \frac{F_1(\tau|a,w) - F_1(t|a,w)}{S(t-|a,w)G(t-|a,w)} dM_2(t|a,w)\right) + F_1(\tau|a,w) \tag{1}$$

where $\nu = (\Lambda_1, \Lambda_2, \Lambda_c, \pi)$ is the nuisance parameter. We denote $\varphi_\tau = \varphi_\tau^1 - \varphi_\tau^0$. Here, M_1 and M_2 correspond to the martingales associated with the counting processes for the event of interest and the competing event, respectively, and are given by $M_j(t|a,w) = N_j(t) - \Lambda_j(t|a,w)\mathbb{1}\{\tilde{T} \geq t\}$, $j = 1, 2$. In the following we will use the EIF to construct an estimator for the target parameter. The estimation will be carried out for ψ_τ^a for $a = 0, 1$ separately to obtain $\hat{\psi}_\tau = \hat{\psi}_\tau^1 - \hat{\psi}_\tau^0$.

4.1 One step estimator

The one-step estimator for $\psi_\tau^a(P)$ is found by adding the first order bias from a Von-Mises expansion to a plugin estimator, where the bias is given by the empirical measure of the EIF (Van der Vaart, 2000). When the EIF is linear in the target parameter this corresponds to the estimating equation based estimator found by solving $\mathbb{P}_n\tilde{\psi}_\tau^a(O) = 0$ in ψ_τ^a. Our estimator is thus given by

$$\begin{aligned}\hat{\psi}_\tau^a &= \frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{1}\{A_i = a\}}{\hat{\pi}(a|w_i)}\Bigg(\int_0^\tau \frac{\hat{S}(t-|a,w_i) - \hat{F}_1(\tau|a,w_i) + \hat{F}_1(t|a,w_i)}{\hat{S}(t-|a,w_i)\hat{G}(t-|a,w_i)}d\hat{M}_1(t|a,w_i) \\ &- \int_0^\tau \frac{\hat{F}_1(\tau|a,w_i) - \hat{F}_1(t|a,w_i)}{\hat{S}(t-|a,w_i)\hat{G}(t-|a,w_i)}d\hat{M}_2(t|a,w_i)\Bigg) + \hat{F}_1(\tau|a,w_i) \\ &= \mathbb{P}_n\varphi_\tau^a(\hat{\nu}) \end{aligned} \tag{2}$$

where $\hat{\nu}$ is the estimated nuisance parameter. The estimator of the ATE is then obtained from $\hat{\psi}_\tau = \hat{\psi}_\tau^1 - \hat{\psi}_\tau^0$.

In order to show that the one-step estimator is asymptotically linear with the EIF as its influence function, one typically relies on an expansion of the estimator as follows (Kennedy, 2022)

$$\hat{\psi}_\tau - \psi_\tau = \mathbb{P}_n\tilde{\psi}_\tau + \underbrace{(\mathbb{P}_n - P)[\varphi_\tau(\hat{\nu}) - \varphi_\tau(\nu)]}_{\text{empirical process term}} + \underbrace{P\varphi_\tau(\hat{\nu}) - \psi_\tau(P)}_{\text{remainder term}}$$

where $Pf = \int f\,\mathrm{d}P$. Here, the empirical process term and the remainder term are required to be $o_p(n^{-1/2})$.

For the remainder term, $o_p(n^{-1/2})$-convergence depends on the convergence rates of the nuisance parameter estimators. In a related study on estimation of the treatment effect defined as a regression parameter in an additive hazards model with high-dimensional covariates, Hou et al. (2021) shows that the condition on the convergence rates of the nuisance estimator obtained from penalized additive hazards model and logistic regression amounts to a condition on the sparsity of the underlying nuisance functions, i.e., how many non-zero coefficients appear in the true regression models. Similar results have not been shown for the estimator considered here, and in Section 5, we investigate the finite-sample distribution of the target parameter estimator in relation to the underlying sparsity of the nuisance parameters.

The $o_p(n^{-1/2})$-convergence of the empirical process term is achieved if the nuisance estimators belong to a Donsker class (Rytgaard et al., 2023), which restricts us to nuisance parameter estimators that are not too flexible. As advocated by several authors, the Donsker class condition may fail to hold for high-dimensional models and instead suggest to use a special kind of sample splitting, termed cross-fitting, to ensure $o_p(n^{-1/2})$-convergence of the empirical process term (see e.g. Chernozhukov et al., 2018). In the next section, we will describe a cross-fitted version of the proposed estimator.

Along with the estimator $\hat{\psi}_\tau$, we define an estimator for the variance $\mathrm{E}\,\tilde{\psi}_\tau(O)^2$ by

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(\varphi_\tau(\hat{\nu})(O_i) - \hat{\psi}_\tau)^2.$$

Based on this estimator we can estimate the standard error of $\hat{\psi}_\tau$ by $\sqrt{\hat{\sigma}^2/n}$.

4.2 Cross-fitting

To alleviate the Donsker conditions needed for $o_p(n^{-1/2})-$convergence of the empirical process term, we consider a general form of sample splitting in estimating our target parameter. The cross-fitting procedure works by estimating $\hat{\nu}$ and $\mathbb{P}_n$ on independent subsamples of the data from which one obtains $\sqrt{n}$-convergence under much milder conditions compared to Donsker type assumptions. To combat the loss in efficiency coming from the data splitting, the roles of the subsamples are reversed to obtain another estimate of the target parameter and the two estimates are then averaged in a final target parameter estimate. The procedure can then be extended to finer partitions of the data, and it is commonly referred to as K-fold cross-fitting, when done over K subsamples. It is important that the number of subsamples, K, is assumed fixed and not depending on the number of observations. For details and discussions of cross-fitting see e.g. Chernozhukov et al. (2018), Kennedy (2022).

In cross-fitting we split the data into K disjoint folds $\mathcal{V}_k$, $k = 1, \dots, K$, and let N_k be the number of observations in fold k. Let $\mathbb{P}_n^k$ be the empirical measure on the k'th split and $\hat{\nu}_{-k}$ be the nuisance parameter estimates based on observations not in the k'th split. For each split, k, define the estimator $\hat{\psi}^a_{\tau,k} = \mathbb{P}_n^k \varphi^a_\tau(\hat{\nu}_{-k})$. The cross-fitted estimator is then given by

$$\hat{\psi}^a_{\tau,CF} = \sum_{k=1}^K \frac{N_k}{n} \hat{\psi}^a_{\tau,k} = \sum_{k=1}^K \frac{N_k}{n} \mathbb{P}_n^k \varphi^a_\tau(\hat{\nu}_{-k}) = \frac{1}{n} \sum_{k=1}^K \sum_{i \in \mathcal{V}_k} \varphi^a_\tau(\hat{\nu}_{-k})(O_i). \tag{3}$$

Similar to the one-step estimator, we define the cross-fitted variance estimator by

$$\hat{\sigma}^2_{CF} = \sum_{k=1}^K \frac{N_k}{n} \mathbb{P}_n^k [\varphi_\tau(\hat{\nu}_{-k}) - \hat{\psi}_{\tau,CF}]^2$$

which gives the standard error of the estimator $\hat{\psi}_{\tau,CF}$, when it is asymptotically linear with the EIF as its influence function.

4.3 Penalized nuisance parameters

The one-step estimators defined in equations (2) and (3) both rely on an initial estimate of the nuisance parameters $\nu = (\Lambda_1, \Lambda_2, \Lambda_c, \pi)$. To handle potentially high-dimensional covariates, we adopt the elastic net procedure (Zou and Hastie, 2005), where ridge and lasso regression are special cases. The procedure is developed for both generalized linear models and Cox regression and implemented collectively in the R-package `glmnet` which we will employ (see e.g. Simon et al., 2011 and Friedman et al., 2010 for software implementation and overview). In the case of $(\Lambda_1, \Lambda_2, \Lambda_c)$ we can use the obtained hazard ratios together with a Breslow estimator. For clarity, we describe the procedure in the two settings; hazard functions and logistic regression, starting with hazard function estimation.

We need an estimate for each of the cumulative hazard functions $(\Lambda_1, \Lambda_2, \Lambda_c)$. This amounts to estimating a cause-specific penalized Cox-regression model for each hazard function, including the penalized partial likelihood estimators of the log-hazard ratios and the corresponding Breslow estimators for the baseline hazard functions, in order to obtain an estimate of $(\Lambda_1, \Lambda_2, \Lambda_c)$. For ease of notation we describe the procedure for a generic hazard function. The method is presented in Van Houwelingen et al. (2006) an we will give a brief description.

Consider the task of estimating the log-hazard ratio β in a generic hazard model

$$\lambda(t|A_i, W_i) = \lambda_0(t)\exp(Z_i^T\beta),$$

where $Z_i = (A_i, W_i)^T$. Let $pl(\beta)$ be the log partial likelihood and consider the penalized log partial likelihood

$$l_{pen}^{\alpha}(\lambda, \beta) = pl(\beta) + \lambda P_\alpha(|\beta|)$$

where

$$\lambda P_\alpha(|\beta|) = \lambda\left(\alpha\sum_{i=1}^{p}|\beta_i| + (1-\alpha)\frac{1}{2}\sum_{i=1}^{p}\beta_i^2\right)$$

is known as the elastic net penalty, which is a mixture of ℓ_1 and ℓ_2 penalties. Here, we consider α to be a hyperparameter to be chosen a priori and it is considered fixed for the remainder of this section. Note that $\alpha = 0$ corresponds to ridge regression and $\alpha = 1$ corresponds to lasso. For a given λ, the β-estimate is obtained by maximizing l_{pen}^{α}:

$$\hat{\beta}^{\lambda} = \text{argmax}_\beta\; l_{pen}^{\alpha}(\lambda, \beta).$$

In order to select the optimal λ, we will use K-fold cross validation. That is, divide the data into K parts of roughly equal size. Let $\hat{\beta}_{-k}^{\lambda}$ be the estimate obtained by maximizing the penalized log-likelihood when leaving out the k'th part of the data for a given value of λ and define the cross-validated partial log-likelihood by

$$cvpl(\lambda) = \sum_{k=1}^{K} pl(\hat{\beta}_{-k}^{\lambda}) - pl_{(-k)}(\hat{\beta}_{-k}^{\lambda}),$$

where $pl_{(-k)}$ is the partial likelihood computed on the data when leaving out the k'th part. The optimal penalty parameter is then found by maximizing the cross-validated partial log-likelihood in λ

$$\lambda^{CV} = \text{argmax}_\lambda cvpl(\lambda)$$

from which we obtain and estimate for β as $\hat{\beta}^{\lambda^{CV}} = \text{argmax}_\beta\; l_{pen}^{\alpha}(\lambda^{CV}, \beta)$.

The $cvlp(\lambda)$ was proposed in Verweij and Van Houwelingen, 1993 to accommodate survival data in cross-validation. This method of validation is the one used by the R-package `glmnet`. The β-estimates are then plugged into a Breslow estimator to obtain an estimate of the cumulative hazard function

$$\hat{\Lambda}(t \mid z) = \hat{\Lambda}_0(t)\exp(z^T\hat{\beta}^{\lambda^{CV}})$$

where

$$\hat{\Lambda}_0(t) = \sum_{i=1}^{n}\int_0^t \frac{\mathrm{d}N_i(u)}{\sum_{j=1}^{n}\mathbb{1}(\tilde{T}_j > u)\exp(z_j^T\hat{\beta}^{\lambda^{CV}})}.$$

We estimate the propensity score by employing the elastic net penalization to a logistic regression for the treatment variable. The procedure is defined analogously to the hazard setting, where we consider the logistic model for the propensity

$$\pi(1 \mid z) = P(A = 1 \mid Z) = \text{expit}(Z^t\beta_\pi)$$

together with the penalized log-likelihood

$$l^{\alpha}_{pen,\pi}(\lambda, \beta_\pi) = \left[\sum_{i=1}^{n} A_i z_i^t \beta_\pi - \log(1 + e^{z_i^t \beta_\pi})\right] + \lambda P_\alpha(|\beta_\pi|).$$

5 Simulation study

We consider simulation studies of estimators of the target parameter

$$\psi_\tau(P) = \mathrm{E}(F_1(\tau \mid A = 1, W)) - \mathrm{E}(F_1(\tau \mid A = 0, W))$$

for a specific τ (in the following, we set $\tau = 1$). This requires simulation of our observed data which is comprised of $\mathcal{O}_i = (\tilde{T}_i, \tilde{\Delta}_i, A_i, W_i)$, for subject $i = 1, ..., n$, where $W_i = (W_{i1}, ..., W_{id})$. The distribution of $(\tilde{T}_i, \tilde{\Delta}_i)$ is determined by λ_1, λ_2 and λ_c, and the distribution of A_i by π. In the following we let $d = dim(W_i)$ be the dimension of the covariates and s_k, $k = \lambda_1, \lambda_2, \lambda_c, \pi$, be the sparsity level for a given nuisance parameter. That is, if, for example, $\pi(a|W) = \text{expit}(\gamma^T W)$ then $s_\pi = \sum_{i=1}^{d} 1\{\gamma_i \neq 0\}$. In all simulation studies we consider estimators $\hat{\psi}_\tau$ and $\hat{\psi}_{\tau,CF}$ with nuisance parameter estimators defined according to the penalized estimators in Section 4.3, including all covariates for estimation. The penalty functions were chosen corresponding to unpenalized, lasso, ridge and elastic net, where elastic net is chosen as $\alpha = 0.5$. For the three latter penalty functions, the penalty term is chosen by 10-fold cross validation and the estimator $\hat{\psi}_{\tau,CF}$ is based on 10-fold cross-fitting. For comparison, we also include an oracle estimator based on correctly specified Cox regressions for the hazard functions and logistic regression for the treatment. Finally, we generate covariates from a normal distribution $Z \sim \mathcal{N}(0, \Sigma)$, where $\Sigma_{ij} = 0.5^{|i-j|}$ with half of them being dichotomized, such that $W_{ij} = Z_{ij}$ if j is odd and $W_{ij} = 1(Z_{ij} > 0.3)$ if j is even. For each simulation study, the different penalization schemes are compared based on the sampling distribution of the target parameter estimator and coverage of the corresponding confidence intervals.

5.1 Increasing dimension

We consider a case with a sample size of $n = 500$ observations and varying covariate dimension $d = 10, 30, 50, \ldots, 400$ and relative sparse nuisance parameters with distributions given below. The aim here is to see how regularization of the nuisance estimators compares to their unregularized counterparts in low dimensional settings and how it stabilizes the estimation of the target parameter when the dimension is increased. For each value of d, we simulated $N = 1000$ data sets from the following models:

- $P(A = 1|W) = \text{expit}(W_1 + 0.3W_2 - 0.4W_3 - 0.5W_4)$
- $\lambda_c(t|A, W) = \lambda_{0,c}(t) \exp(0.5W_1 + 0.3W_2)$
- $\lambda_1(t|A, W) = \lambda_{0,1}(t) \exp(0.4W_1 + 0.5W_2 + 0.1W_3 - 0.1W_4 + 0.5A)$
- $\lambda_2(t|A, W) = \lambda_{0,2}(t) \exp(0.2W_1 + 0.8W_2 + 0.2W_3 + 0.5A)$

where $\lambda_{0,k}$ $(k \in \{c, 1, 2\})$ follows a Weibull distribution. The true value of the target parameter is $\psi_1 = 0.1287$. The biases of the estimators are estimated by $\frac{1}{N}\sum_{i=1}^{N} \hat{\psi}_1 - \psi_1$, where

$N = 1000$ is the number of simulations (here, $\hat{\psi}_1$ is meant as a generic estimator), and they are reported in Figure 1. For the lower dimension settings, the estimators based on lasso and elastic net provide similar results in terms of bias compared to the oracle, whereas the estimator based on ridge penalization is only similar in the lowest dimension settings. The lasso and elastic net estimators are seen to be stable across increasing dimensions, with their cross-fitted versions being closest to the oracle and with the cross-fitted lasso providing the best results.

The coverage of the different estimators is displayed in figure 2. The cross-fitted estimators are seen to provide better coverage than the un-cross-fitted versions, where elastic net and lasso seem to provide stable coverage across increasing dimension.

Figure 1 and 2 together suggest that the estimator based on lasso penalization and cross-fitting performs in line with the oracle across varying covariate dimensions, when the underlying true regression functions are sparse.

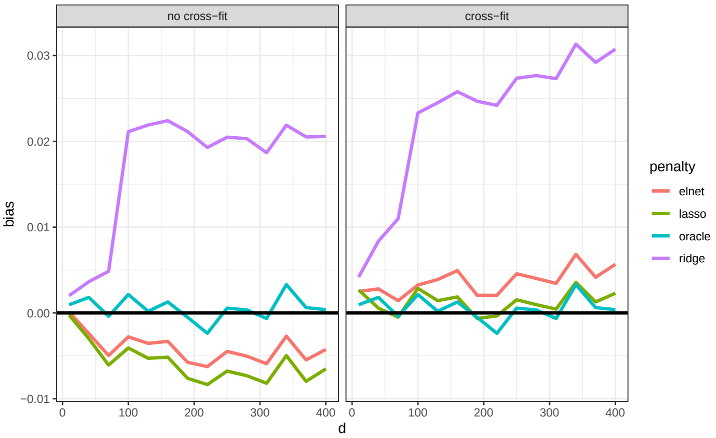


Figure 1: Bias of the estimators under varying covariate dimension d based on 1000 simulations.

5.2 Sparsity double robustness

We consider a high dimensional case with $n = 400$ and $d = 600$ and different combinations of sparsity levels for the nuisance parameters. The aim is to investigate if the double robustness related to the sparsity of nuisance parameters as suggested in Hou et al. (2021) relates to the double robustness of our target parameter estimators as described in Rytgaard et al., 2023. In our setting, the double robustness is given in terms of convergence rates for the pairs π *and* λ_c, and, λ_1 *and* λ_2, and we thus consider the following scenarios:

(i) *ultra-sparse nuisance parameters*: $s_\pi = s_{\lambda_c} = s_{\lambda_1} = s_{\lambda_2} = 3$.

(ii) *ultra-sparse intervention parameters*: $s_\pi = s_{\lambda_c} = 3$ and $s_{\lambda_1} = s_{\lambda_2} = 30$

(iii) *ultra-sparse event parameters*: $s_\pi = s_{\lambda_c} = 30$ and $s_{\lambda_1} = s_{\lambda_2} = 3$

(iv) *no ultra-sparseness*:$s_\pi = s_{\lambda_c} = s_{\lambda_1} = s_{\lambda_2} = 30$.

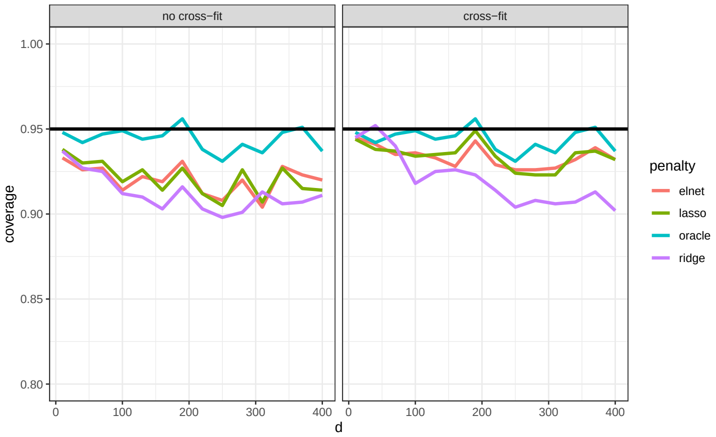


Figure 2: Coverage probability of confidence intervals of different estimators under varying covariate dimension d based on 1000 simulations.

Given the product structure of the remainder term (Rytgaard et al., 2023) and the sparsity conditions in Hou et al. (2021), we would expect that the desired convergence rate is obtained if either π *and* λ_c or λ_1 *and* λ_2 are sparse and the converse is of moderate sparsity. Thus, we expect good performance for scenario (i)-(iii) but not for (iv). The true values of the target parameter under the four scenarios are approximated as 0.180, 0.168, 0.180, 0.177, respectively, in the order they appear above. To ensure confounding in each of the scenarios, we let the propensity score along with the event hazards depend on the same three covariates, and additionally let the event hazards depend on the treatment. The sample distribution of the different estimators are presented in figure 3. In general, elastic net and lasso penalization provide similar results, with their cross-fitted versions performing better in terms of bias compared to their un-cross-fitted versions, except when none of the nuisance parameters are ultra sparse. As in the increasing dimension setting, ridge penalization provides biased estimates of the target parameter with the cross-fitted version being more biased than the un-cross-fitted. From figure 4 we see an overall large improvement in coverage in the cross-fitted versions of the elastic net and lasso compared to their un-cross-fitted versions. The results indicate that elastic net and lasso estimators with cross-fitting achieve asymptotic linearity when either the intervention or event nuisance parameters are sparse.

5.3 Misspecification

We consider a high dimensional setting with $n = 450$ and $d = 600$ and three different scenarios for the nuisance parameters. Each scenario corresponds to misspecification in one of the nuisance parameters, i.e, one of the hazard functions not following a Cox model or the propensity not following a logistic regression. As in the varying sparsity case, the aim is to investigate the double robustness of the target parameter estimator, but now considering that, on top of dealing with a high-dimensional covariate, one might misspecify the functional form of the nuisance parameters. Indeed, the considered nuisance parameter models only account for sparsity in the linear case (on log-hazard and log-odds scale, respectively) and do not account for non-linearities or interactions. The three scenarios are given by alterations to the setup from the varying dimension case, and they are stated below in terms of the alterations only:

(i) *misspecified hazard for event of interest*:

$$\lambda_1(t|A,W) = \lambda_{0,1}(t)\exp(0.4W_1^2 + 0.5W_2W_3 + 0.05W_3A - 0.05W_4 + 0.5A)$$

(ii) *misspecified hazard for competing event*:

$$\lambda_2(t|A,W) = \lambda_{0,2}(t)\exp(0.2W_1^2 + 0.8W_2W_3 + 0.05W_3A - 0.05W_4 + 0.5A)$$

(iii) *misspecified propensity*: $P(A = 1 \mid W) = \Phi(1.2W_1 - 0.8W_2 + 0.05W_3 + 0.05W_4)$

where Φ is the distribution function of a standard normal distribution. Figure 5 displays the sampling distribution of the different estimators under the three scenarios. Again, the cross-fitted versions of the estimators decrease the bias, with ridge penalization generally producing biased estimates, and figure 6 shows that cross-fitting provides better coverage for each of the misspecification scenarios.

6 Results on non-response using penalized nuisance estimators

The original analysis was done in three settings, one primary and two sensitivity analyses, where the difference between settings corresponded to different exclusion criteria. We will concern our selves with the first of the sensitivity analyses (table 2aS in Kessing et al., 2023), where patients with a purchase of antidepressants within one year prior to their initial diagnosis with major depressive disorder are excluded. Additionally, we constrain the study period to 2005-2018, and somatic disease histories are defined within a period 10 years prior to index date in terms of ICD-10 codes giving a total of 1170 different observed somatic diseases (the ICD-10 is fully implemented in the Danish national health registers from 1995, and by restricting the study period to 2005-2018, we avoid a mixture of ICD-8 and ICD-10 codes). Based on the simulation studies, we chose to redo the sensitivity analysis with the cross-fitted estimator $\hat{\psi}_{\tau,CF}$ with $K = 10$ sample splits, and nuisance estimators given by 10-fold cross-validated lasso estimators, described in Section 4.3. Compared to the original study, the somatic disease histories were included based on individual disease levels in order to safeguard against confounding. The results are presented in Table 1 and the results from Kessing et al. (2023) are included in Table 2 for reference. For the SSRI group, the average treatment effect estimates are not much different compared to the results from the original study. Interestingly, though, is that the estimated absolute risk of non-response are generally lower when compared to the estimates in the original study, but as the ATE is comprised of the difference of the absolute risk estimate for the drug of interest and reference drug, the ATE estimate is largely unchanged. For the groups NARI, there were too few observations in the Reboxetine treatment group ($n = 37$) for our cross-fitted estimator to provide meaningful results. For the rest of the treatment groups, the picture is similar to the SSRI group. One thing to note is that the confidence intervals based on the cross-fitted estimator is generally wider compared to the original study. This should come as no surprise though, as the EIF provides lower information bounds in the nonparametric model, whereas the estimator in the original study assumes low-dimensional (semi)parametric nuisance models.

	Absolute risk: drug of interest	**Absolute risk: reference drug**	**Absolute risk difference**
SSRI (reference: Setraline)			
Citalopram	0.378 (0.363, 0.393)	0.363 (0.350, 0.375)	0.015 (-0.004, 0.035)
Fluoxetine	0.380 (0.354, 0.405)	0.353 (0.341, 0.366)	0.026 (-0.002, 0.055)
Paroxetine	0.454 (0.377, 0.531)	0.359 (0.347, 0.372)	0.095 (0.017, 0.172)
Escitalopram	0.453 (0.431, 0.475)	0.361 (0.349, 0.374)	0.092 (0.066, 0.117)
NARI (reference: Sertraline)			
Reboxetine	NA	NA	NA
SNRI (reference: Venlafaxine)			
Duloxetine	0.466 (0.431, 0.502)	0.452 (0.428, 0.475)	0.015 (-0.03, 0.057)
NaSSA (reference: Mirtazapine)			
Mianserin	0.552 (0.473, 0.630)	0.470 (0.450, 0.489)	0.082 (0.001, 0.162)
TCA (reference: Amitriptyline)			
Nortriptyline	0.558 (0.452, 0.664)	0.338 (0.290, 0.386)	0.220 (0.104, 0.336)
Imipramine	0.424 (0.248, 0.599)	0.329 (0.282, 0.375)	0.095 (-0.087, 0.277)
Clomipramine	0.631 (0.495, 0.766)	0.336 (0.290, 0.381)	0.295 (0.152, 0.438)
Dosulepin	0.652 (0.263, 1.000)	0.341 (0.294, 0.387)	0.311 (-0.080, 0.703)
Others (reference: Sertraline)			
Vortioxitine	0.342 (0.220, 0.464)	0.375 (0.355, 0.388)	-0.029 (-0.153, 0.094)
Agomelatine	0.498 (0.424, 0.572)	0.365 (0.351, 0.378)	0.133 (0.058, 0.209)

Table 1: Estimates of absolute risk and absolute risk differences of non-response at time $t = 2$ years with 95% confidence intervals. The cross-fitted estimator in Section 4.2 with $K = 10$ folds are used in the estimation with nuisance parameters estimated by the cross-validated lasso as described in Section 4.3.

	Absolute risk: drug of interest	**Absolute risk: reference drug**	**Absolute risk difference**
SSRI (reference: Setraline)			
Citalopram	0.42 (0.41, 0.43)	0.41 (0.40, 0.42)	0.01 (-0.01, 0.02)
Fluoxetine	0.47 (0.45, 0.49)	0.40 (0.39, 0.41)	0.07 (0.04, 0.10)
Paroxetine	0.44 (0.40, 0.49)	0.41 (0.39, 0.42)	0.03 (-0.01, 0.08)
Escitalopram	0.48 (0.46, 0.50)	0.41 (0.40, 0.42)	0.07 (0.05, 0.10)
NARI (reference: Sertraline)			
Reboxetine	0.60 (0.43, 0.76)	0.40 (0.39, 0.42)	0.19 (0.03, 0.35)
SNRI (reference: Venlafaxine)			
Duloxetine	0.52 (0.48, 0.56)	0.52 (0.50, 0.55)	0.00 (-0.05, 0.04)
NaSSA (reference: Mirtazapine)			
Mianserin	0.62 (0.56, 0.53)	0.53 (0.52, 0.55)	0.09 (0.02, 0.15)
TCA (reference: Amitriptyline)			
Nortriptyline	0.60 (0.56, 0.63)	0.40 (0.35, 0.44)	0.20 (0.14, 0.26)
Imipramine	0.46 (0.31, 0.61)	0.39 (0.34, 0.43)	0.07 (-0.08, 0.23)
Clomipramine	0.67 (0.59, 0.75)	0.40 (0.35, 0.44)	0.27 (0.18, 0.37)
Dosulepin	0.57 (0.46, 0.68)	0.40 (0.36, 0.45)	0.17 (0.05, 0.28)
Others (reference: Sertraline)			
Vortioxitine	0.37 (0.24, 0.49)	0.41 (0.39, 0.43)	-0.05 (-0.17, 0.08)
Agomelatine	0.49 (0.42, 0.57)	0.41 (0.39, 0.42)	0.08 (0.01, 0.16)

Table 2: Estimates of absolute risk and absolute risk differences of non-response at time $t = 2$ years with 95% confidence intervals from the original study in Kessing et al. (2023). The G-formula is used for estimation of the target parameter with the cause-specific hazards estimated by Cox regression.

7 Discussion

The study in Kessing et al. (2023) showcases some of the challenges in estimating the ATE based on registry data. In order to adhere to the assumption of no unmeasured confounding a potentially large set of covariates has to be included in the estimation of the ATE. This prevents the use of ordinary parametric models, and methods aimed at high-dimensional data has to be incorporated. When the outcome of interest is defined as time to an event, high-dimensional covariates can be incorporated using penalized Cox regression (Wu, 2012). In order to obtain valid inference on the resulting ATE estimate, methods from the semiparametric efficiency literature has to be employed.

In this article, we have constructed two estimators based on the EIF corresponding to the ATE defined as the mean difference in cumulative incidence functions. The estimators allow for the inclusion of penalized regression for estimation of the nuisance parameters given by the hazard functions for the event of interest, competing event and censoring, respectively, and treatment propensity. The first estimator is constructed as a one-step estimator based on the EIF (Kennedy, 2022) and the second is given by its cross-fitted version. The simulation studies in Section 5 indicate that the cross-fitted estimator is asymptotically linear with the EIF as its influence functions, under similar sparsity constraints as considered in Hou et al. (2021). Specifically, when either the treatment and censoring models or the models for the cause-specific hazards are sparse, with the other being of moderate sparsity compared to the sample size, the desired properties are achieved.

The proposed cross-fitted estimator was applied to the data from Kessing et al. (2023) with lasso estimation of the nuisance parameters, treating the somatic disease histories as a high-dimensional covariate. The ATE estimates were largely similar to the original study, but estimates of the risk of non-response under a given treatment were generally lower. Furthermore, the confidence intervals corresponding to the cross-fitted ATE estimator were generally wider compared to the original study. This is a price to pay, since the variance of the estimator is given by the EIF corresponding to ψ_τ defined on a nonparamteric model, where the EIF characterizes the lower information bound (Van der Vaart, 2000). Hence, by assuming a larger model, the variance of the cross-fitted estimator is naturally larger compared to the G-formula estimator used in Kessing et al. (2023), which assumes correctly specified Cox models for the event hazards.

The performance of the estimators derived in this paper are based on simulation studies which suggest that the double robustness of the remainder term shown in Rytgaard et al. (2023) can be related to the sparsity of the nuisance parameters, analogous to the results in Hou et al. (2021). Further theoretical analysis is needed to confirm that this relationship holds for the cross-fitted estimator considered here, but we leave this for future work.

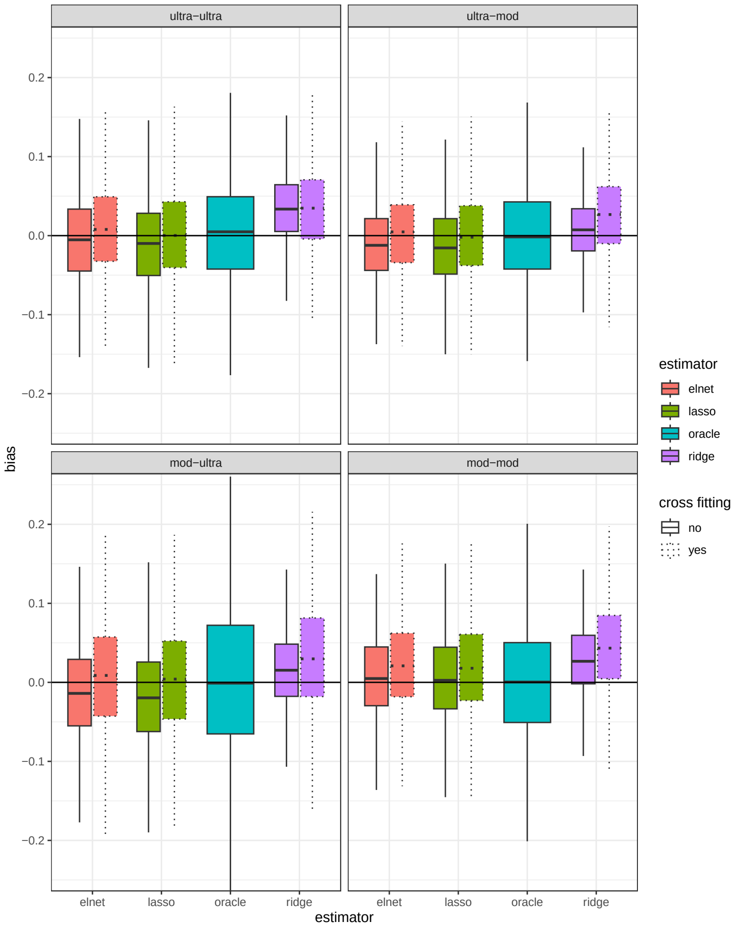


Figure 3: Sample distribution of estimators under varying nuisance sparsity according to the sparsity schemes described in Section 5.2. The plots shows boxplots corresponding to the (0.25,0.50,0.75)-quantile without extreme outliers for the different estimators based on 1000 simulations. The abbreviations in the title of the plots are read as A-B, where A corresponds to the sparsity of the intervention parameters and B corresponds to the event parameters. Thus, the upper left hand panel corresponds to (i), the upper right to (ii), the lower left to (iii) and the lower right to (iv).

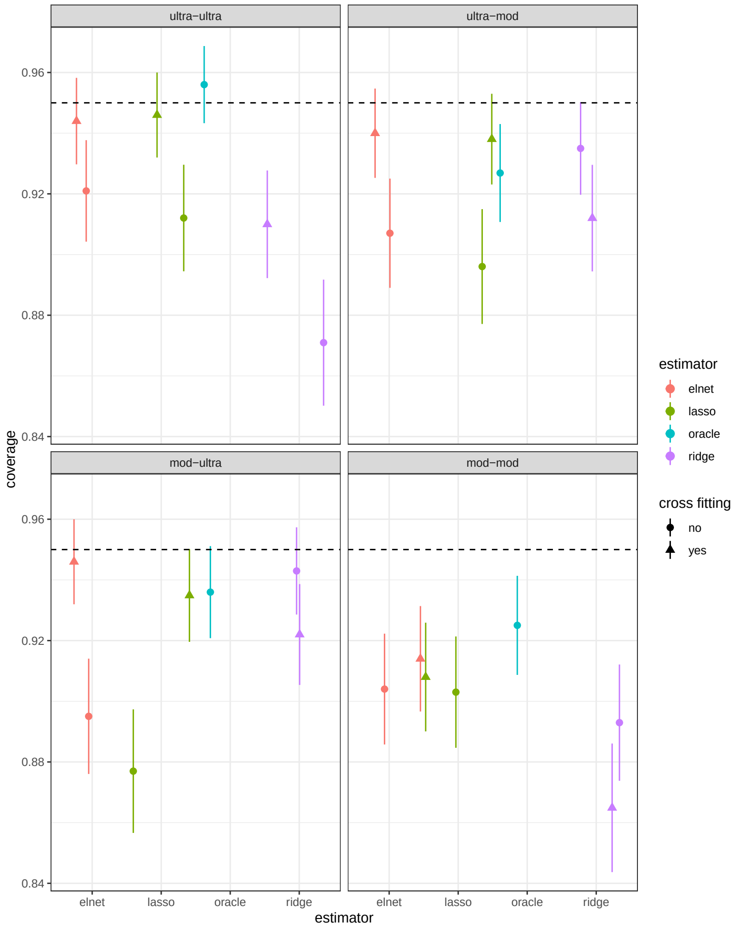


Figure 4: Coverage of estimators under varying nuisance sparsity based on 1000 simulations. Dots and triangles are the estimated coverage probabilities and the dashed bars denotes the Monte-Carlo uncertainty. The abbreviations in the title of the plots are read as A-B, where A corresponds to the sparsity of the intervention parameters and B corresponds to the event parameters. Thus, the upper left hand panel corresponds to (i), the upper right to (ii), the lower left to (iii) and the lower right to (iv).

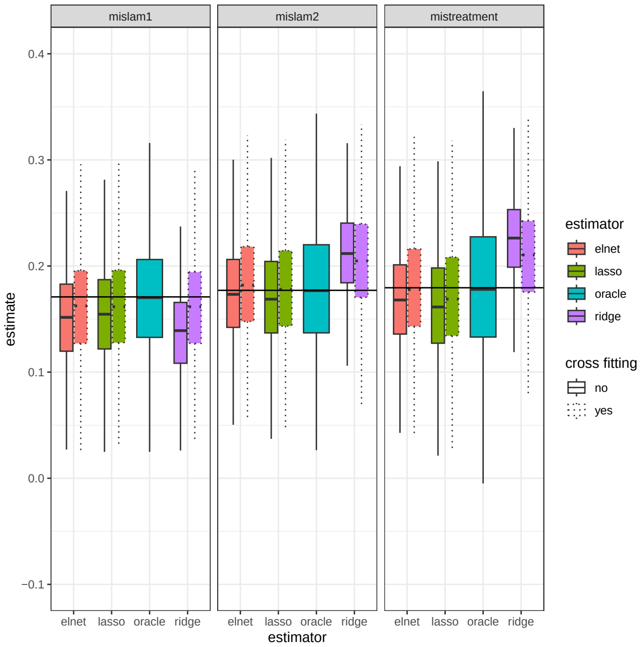


Figure 5: Sample distribution of the estimators under different types of misspecification of models for the nuisance parameters described in Section 5.3. The plots show boxplots corresponding to the (0.25,0.50,0.75)-quantile without extreme outliers for the different estimators based on 1000 simulations. The panel on the left corresponds to the misspecification given in (i), the middel panel corresponds to the setting given in (ii) and the right panel corresponds to setting given in (iii).

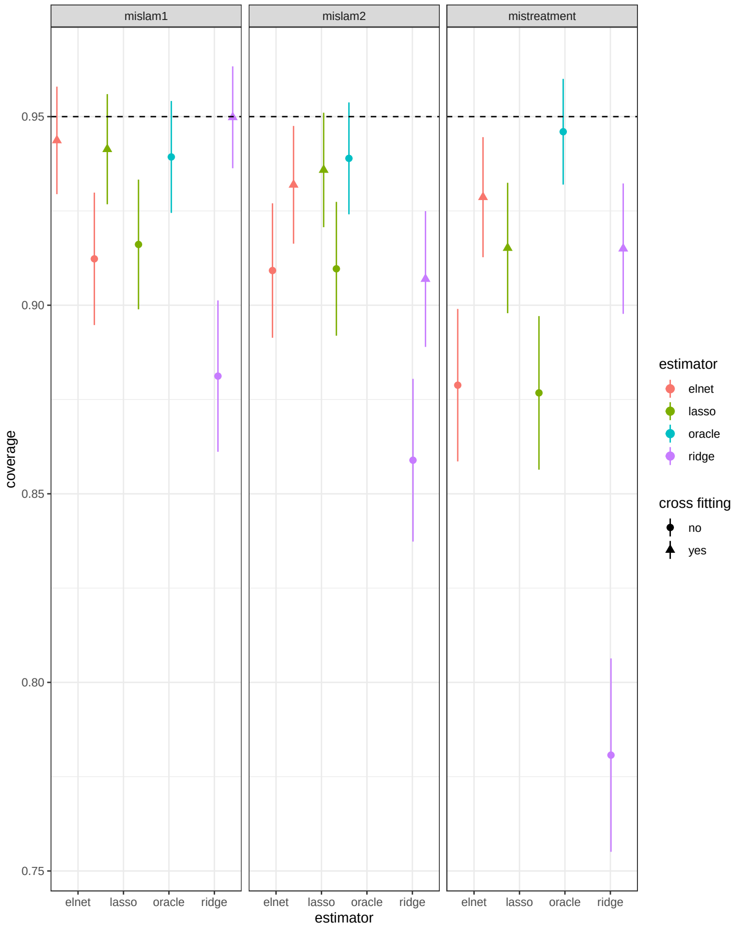


Figure 6: Coverage probability of confidence intervals of estimators under different types of misspecification of models for the nuisance parameters described in Section 5.3 based on 1000 simulations. Dots and triangles are the estimated coverage probabilities and the dashed bars denotes the Monte-Carlo uncertainty. The panel on the left corresponds to the misspecification given in (i), the middel panel corresponds to the setting given in (ii) and the right panel corresponds to setting given in (iii).

References

Bickel, P. J., Klaassen, C. A., Bickel, P. J., Ritov, Y., Klaassen, J., Wellner, J. A., & Ritov, Y. (1993). *Efficient and adaptive estimation for semiparametric models* (Vol. 4). Springer.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, *33*(1), 1.

Hines, O., Dukes, O., Diaz-Ordaz, K., & Vansteelandt, S. (2022). Demystifying statistical learning based on efficient influence functions. *The American Statistician*, *76*(3), 292–304.

Hou, J., Bradic, J., & Xu, R. (2021). Treatment effect estimation under additive hazards models with high-dimensional confounding. *Journal of the American Statistical Association*, 1–16.

Huang, J., Sun, T., Ying, Z., Yu, Y., & Zhang, C.-H. (2013). Oracle inequalities for the lasso in the cox model. *Annals of statistics*, *41*(3), 1142.

Kennedy, E. H. (2022). Semiparametric doubly robust targeted double machine learning: A review. *arXiv preprint arXiv:2203.06469*.

Kessing, L. V., Ziersen, S. C., Andersen, F. M. A., Gerds, T. A., & Budtz-Jørgensen, E. (2023). Comparative responses to 15 different antidepressants in major depressive disorder – results from a 2-year long-term nation-wide population-based study emulating a randomised trial. *BA*.

Martinussen, T., & Scheike, T. H. (2006). *Dynamic regression models for survival data* (Vol. 1). Springer.

Munch, A., Gerds, T. A., van der Laan, M. J., & Rytgaard, H. C. (2024). Estimating conditional hazard functions and densities with the highly-adaptive lasso. *arXiv preprint arXiv:2404.11083*.

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

Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical modelling*, *7*(9-12), 1393–1512.

Rytgaard, H. C. W., Eriksson, F., & van der Laan, M. (2023). Estimation of time-specific intervention effects on continuously distributed time-to-event outcomes by targeted maximum likelihood estimation. *Biometrics*.

Rytgaard, H. C. W., & van der Laan, M. J. (2022). Targeted maximum likelihood estimation for causal inference in survival and competing risks analysis. *Lifetime Data Analysis*, 1–30.

Simon, N., Friedman, J., Hastie, T., & Tibshirani, R. (2011). Regularization paths for cox's proportional hazards model via coordinate descent. *Journal of statistical software*, *39*(5), 1.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *58*(1), 267–288.

Van Der Vaart, A. (1991). On differentiable functionals. *The Annals of Statistics*, 178–204.

Van der Vaart, A. W. (2000). *Asymptotic statistics* (Vol. 3). Cambridge university press.

van der Laan, M. J., & Robins, J. M. (2003). *Unified methods for censored longitudinal data and causality*. Springer.

van der Laan, M. J., & Rose, S. (2011). *Targeted learning*. Springer.

Van Houwelingen, H. C., Bruinsma, T., Hart, A. A., Van't Veer, L. J., & Wessels, L. F. (2006). Cross-validated cox regression on microarray gene expression data. *Statistics in medicine*, *25*(18), 3201–3216.

Verweij, P. J., & Van Houwelingen, H. C. (1993). Cross-validation in survival analysis. *Statistics in medicine*, *12*(24), 2305–2314.

Westling, T., Luedtke, A., Gilbert, P. B., & Carone, M. (2023). Inference for treatment-specific survival curves using machine learning. *Journal of the American Statistical Association*, (just-accepted), 1–26.

Wu, Y. (2012). Elastic net for cox's proportional hazards model with a solution path algorithm. *Statistica Sinica*, *22*, 27.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, *67*(2), 301–320.

Manuscript II

Variable importance measures for heterogeneous treatment effect with survival outcome

Simon Christoffer Ziersen & Torben Martinussen

Details: In preparation

Variable importance measures for heterogeneous treatment effects with survival outcome

Simon Christoffer Ziersen & Torben Martinussen

Section of Biostatistics, University of Copenhagen

Abstract

Treatment effect heterogeneity plays an important role in many areas of causal inference and within recent years, estimation of the conditional average treatment effect (CATE) has received much attention in the statistical community. While accurate estimation of the CATE-function through flexible machine learning procedures provides a tool for prediction of the individual treatment effect, it does not provide further insight into the driving features of potential treatment effect heterogeneity. Recent papers have addressed this problem by providing variable importance measures for treatment effect heterogeneity. Most of the suggestions have been developed for continuous or binary outcome, while little attention has been given to censored time-to-event outcome, although this is prominent in many biostatistical applications. In this paper we extend the treatment effect variable importance measure (TE-VIM) proposed in Hines, Diaz-Ordaz, and Vansteelandt (2022) to a survival setting with censored outcome. We consider two measures of treatment effect and derive estimators in the new setting. The estimators share some structure to the one proposed in Hines, Diaz-Ordaz, and Vansteelandt (2022), thereby suggesting future extensions to other treatment effect measures. Along with TE-VIM, we derive a new measure of treatment effect heterogeneity based on the best partially linear projection of the CATE and we provide an estimator for that projection. All estimators are based on semiparametric efficiency theory, and we give conditions under which they are asymptotically linear. The finite sample performance of the derived estimators are investigated in a simulation study. Finally, the estimators are applied and contrasted in a real data example.

1 Introduction

Treatment effect heterogeneity plays an important role in medical research, as an understanding of such can be used in personalizing individual treatment plans as well as informing further research. The former point has received much attention in the causal inference community within the past decade, see for example Kennedy (2022b) and Wager and Athey (2018). Much work has focused on the Conditional Average Treatment Effect (CATE) given by the difference $\tau(x) = \mathrm{E}(Y^1 - Y^0 \mid X = x)$, where Y^1 and Y^0 are the counterfactual outcomes under treatment and no treatment, respectively. Under standard assumptions from the causal inference literature, including the assumption of no unmeasured confounding, the CATE can be identified from the observed data $\mathcal{O} = (Y_i, A_i, X_i)_{i=1}^n$, where Y_i, A_i, X_i correspond to the outcome, treatment and covariates of individual i, as $\tau(x) = \mathrm{E}(Y \mid A = 1, X = x) - \mathrm{E}(Y \mid A = 0, X = x)$.

Survival analysis is often complicated by the fact that one does not observe the full data, but only a censored version. Considering counterfactual survival times T^1 and T^0, and letting $Y^a(t) = \mathbb{1}(T^a \geq t)$, $a = 0, 1$, the CATE can be defined as $\tau(x; t) = \mathrm{E}(Y^1(t) - Y^0(t) \mid X = x)$, which, under the same causal assumptions, is identified by the observed data as $\tau(x; t) = \mathrm{E}(\mathbb{1}(T \geq t) \mid A = 1, X = x) - \mathrm{E}(\mathbb{1}(T \geq t) \mid A = 0, X = x)$ for a specific time horizon t. As the data is censored, the observed data is given by $\mathcal{O} = (\tilde{T}_i, \Delta_i, A_i, X_i)_{i=1}^n$ where $\tilde{T}_i = T_i \wedge C_i$ for a given censoring time C_i and $\Delta_i = \mathbb{1}(T_i \leq C_i)$. Under the additional assumption of conditional (on A and X) independent censoring, the CATE is still identified from the observed data as the difference in conditional survival functions: $\tau(x; t) = S(t \mid A = 1, X = x) - S(t \mid A = 0, X = x)$. Estimation of the CATE in the survival context has received some attention in the recent years: Cui et al. (2023) extend the work of Wager and Athey (2018) to a survival setting, Hu et al. (2021) compares different machine learning methods for estimating the CATE in a survival setting and Xu et al. (2023) discuss the use of different meta-learners in combination with arbitrary machine learning methods.

CATE estimation provides a tool for prediction of the individual treatment effect, but as the methods of obtaining such estimates are often based on machine learning, it provides little information as to which features are driving the observed heterogeneity (if any at all). As such, Levy et al. (2021) derives a measure of overall treatment effect heterogeneity as the variance of the treatment effect (VTE), given by $\mathrm{var}(\tau(X))$, Wei et al. (2023) derives an estimator for sub-group treatment effects, and Boileau et al. (2023) constructs a general framework for identification of treatment effect modifiers, as a weighted covariance of individual covariates and the CATE, which they also extend to a survival setting. Their approach can be viewed in terms of the *best linear projection* of the CATE-function, an approach also discussed in Van der Laan (2006) and Semenova and Chernozhukov (2021), but where the projection is used to approximate a target function (such as the CATE-function) rather than summary statistics of the CATE itself. Finally, Hines, Diaz-Ordaz, and Vansteelandt (2022) develop a treatment effect variable importance measure (TE-VIM), which measures the amount of the VTE explained by a given subset of covariates. Their derived estimand has the interpretation of a non-parametric ANOVA and can employ arbitrary machine learning methods for nuisance parameter estimation.

In this paper, we extend the TE-VIM of Hines, Diaz-Ordaz, and Vansteelandt (2022) for two different CATE functions for survival data. The derived estimator is based on semiparametric efficiency theory, and the efficient influence function (EIF) corresponding to the TE-VIM with censored data is seen to share some structure to the one proposed by Hines, Diaz-Ordaz, and Vansteelandt (2022). This connection is found to hold for essentially all $\tau(x)$, when the EIF corresponding to the ATE, $\mathrm{E}\{\tau(X)\}$, is linear in the ATE. Furthermore, we derive a new measure of treatment effect heterogeneity inspired by the assumption lean inference approach (Vansteelandt and Dukes, 2022a) and derive an estimator based on its corresponding efficient influence function. The new measure is derived as the *best partially linear projection* of the CATE and it can be interpreted as a regression coefficient, expressing the association between the CATE and a single covariate of interest. Other authors have suggested a similar approach (Boileau et al., 2023, Cui et al., 2023) for treatment effect variable importance, using the best linear projection of the CATE as a measure of heterogeneity, but, as we discuss in the Appendix, the error made by the projecting the CATE onto the linear model is larger compared to the projection onto the partially linear model, thus showing that our approach captures more of the heterogeneity through a single covariate compared the best linear projection. Furthermore, the derived parameter is seen to provide a natural interpretation of the association between the CATE and a given covariate when the partially linear model does not hold for the CATE function, as it is given as weighted average of the conditional covariance of the CATE and the covariate in question.

We give assumptions under which the proposed estimators are asymptotically normal and locally efficient, and investigate their finite sample performance in a simulation study, using random survival forests (Ishwaran et al., 2008) for nuisance estimation. Finally, we illustrate and contrast the two approaches in a data example also studied in Cui et al. (2023) and Hines, Diaz-Ordaz, and Vansteelandt (2022).

2 Notation and Setup

We consider a survival setup where T and C denote the survival and censoring time, respectively. Due to censoring, we do not observe T, but rather, we observe the censored time $\tilde{T} = T \wedge C$ together with the event indicator $\Delta = \mathbb{1}\{T \leq C\}$. Furthermore let $A \in \{0,1\}$ denote a binary treatment variable at baseline and let $X = (X_1, ..., X_d) \in \mathbb{R}^d$ denote baseline covariates. The observed data is $\mathcal{O} = (\tilde{T}_i, \Delta_i, A_i, W_i)_{i=1,..,n}$ where $O_1, ..., O_n$ are assumed to be i.i.d. with distribution $P_0 \in \mathcal{M}$, where $\mathcal{M}$ is the set of all probability measures corresponding to a non-parametric model.

Let $N(t) = \mathbb{1}\{\tilde{T} \leq t, \tilde{\Delta} = 1\}$ be the observed counting process for the event of interest and let $\lambda(t|a,x)$, $\lambda_c(t|a,x)$ denote the conditional hazard for the survival and censoring distribution, respectively, and let $\Lambda(t|a,x) = \int_0^t \lambda(s \mid a,x)\,\mathrm{d}s$, $\Lambda_c(t|a,x) = \int_0^t \lambda_c(s \mid a,x)\,\mathrm{d}s$ denote the corresponding cumulative hazard functions. Furthermore we denote $S(t|a,x) = \exp(-\Lambda(t|a,x))$ the survival function, $S_c(t|a,x) = \exp(-\Lambda_c(t|a,x))$ the survival function of the censoring distribution, $\pi(a|x) = P(A = a|X = x)$ is the propensity score and μ is the distribution of X. Throughout, we denote $M(t \mid A, X)$ as the martingale related to $N(t)$, where $dM(t \mid A,X) = dN(t) - \mathbb{1}(\tilde{T} \geq t)d\Lambda(t|A,X)$ is the martingale increment given A and X.

To define causal parameters, we introduce the variable $Y(t) = \mathbb{1}\{T \geq t\}$ and define $Y^a(t)$ as the counterfactual outcome, that is, the outcome of a person if he or she, possibly contrary to the fact, had received treatment a. Let

$$\tau(x;t) = \mathrm{E}(Y^1(t) - Y^0(t)|X = x)$$

be the CATE function, i.e. the average treatment effect conditional on the event $X = x$ for some fixed time-horizon t, which is left out from the notation through out the paper, so we write $\tau(x) = \tau(x;t)$. Under suitable causal assumptions we can identify τ through the observed data as

$$\tau(x) = S(t|A = 1, X = x) - S(t|A = 0, X = x). \tag{1}$$

Another, and maybe more interesting $\tau(x)$, is

$$\tau(x) = E\left(T^1 \wedge t - T^0 \wedge t \mid X = x\right) = \int_0^t S(u \mid 1,x)\,du - \int_0^t S(u \mid 0,x)\,du. \tag{2}$$

We will consider both in what follows, where we will refer to the first as the *survival function setting* and to the second as the *restricted mean survival time setting (RMST).* Note that $\mathrm{E}\{\tau(X)\}$ is simply the average treatment effect. Furthermore we define

$$\tau_l(x) = \mathrm{E}(\tau(X)|X_{-l} = x_{-l})$$

as the conditional expectation of the CATE-function, where we fix all variables except X_l, $l \subseteq \{1, ..., d\}$, as we define X_{-l} to be the covariates with an index not contained in l. We will use the notation $\tau_d = \mathrm{E}(\tau(X))$ to denote the average treatment effect.

Furthermore, we introduce the nuisance parameter $\nu = (\Lambda, \Lambda_c, \tau_l, \mu)$.

3 Target parameter

3.1 Treatment effect variable importance measure

As in Hines, Diaz-Ordaz, and Vansteelandt, 2022 we define

$$\Theta_l \equiv \mathrm{E}(\mathrm{var}(\tau(X) \mid X_{-l} = x_{-l})) = \mathrm{var}(\tau(X)) - \mathrm{var}(\tau_l(X)) \geq 0.$$

With a slight abuse of notation we denote the VTE: $\Theta_d = \mathrm{var}(\tau(X))$. We note that Θ_l can be interpreted as the amount of heterogeneity not already explained by X_{-l}, as Θ_l is large when a large amount of the VTE is explained by X_l. The proposed treatment effect variable importance measure (TE-VIM) is defined by re-scaling Θ_l by the VTE:

$$\Psi_l \equiv \frac{\Theta_l}{\mathrm{var}(\tau(X))} = 1 - \frac{\mathrm{var}(\tau_l(X))}{\mathrm{var}(\tau(X))}$$

with values in $[0, 1)$. We can interpret Ψ_l as a nonparametric analog of an ANOVA statistic, which is close to one when a large amount of the VTE is explained by X_l and close to zero when a small amount of the VTE is explained by X_l.

3.2 Best partially linear projection

Along with the TE-VIM, we also consider an alternative target parameter inspired by Vansteelandt and Dukes (2022b), which is given by

$$\Omega_j = \frac{\Gamma_j}{\chi_j} = \frac{\mathrm{E}\left(\mathrm{cov}(X_j, \tau(X) \mid X_{-j})\right)}{\mathrm{E}(\mathrm{var}(X_j \mid X_{-j}))}$$

for a single covariate X_j, corresponding to the singleton set $\{j\} \in \{1, \ldots, d\}$. It is seen that the parameter depends on the scale of the covariate of interest, X_j, and as such, the variable importance of X_j will be based on the p-value for the test of the hypothesis $H : \Omega_j = 0$.

In contrast to Ψ_l, the parameter Ω_j measures the heterogeneity explained by a single covariate X_j, $j \in \{1, \cdots, d\}$, whereas Ψ_l determines the heterogeneity explained by, possibly non-singular, sets of covariates. This makes Ψ_l suited for incorporating subject matter knowledge, where naturally correlated covariates can be grouped together, where Ω_j serves as a variable importance measure to be used for single covariates.

The estimand Ω_j can be expressed by the linear term in the projection of τ onto the space of partially linear functions. To elaborate, let $\beta \in \mathbb{R}$ and let w be some measurable function of X_{-j} with finite variance. Without loss of generality, define

$$\tau(x) = \beta x_j + w(x_{-j}) + R(x_j, x_{-j})$$

for some function R and let

$$(\beta^*, w^*) = \arg\min_{\beta,w} \mathrm{E}\{R(X_j, X_{-j})^2\} = \arg\min_{\beta,w} \mathrm{E}\{[\tau(X) - \beta X_j - w(X_{-j})]^2\}$$

be the least squares projection of τ onto the partially linear model. Then $\Omega_j = \beta^*$.

If $R = 0$, then for a given level of x_{-j}, the parameter β denotes the treatment effect modification given by x_j. When $R \neq 0$, Ω_j is the treatment effect modification parameter that minimizes the error made by summarizing the effect of x_j in a single value. We note that other authors have looked at the *best linear projection* as a treatment effect variable importance measure (Semenova and Chernozhukov, 2021, Boileau et al., 2023, Cui et al., 2023, Van der Laan, 2006). In our setting, this corresponds to the least squares projection of τ onto the space of linear models. In the Appendix, we give a discussion of partially linear versus linear projections of τ in terms of the size of the error given by the remainder term R.

4 Estimation

The estimation of Ψ_l goes through estimation of Θ_l and Θ_d, separately. Likewise, an estimator of Ω_j is obtained from estimators of Γ_j and χ_j. The estimation of individual parameters is based on semi(non)-parametric efficiency theory. For an introduction to this methodology see for instance (Kennedy, 2022a, Hines, Dukes, et al., 2022, van der Laan and Robins, 2003, van der Vaart, 2000 ch. 25). The theory revolves around the so-called efficient influence function (EIF), which characterizes the lower bound on the asymptotic variance of any regular estimator of a pathwise differentiable parameter in a non-parametric setting. The EIF is related to the target parameter and the model $\mathcal{M}$, and it can be calculated without reference to any estimator. Once it is known, it can be leverage to construct an estimator that is asymptotically linear with the EIF as its influence function. Several techniques exist for constructing such estimators, and they all share the convenient property that it is possible to use data-adaptive nuisance parameter estimators (under some conditions), while still obtaining parametric-like inference on the target parameter.

Hence, estimation of Ψ_l and Ω_j will follow the same pattern, where the EIF is calculated at first, to then be used in the construction of an estimator for the target parameter in question.

4.1 Estimation of Ψ_l

4.1.1 Efficient influence function

The two target parameters Θ_l and Ψ_l are functions of $\mathrm{var}(\tau(X))$ and $\mathrm{var}(\tau_l(X))$ so their efficient influence functions can be derived from the EIFs of $\mathrm{var}(\tau(X))$ and $\mathrm{var}(\tau_l(X))$ using the chain rule (cf. van der Vaart, 2000 ch. 25.7.). Define

$$H(u, t \mid a, x) = \int_u^t S(u \mid a, x)\, \mathrm{d}u$$

and

$$g(A, X) = \left(\frac{\mathbb{1}(A = 1)}{\pi(1 \mid X)} - \frac{\mathbb{1}(A = 0)}{\pi(0 \mid X)} \right),$$

where $\pi(a \mid X) = P(A = a \mid X)$. We have the following result.

Theorem 1. *Let $\tau(x)$ be given by (1). The efficient influence functions of $var(\tau(X))$ and $var(\tau_l(X))$ are given by $\tilde{\psi}_{var(\tau(X))}$ and $\tilde{\psi}_{var(\tau_l(X))}$, respectively, where*

$$\begin{aligned}\tilde{\psi}_{var\{\tau(X)\}} =& [\tau(X) - E\{\tau(X)\}]^2 - var\{\tau(X)\} - 2[\tau(X) - E\{\tau(X)\}] \\ & \times g(A, X) \int_0^t \frac{S(t \mid A, X)}{S(u \mid A, x) S_c(u \mid A, X)} \mathrm{d}M(u \mid A, X) \\ \tilde{\psi}_{var\{\tau_l(X)\}} =& [\tau_l(X) - E\{\tau_l(X)\}]^2 - var\{\tau(X)\} - 2[\tau_l(X) - E\{\tau_l(X)\}] \\ & \times \left(\tau_l(X) - \tau(X) + g(A, X) \int_0^t \frac{S(t \mid A, X)}{S(u \mid A, X) S_c(u \mid A, X)} \mathrm{d}M(u \mid A, X)\right).\end{aligned}$$

For $\tau(x)$ given by (2) we have

$$\begin{aligned}\tilde{\psi}_{var\{\tau(X)\}} =& [\tau(X) - E\{\tau(X)\}]^2 - var\{\tau(X)\} - 2[\tau(X) - E\{\tau(X)\}] \\ & \times g(A, X) \int_0^t \frac{H(u, t, A, X)}{S(u \mid A, X) S_c(u \mid A, X)} \mathrm{d}M(u \mid A, X) \\ \tilde{\psi}_{var\{\tau_l(X)\}} =& [\tau_l(X) - E\{\tau_l(X)\}]^2 - var\{\tau(X)\} - 2[\tau_l(X) - E\{\tau_l(X)\}] \\ & \times \left(\tau_l(X) - \tau(X) + g(A, X) \int_0^t \frac{H(u, t, A, X)}{S(u \mid A, X) S_c(u \mid A, X)} \mathrm{d}M(u \mid A, X)\right).\end{aligned}$$

In both the survival function and RMST setting the EIF's corresponding to Θ_l and Ψ_l are given by $\tilde{\psi}_{\Theta_l}$ and $\tilde{\psi}_{\Psi_l}$, respectively, where

$$\begin{aligned}\tilde{\psi}_{\Theta_l} &= \tilde{\psi}_{var(\tau(X))} - \tilde{\psi}_{var(\tau_l(X))}, \\ \tilde{\psi}_{\Psi_l} &= \frac{1}{var(\tau(X))} \left(\tilde{\psi}_{\Theta_l}(O) - \Psi_l \tilde{\psi}_{var(\tau(X))}(O)\right).\end{aligned}$$

Proof. See Appendix B. □

Before moving to estimation of Θ_l (and Θ_d) we state some results from ATE-estimation. Recall the average treatment effect as the mean of the CATE;

$$\tau_d = \mathrm{E}\{\tau(X)\}.$$

The parameter τ_d has an EIF known from the literature in the survival function setting (e.g. Rytgaard et al., 2023 and Westling et al., 2023), and we write it in terms of the uncentered EIF, φ, defined as:

$$\varphi(O) - \tau_d = \varphi_1(O) - \varphi_0(O) - \tau_d \tag{3}$$

with

$$\varphi_a(O) = S(t \mid A = a, X) - \frac{\mathbb{1}(A = a)}{\pi(a \mid X)} \int_0^t \frac{S(t \mid A, X)}{S(u- \mid A, X) S_C(u- \mid A, X)} dM(u \mid A, X). \tag{4}$$

Lemma B.1 in Appendix B, gives the Gateaux derivative of $\tau(x)$ in the RMST setting as

$$\frac{\mathbb{1}(X = x)}{f(x)} \frac{\mathbb{1}(A = a)}{\pi(a \mid X)} \int_0^t \frac{-H(u, t \mid A, X))}{S(u- \mid A, X) S_C(u- \mid A, X)} dM(u \mid A, X)$$

from which it follows that τ_d has efficient influence function given by

$$\varphi(O) - \tau_d = \varphi_1(O) - \varphi_0(O) - \tau_d$$

with

$$\varphi_a(O) = \int_0^t S(u \mid a, X)\,\mathrm{d}u - \frac{\mathbb{1}(A = a)}{\pi(a \mid X)} \int_0^t \frac{H(u, t \mid A, X))}{S(u- \mid A, X) S_C(u- \mid A, X)} dM(u \mid A, X), \quad (5)$$

analogous to the survival function setting. The uncentered EIF, φ, can be parameterized by parts of the nuisance parameter, $(\pi, \Lambda, \Lambda_c)$, and we write $\varphi(\pi, \Lambda, \Lambda_c)$ when we want to be explicit about the nuisance parameters considered, which will be the case when we consider estimators for φ, where $\hat{\varphi} = \varphi(\hat{\pi}, \hat{\Lambda}, \hat{\Lambda}_c)$ is an obvious candidate.

We can now restate the EIF's given in Theorem 1 so that the structure is similar to that given in Hines, Diaz-Ordaz, and Vansteelandt (2022), but with the φ, τ and τ_l having different expressions. We adopt their notation of the VTE as $\Theta_d = \mathrm{var}\{\tau(X)\}$

Corollary 1. *The EIF of Θ_l and Θ_d is given by $\tilde{\psi}_{\Theta_l}$ and $\tilde{\psi}_{\Theta_d}$, respectively, where*

$$\tilde{\psi}_{\Theta_l} = (\varphi(O) - \tau_l(X))^2 - (\varphi(O) - \tau(X))^2 - \Theta_l \quad (6)$$

$$\tilde{\psi}_{\Theta_d} = (\varphi(O) - \tau_d)^2 - (\varphi(O) - \tau(X))^2 - \Theta_d \quad (7)$$

$$\tilde{\psi}_{\Psi_l} = \frac{1}{\Theta_d}\left(\tilde{\psi}_{\Theta_l} - \Psi_l \tilde{\psi}_{\Theta_d}\right) \quad (8)$$

Proof. Note that in both the survival and RMST setting, the EIFs of $\mathrm{var}\{\tau(X)\}$ and $\mathrm{var}\{\tau_l(X)\}$ can be written as

$$\tilde{\psi}_{\mathrm{var}\{\tau(X)\}} = [\tau(X) - \mathrm{E}\{\tau(X)\}]^2 + 2[\tau(X) - \mathrm{E}\{\tau(X)\}][\varphi(O) - \tau(X)] - \mathrm{var}\{\tau(X)\}$$

and

$$\tilde{\psi}_{\mathrm{var}\{\tau_l(X)\}} = [\tau(X) - \mathrm{E}\{\tau(X)\}]^2 + 2[\tau(X) - \mathrm{E}\{\tau(X)\}][\varphi(O) - \tau_l(X)] - \mathrm{var}\{\tau_l(X)\}.$$

A simple rewriting of the above EIFs gives

$$\begin{aligned}\tilde{\psi}_{\mathrm{var}\{\tau(X)\}} =& [\tau(X) - \mathrm{E}\{\tau(X)\}]^2 - 2\tau(X)^2 + 2\tau(X)\,\mathrm{E}\{\tau(X)\} \\ &+ 2[\tau(X) - \mathrm{E}\{\tau(X)\}]\varphi(O) - \mathrm{var}\{\tau(X)\} \\ =& \mathrm{E}\{\tau(X)\}^2 - \tau(X)^2 + 2[\tau(X) - \mathrm{E}\{\tau(X)\}]\varphi(O) - \mathrm{var}\{\tau(X)\} \\ =& [\varphi(O) - \tau_d]^2 - [\varphi(O) - \tau(X)]^2 - \mathrm{var}\{\tau(X)\}\end{aligned}$$

and analogously for $\mathrm{var}\{\tau_l(X)\}$:

$$\tilde{\psi}_{\mathrm{var}\{\tau(X)\}} = [\varphi(O) - \tau_d]^2 - [\varphi(O) - \tau_l(X)]^2 - \mathrm{var}\{\tau_l(X)\}.$$

Subtracting the two gives the EIF for Θ_l and the chain rule gives the EIF for Ψ_l. □

Note that the above EIFs has the same structure whether we are in the survival function setting or the RMST setting, but with φ having a different expression. For the rest of the paper we will use the form of the EIFs given in Corollary 1.

Remark 1. The fact that the structure of the EIFs is identical to the one in Hines, Diaz-Ordaz, and Vansteelandt (2022), stems from the definition of EIFs as derivatives for which the chain-rule apply. From the derivations of the EIFs in the Appendix, it is seen that for any function $\tau(x)$ with Gateaux derivative given by $\frac{\mathbb{1}(X=x)}{f(x)}g(z)$, for some function g and some variable z, the EIFs will have the same structure as in (6) and (7) with φ being the uncentered EIF of $\mathrm{E}\{\tau(X)\}$ where $\varphi = \tau + g$. Thus, the framework of Hines, Diaz-Ordaz, and Vansteelandt (2022) can readily be extended to any data setting by calculating EIF of the ATE in that setting and denoting the uncentered version φ. The properties of estimators derived by this approach will have to be studied case by case, though, as will be apparent in the following.

4.1.2 Cross-fitted one-step estimators

The EIFs of Θ_l and Θ_d are used to construct estimators that are asymptotically linear with influence function given by the EIFs above, from which they are seen to be locally asymptotically efficient and asymptotically normal distributed. Given two such estimators, $\hat{\Theta}_l$ and $\hat{\Theta}_d$, an application of the delta method gives that $\hat{\Psi}_l = \frac{\hat{\Theta}_l}{\hat{\Theta}_d}$ is asymptotically linear with influence function given by (8) (see van der Vaart, 2000 ch. 25.7). For readability and ease of notation we only consider construction of an estimator for Θ_l in the following, but since the EIFs of Θ_l and Θ_d have a similar structure, the derived estimators will be the same with l replaced by d.

There are different ways of constructing such estimators; one-step estimators, estimating equation based, and targeted minimum loss-based estimators (TMLE). All of them require that the nuisance parameters are estimated fast enough such that the resulting remainder term and empirical process term (see Section C in the Appendix) converge at rate $n^{-1/2}$. We will focus on the estimating equation based estimator, which is given as the solution to $\mathbb{P}_n\tilde{\psi}_{n,\Theta_l} = 0$ in Θ_l, where $\tilde{\psi}_{n,\Theta_l}$ denotes the EIF with estimated nuisance parameters. Because the EIF is linear in Θ_l, this will correspond to the one-step estimator, where $\mathbb{P}_n\tilde{\psi}_{n,\Theta_l}$ is added to a plug-in estimate of Θ_l:

$$\hat{\Theta}_l = \mathbb{P}_n(\hat{\varphi} - \hat{\tau}_l)^2 - (\hat{\varphi} - \hat{\tau})^2,$$

and analogously

$$\hat{\Theta}_d = \mathbb{P}_n(\hat{\varphi} - \hat{\tau}_d)^2 - (\hat{\varphi} - \hat{\tau})^2,$$

where $\hat{\varphi} = \varphi(\hat{\pi}, \hat{\Lambda}, \hat{\Lambda}_c)$. Note that the estimation of τ_l can be obtained as a regression of $\hat{\tau}(X)$ onto X_{-l}, whereas the estimation of $\tau_d = \mathrm{E}(\tau(X))$ can be obtained by the mean of $\hat{\tau}(X)$, i.e., the marginal distribution, μ is estimated with the empirical measure $\mathbb{P}_n$. Or, as τ_d is itself a differentiable parameter, more sophisticated methods can be used in constructing estimators $\hat{\tau}_d$ (see section 4.3). The $n^{-1/2}$-convergence of the empirical process term related to the one-step estimator depends on the flexibility of the nuisance estimators, in the sense that, e.g., working parametric models ensure $n^{-1/2}$-convergence, which is not the case for some data-adaptive estimators. More specifically, if the nuisance estimators falls in a Donsker class which also contains the true nuisance parameter, then $n^{-1/2}$-convergence of the empirical process term is obtained. To alleviate the Donsker class condition, we employ a type of sample splitting (coined cross-fitting, Chernozhukov et al., 2018) which ensures the desired convergence as long as the nuisance estimators are $L_2(P)$-consistent. We will now detail the sample splitting, but we note that this is a general construction of cross-fitted one-step estimators (Kennedy, 2022a).

Split the index set $\{1,\dots,n\}$ (uniformly at random) into K disjoint sets $\mathcal{T}_1,\mathcal{T}_2,\dots,\mathcal{T}_K$, such that $\{1,\dots,n\}=\dot\cup_{k=1}^{k}\mathcal{T}_k$. Let $\mathcal{V}_k$ denote the subset of the observed data corresponding to the k'th index set, $\mathcal{T}_k$, i.e., $\mathcal{V}_k=\{O_i:\ i\in\mathcal{T}_k\}$, such that $\mathcal{O}=\dot\cup_{k=1}^{K}\mathcal{V}_k$. Let $\mathbb{P}_n^k$ be the empirical measure in the sample $\mathcal{V}_k$ and let $\phi_{\Theta_l}=(\varphi-\tau_l)^2-(\varphi-\tau)^2$ denote the uncentered EIF of Θ_l with $\hat\phi_{\Theta_l}$ being an estimate obtained by plugging in estimated nuisance parameters in the expression for ϕ. Let $\hat\phi_{\Theta_l,-k}$ be the estimate of ϕ_{Θ_l} based on data in $\mathcal{V}_{-k}=\cup_{i\neq k}\mathcal{V}_i$. The cross-fitted one-step estimator is then given by

$$\hat\Theta_l^{CF}=\sum_{i=k}^{K}\frac{n_k}{n}\mathbb{P}_n^k\hat\phi_{\Theta_l,-k}=\frac{1}{n}\sum_{k=1}^{K}\sum_{i\in\mathcal{T}_k}\left\{(\hat\varphi_{-k}(O_i)-\hat\tau_{s,-k}(X_i))^2-(\hat\varphi_{-k}(O_i)-\hat\tau_{-k}(X_i))^2\right\}$$

where n_k is the number of observations in $\mathcal{V}_k$ and $\hat\varphi_{-k}=\varphi(\hat\Lambda_{-k},\hat\Lambda_{C,-k},\hat\pi_{-k})$ with $(\hat\Lambda_{-k},\hat\Lambda_{C,-k},\hat\pi_{-k})$ being the nuisance estimators obtained from the sample $\mathcal{V}_{-k}$. We let $\hat\Theta_d^{CF}$ be defined in the same way with d instead of s in the above formula and note that $\hat\tau_{d,-k}$ corresponds to the ATE estimate based on the data in $\mathcal{V}_{-k}$ (see 4.3).

Next we state a set of conditions from which the asymptotic distribution of $\hat\Theta_l^{CF}$ is obtained.

Assumption A (Nuisance parameters)**.** *Let $g(s\mid a,x)=\pi(a\mid x)S_c(s\mid a,x)$. For nuisance estimates $\hat\pi,\hat\Lambda$, $\hat\Lambda_C$ define*

- $\hat\tau(x)=e^{-\hat\Lambda(t|A=1,x)}-e^{-\hat\Lambda(t|A=0,x)}$
- $\hat g(s\mid a,x)=\hat\pi(a\mid x)e^{-\hat\Lambda_c(s|a,x)}$
- $\hat\tau_l(x)=\hat E_n(\hat\tau(X)\mid X_{-l}=x_{-l})$

where $\hat E_n$ is some regression of $\hat\tau(X)$ onto X_{-l}. Define $\hat L(s,t\mid a,x)=\frac{S(s|a,X)}{\hat S(s|a,X)}\hat S(t\mid a,X)$ in the survival function setting and $\hat L(s,t\mid a,x)=\frac{\hat H(s,t|a,X)S(s|a,X)}{\hat S(s|a,X)}$ in the RMST setting. Assume that the nuisance parameters are chosen such that

A1 $\exists\eta>0,\ s.t.\ \eta<\hat g(s\mid a,x)$ *and* $\eta<e^{-\hat\Lambda(s|a,x)}\quad\forall(s,a,x)\in[0,t]\times\{0,1\}\times\mathcal{X}$.

A2 $\|\hat\tau(x)-\tau(x)\|_{L_2(P)}=o_p(n^{-\frac{1}{4}})$.

A3 $\|\hat\tau_l(x)-\tau_l(x)\|_{L_2(P)}=o_p(n^{-\frac{1}{4}})$.

A4 $E\Big\{\int_0^t\Big(1-\frac{g(s|a,X)}{\hat g(s|a,X)}\Big)\hat L(s,t\mid a,x)\,\mathrm{d}\Big[\Lambda(s\mid a,X)-\hat\Lambda(s\mid a,X)\Big]\Big\}=o_p(n^{-\frac{1}{2}})$.

A5 $\|\sup_{s<t}|\hat g(s\mid a,x)-g(s\mid a,x)|\,\|_{L_2(P)}=o_p(1)$.
$\Big\|\sup_{s<t}\Big|\hat\Lambda(s\mid a,x)-\Lambda(s\mid a,x)\Big|\,\Big\|_{L_2(P)}=o_p(1)$.

A6 $|\hat\tau_l(x)-\hat\tau(x)|\leq\delta<\infty$ *for almost all x.*

Theorem 2. *Assume that assumption A hold for the nuisance estimators in each data split $\mathcal{V}_{-k}$ and assume $\Theta_l > 0$. Then $\hat\Theta_l^{CF}$ is asymptotically linear with influence function given by $\tilde\psi_{\Theta_l}$,* (6), *and*

$$\sqrt{n}(\hat\Theta_l^{CF} - \Theta_l) \xrightarrow{D} \mathcal{N}(0, P\tilde\psi^2_{\Theta_l})$$

Proof. See Section C. □

Before proceeding to estimation of Θ_d, some comments about assumption A are in order. A1 is a common positivity assumption, which states that all individuals have a positive probability of receiving treatment and being under observation for the entire time horizon. Assumption A2 and A3 refer to convergence rates of the CATE function estimate as well as the conditional CATE function estimate. For CATE estimates given by $\hat\tau(x) = e^{-\hat\Lambda(t|A=1,x)} - e^{-\hat\Lambda(t|A=0,x)}$, the assumption boils down to an assumption on the convergence rate of the survival function estimate, but this is seen to be a mild assumption for which many ML-methods concur (see e.g. the discussion in Section 4.3 in Kennedy, 2022a). We note though, that assumption A2 and A3 imply that the estimator is not double robust in the sense that one only needs the outcome or the censoring and propensity to be correctly specified in order to obtain a consistent estimator, which is in contrast to other related estimands (e.g. the ATE, see Westling et al., 2023, theorem 2). From the structure of the remainder term in the Appendix, it is seen that the estimator is consistent if Λ, τ and τ_l are consistent, but that it is not the case if only the censoring and propensity are consistent. Thus, it is important to employ flexible methods for obtaining nuisance estimators $\hat\Lambda$, $\hat\tau$, and $\hat\tau_l$. Furthermore, since $\tau_l(x) = \mathrm{E}(e^{-\Lambda(t|1,X)} - e^{-\Lambda(t|0,X)} \mid X = x)$ it will generally be a complicated function, even for a correctly specified $\hat\Lambda$ (e.g. as a Cox regression with a Breslow baseline hazard), emphasizing the need for a flexible estimator $\hat{\mathrm{E}}_n$.

The assumption A4 corresponds to a bound on the aforementioned remainder term (see proof of Theorem 2 in the appendix). In studies on related target parameters (e.g. the ATE) with uncensored outcome, the related bound on the remainder term is seen to have a product structure, in the sense that the product of the $L_2(P)$-norms of the outcome regression and the propensity estimator needs to be $o_p(n^{-1/2})$ (see Kennedy, 2022a). This is then achieved if both estimators converge on $n^{-1/4}$ rate or, e.g., if one estimator is bounded in probability and the other converges on parametric rate. In our case $\hat\Lambda$ will often be a step function (see next subsection) and one has to study A4 in greater detail in order to obtain a product structure result analogous to the uncensored case. This is beyond the scope of this paper, but we will expect it to be the case in many settings. A5 corresponds to uniform consistency of the time-to-event nuisance parameters. Assumption A6 is a technical assumption, which we would expect to hold for most reasonable choices of $\hat{E}_n$.

Finally, we state a distribution result for an estimate of Ψ_l based on the cross-fitted estimators $\hat\Theta_l^{CF}$ and $\hat\Theta_d^{CF}$

Corollary 2. *Let $\hat\Psi_l^{CF} = \frac{\hat\Theta_l^{CF}}{\hat\Theta_d^{CF}}$. Under assumption A and $\|\hat\tau_d - \tau_d\|_{L_2(P)} = o_p(n^{-\frac{1}{4}})$, $\hat\Psi_l^{CF}$ is asymptotically efficient with influence function given by* (8) *and*

$$\sqrt{n}(\hat\Psi_l^{CF} - \Theta) \xrightarrow{D} \mathcal{N}(0, P\tilde\psi^2_{\Psi_l}).$$

To estimate the variance of $\hat\Psi_l^{CF}$, we define the cross-fitted plug-in estimator of $\sigma^2_{\Psi_l} = P\tilde\psi^2_{\Psi_l}$. Let $\hat{\tilde\psi}_{\Psi_l,-k}$ denote the estimate of the EIF $\tilde\psi_{\Psi_l}$ based on data from $\mathcal{V}_{-k}$. Define the variance estimator

$$\hat{\sigma}^{2,CF}_{\Psi_l} = \sum_{i=k}^{K} \frac{n_k}{n} \mathbb{P}_n^k \hat{\tilde{\psi}}^2_{\Psi_l,-k} = \frac{1}{n} \sum_{k=1}^{K} \sum_{i \in \mathcal{T}_k} \left[\frac{1}{\hat{\Theta}_d^{CF}} \left(\hat{\phi}_{\Theta_l,-k}(O_i) - \hat{\Theta}_l^{CF} - \hat{\Psi}_l(\hat{\phi}_{\Theta_d}(O_i) - \hat{\Theta}_d^{CF}) \right) \right]^2 .$$

Lemma 1 in the next section gives that the variance estimator above is consistent, and a confidence interval of $\hat{\Psi}_l^{CF}$ is then constructed as $\hat{\Psi}_l^{CF} \pm 1.96\sqrt{\hat{\sigma}^{2,CF}_{\Psi_l}/n}$

4.1.3 On estimation of logit transformation of Ψ_l

The target parameter Ψ_l is restricted to $[0,1)$, but the estimator $\hat{\Psi}_l^{CF}$ is unrestricted, which in practice can result in parameter estimates that are outside the range $[0,1)$, or confidence intervals that contain either 0 or 1. To combat this issue, we construct a cross-fitted one-step estimator of the transformed parameter $\operatorname{logit}(\Psi_l)$. We note that given initial estimators, $\hat{\Theta}_l^{CF}$, $\hat{\Theta}_d^{CF}$, $\hat{\Theta}_l^0$, $\hat{\Theta}_d^0$, where $\hat{\Theta}_l^0$ and $\hat{\Theta}_d^0$ are plug-in estimators, the construction is directly given in Hines, Diaz-Ordaz, and Vansteelandt (2022) and in Theorem 4 in their Appendix, they give additional conditions on the plug-in estimators under which the estimator of $\operatorname{logit}(\Psi_l)$ is asymptotically linear. Hence, we will only sketch the construction, and refer to Hines, Diaz-Ordaz, and Vansteelandt (2022) for a derivation of the asymptotic results.

Define the transformed target parameter $\zeta_l(P) \equiv \operatorname{logit}(\Psi_l(P))$. The efficient influence function of $\zeta(P)$ is given by

$$\tilde{\psi}_{\zeta_l} = \frac{\tilde{\psi}_{\Psi_l}}{\Psi_l(1-\Psi_l)}$$

by the chain rule. Given the plug-in estimators, define $\hat{\Psi}_l^0 = \frac{\hat{\Theta}_l^0}{\hat{\Theta}_d^0}$. The cross-fitted one-step estimator is then given by

$$\hat{\zeta}_l^{CF} = \sum_{k=1}^{K} \frac{n_k}{n} \hat{\zeta}_{l,k}$$

where

$$\hat{\zeta}_{l,k} = \operatorname{logit}(\hat{\Psi}^0_{l,-k}) + \mathbb{P}_n^k \hat{\tilde{\psi}}_{\zeta_l,-k} = \operatorname{logit}(\hat{\Psi}^0_{l,-k}) + \mathbb{P}_n^k \frac{\hat{\tilde{\psi}}_{\Psi_l,-k}}{\hat{\Psi}^0_{l,-k}(1-\hat{\Psi}^0_{l,-k})}$$

with $\hat{\Psi}^0_{l,-k}$ being the plug-in estimator obtained from the sample $\mathcal{V}_{-k}$ and $\hat{\tilde{\psi}}_{\Psi_l,-k}$ being the estimator of the EIF $\tilde{\psi}_{\Psi_l}$ derived from nuisance estimators obtained from $\mathcal{V}_{-k}$. As noted in Hines, Diaz-Ordaz, and Vansteelandt (2022), the estimator, $\hat{\zeta}_l^{CF}$, can be written in terms of the already established estimators with

$$\hat{\zeta}_{l,k} = \operatorname{logit}(\hat{\Psi}^0_{l,-k}) + \frac{\hat{\Theta}_{d,-k}}{\hat{\Theta}^0_{d,-k}} \frac{(\hat{\Psi}_{l,-k} - \hat{\Psi}^0_{l,-k})}{\hat{\Psi}^0_{l,-k}(1-\hat{\Psi}^0_{l,-k})},$$

where $\hat{\Theta}_{d,-k}$ and $\hat{\Psi}_{l,-k}$ are the one-step estimators obtained from $\mathcal{V}_{-k}$.

Under the same assumptions as in Corollary 2, Theorem 4 in Hines, Diaz-Ordaz, and Vansteelandt (2022) gives that $\hat{\zeta}_l^{CF}$ is asymptotically linear with $\tilde{\psi}_{\zeta_l}$ as its influence function. In practice, this can be leverage to obtain an estimate of Ψ_l, where the estimate and the confidence interval are restricted to the interval $[0, 1)$, by an expit-transformation of $\hat{\zeta}_l^{CF}$ and the corresponding confidence interval. By the delta method, the back-transformed estimator then shares the same asymptotic properties as given in Corollary 2.

4.2 Estimation of Ω_j

4.2.1 Efficient influence function

We derive the EIF's of Γ_j and χ_j separately from which the EIF of Ω_j is obtained. This is summarized in the following theorem.

Theorem 3. *Let φ be given as in (4) for the survival function setting, and as in (5) for the RMST setting. Then the EIF's of Γ_j and χ_j are given by $\tilde{\psi}_{\Gamma_j}$ and $\tilde{\psi}_{\chi_j}$, respectively, where*

$$\tilde{\psi}_{\Gamma_j} = [\varphi(O) - \tau_j(X)][X_j - E(X_j \mid X_{-j})] - \Gamma_j \tag{9}$$

and

$$\tilde{\psi}_{\chi_j} = [X_j - E(X_j \mid X_{-j})]^2 - \chi_j. \tag{10}$$

The EIF of Ω_j is given by

$$\tilde{\psi}_{\Omega_j} = \frac{1}{\chi_j}\left(\tilde{\psi}_{\Gamma_j} - \Omega_j \tilde{\psi}_{\chi_j}\right). \tag{11}$$

Remark 2. The EIF of Γ_j is stated in terms the φ, which is the uncentered EIF of the ATE. Thus, the above EIF's can readily be extended to other data settings with different CATE functions, τ. From the calculations in the Appendix, it is seen that the EIF's in Theorem 3 hold as long as the Gateaux derivative of $\tau(x)$ can be expressed as $\frac{\mathbb{1}(X=x)}{f(x)}g(o)$, for some function $g = \varphi - \tau$.

4.2.2 Cross-fitted one-step estimator

As with Ψ_l, we construct cross-fitted one-step estimators for Γ_j and χ_j based on the EIF's in Theorem 3 where we let $\phi_{\Gamma_j} = [\varphi(O) - \tau_j(X)][X_j - \mathrm{E}(X_j \mid X_{-j})]$ and $\phi_{\chi_j} = [X_j - \mathrm{E}(X_j \mid X_{-j})]^2$ denote the uncentered EIF's. Then using the same sample splitting notation as in the construction of $\hat{\Theta}_l^{CF}$ we denote $\hat{E}_{n,-k}^j$ the regression of X_j onto X_{-j} in the sample $\mathcal{V}_{-k}$ and define the estimators

$$\hat{\Gamma}_j^{CF} = \sum_{i=k}^{K} \frac{n_k}{n} \mathbb{P}_n^k \hat{\phi}_{\Gamma_j,-k} = \frac{1}{n} \sum_{k=1}^{K} \sum_{i \in \mathcal{T}_k} \left\{ [\hat{\varphi}_{-k}(O_i) - \hat{\tau}_{j,-k}(X)][X_{i,j} - \hat{E}_{n,-k}^j(X_{i,-j})] \right\}$$

and

$$\hat{\chi}_j^{CF} = \sum_{i=k}^{K} \frac{n_k}{n} \mathbb{P}_n^k \hat{\phi}_{\chi_j,-k} = \frac{1}{n} \sum_{k=1}^{K} \sum_{i \in \mathcal{T}_k} [X_{i,j} - \hat{E}_{n,-k}^j(X_{i,-j})]^2.$$

The two estimators are combined to create an estimator for Ω_j:

$$\hat{\Omega}_j^{CF} = \frac{\hat{\Gamma}_j^{CF}}{\hat{\chi}_j^{CF}}.$$

To state results on the asymptotic distribution of the above estimators, we need a slight modification of assumption A as follows:

Assumption B.

B1 $\left\| \hat{E}_n^j - E(\cdot \mid X_{-j}) \right\|_{L_2(P)} = o_p(n^{-\frac{1}{4}})$.

B2 $(X_j - \hat{E}_n^j(X_{-j}))^2 \leq \delta < \infty, \quad \delta > 0, \quad a.s.$

B3 $var(X_j \mid X_{-j}) < \infty$ *for all* $j \in \{1, \dots d\}$.

Assumption B1 relates to the convergence rate of $\hat{E}_n^j$, and it is similar to assumption A2 and A3. Assumption B2 is a technical assumption, which we would expect to hold for most reasonable estimators $\hat{E}_n^j$, and assumption B3 assumes all the conditional distributions of X_j given X_{-j} to have second moment.

Theorem 4. *Under assumption A with A2 replaced by assumption B, $\hat{\Gamma}_j^{CF}$ is asymptotically linear with influence function given by the EIF (9), and hence locally asymptotically efficient. Thus*

$$\sqrt{n}(\hat{\Gamma}_j^{CF} - \Gamma_j) \xrightarrow{D} \mathcal{N}(0, P\tilde{\psi}_{\Gamma_j}^2).$$

Furthermore, under assumption B, $\hat{\chi}_j^{CF}$ is asymptotically linear with influence function given by the EIF (10), and hence locally asymptotically efficient. Furthermore:

$$\sqrt{n}(\hat{\chi}_j^{CF} - \chi_j) \xrightarrow{D} \mathcal{N}(0, P\tilde{\psi}_{\chi_j}^2).$$

Proof. See Section C in the Appendix. □

And finally, a simple application of the delta method gives the main result for our estimator $\hat{\Omega}_j^{CF}$.

Corollary 3. *Under assumption A with A2 replaced by assumption B, $\hat{\Omega}_j^{CF}$ is asymptotically linear with influence function given by the EIF (11), and hence locally asymptotically efficient. Furthermore:*

$$\sqrt{n}(\hat{\Omega}_j^{CF} - \Omega) \xrightarrow{D} \mathcal{N}(0, P\tilde{\psi}_{\Omega_j}^2). \tag{12}$$

To estimate the variance $P\tilde{\psi}_{\Omega_j}^2$, let $\hat{\tilde{\psi}}_{\Omega_j,-k}$ be the estimate of $\tilde{\psi}_{\Omega_j}$ in the sample $\mathcal{V}_{-k}$. We define the cross-fitted plug-in variance estimator as

$$\hat{\sigma}_{\Omega_j}^{2,CF} = \sum_{i=k}^{K} \frac{n_k}{n} \mathbb{P}_n^k \hat{\tilde{\psi}}_{\Omega_j,-k}^2 = \frac{1}{n} \sum_{k=1}^{K} \sum_{i \in \mathcal{T}_k} \left[\frac{1}{\hat{\chi}_j^{CF}} \left(\hat{\phi}_{\Gamma_j,-k}(O_i) - \hat{\Gamma}_j^{CF} - \hat{\Omega}_j(\hat{\phi}_{\chi_j,-k}(O_i) - \hat{\chi}_j^{CF}) \right) \right]^2.$$

The variance estimator can be used to calculate the standard error of $\hat{\Omega}_j^{CF}$, and the following lemma gives the consistency of the cross-fitted variance estimators considered in this article.

Lemma 1. *Let ψ_1 and ψ_2 be two pathwise differentiable maps from $\mathcal{M}$ to the reals with EIF's given by $\tilde{\psi}_1$ and $\tilde{\psi}_2$, respectively, where $\tilde{\psi}_i(\psi_i, \nu_i) = \varphi_i(\nu_i) - \psi_i$, $i = 1, 2$, for some nuisance* *parameters ν_i. Let $\hat{\psi}_i^{CF}$ denote the cross-fitted one-step estimator for ψ_i and assume that $\|\varphi(\hat{\nu}_{i,-k}) - \varphi(\nu_i)\| = o_p(1)$ for each k and i. Furthermore, we assume that*

$$\hat{\psi}_i^{CF} - \psi_i = \mathbb{P}_n\tilde{\psi}_i + o_p(n^{-1/2}), \quad i = 1, 2.$$

Let $\Psi = \frac{\psi_1}{\psi_2}$ and denote the cross-fitted estimator $\hat{\Psi}^{CF} = \frac{\hat{\psi}_1^{CF}}{\hat{\psi}_2^{CF}}$. Define

$$\tilde{\psi}(\psi_1, \psi_2, \nu_1, \nu_2) = \frac{1}{\psi_2}\left(\varphi_1(\nu_1) - \psi_1 - \frac{\psi_1}{\psi_2}(\varphi_2(\nu_2) - \psi_2)\right).$$

Then

$$\hat{\Psi}^{CF} - \Psi = \mathbb{P}_n\tilde{\psi}(\psi_1, \psi_2, \nu_1, \nu_2) + o_p(n^{-1/2}) \tag{13}$$

and we have the following consistency results for the cross-fitted variance estimators:

$$\hat{\sigma}_{\psi_i}^{2,CF} = \sum_{k=1}^{K}\frac{n_k}{n}\mathbb{P}_n^k\tilde{\psi}_i(\hat{\psi}_i^{CF}, \hat{\nu}_{i,-k})^2 \xrightarrow{P} P\tilde{\psi}_i(\psi_i, \nu_i)^2 \tag{14}$$

$$\hat{\sigma}_{\Psi}^{2,CF} = \sum_{k=1}^{K}\frac{n_k}{n}\mathbb{P}_n^k\tilde{\psi}(\hat{\psi}_1^{CF}, \hat{\psi}_2^{CF}, \hat{\nu}_{1,-k}, \hat{\nu}_{2,-k})^2 \xrightarrow{P} P\tilde{\psi}(\psi_1, \psi_2, \nu_1, \nu_2)^2. \tag{15}$$

Proof. See Appendix C.3. □

As mentioned in Section 3, the target parameter Ω_j is scale sensitive, and rather than comparing the magnitude of Ω_j across different j's, the variable importance is based on a test for the hypothesis $H : \Omega_j = 0$. Using lemma 1, we have the following result:

Corollary 4. *Under the same setup as in Corollary 3, we have under the null-hypothesis, $H_0 : \Omega_j = 0$, that*

$$\frac{\hat{\Omega}_j^{CF}}{\sqrt{\hat{\sigma}_{\Omega_j}^{2,CF}/n}} \xrightarrow{D} \mathcal{N}(0, 1).$$

Proof. Under the H_0, Corollary 3 gives that

$$\sqrt{n}\hat{\Omega}_j^{l,CF} \xrightarrow{D} \mathcal{N}(0, P\tilde{\psi}_{\Omega_j}^2)$$

and since

$$\sqrt{\hat{\sigma}_{\Omega_j^l}^{2,CF}} \xrightarrow{P} \sqrt{P\tilde{\psi}_{\Omega_j^l}^2}$$

by lemma 1 and the continuous mapping theorem, an application of Slutsky's theorem followed by the delta method gives the result. □

4.3 Choice of nuisance parameter estimators

In construction of $\hat{\Psi}_l^{CF}$ and $\hat{\Omega}_j^{CF}$ we need estimators of the nuisance parameters $\Lambda, \Lambda_c, \pi, \tau_l$. For estimation of $\hat{\Psi}_l^{CF}$ we further need an estimator of τ_d, and for estimation of $\hat{\Omega}_j^{CF}$ we further need an estimator $\hat{E}_n^j$. Estimators of $\hat{\Lambda}, \hat{\Lambda}_c, \hat{\pi}, \hat{E}_n^j$ can be chosen by any machine learning methods of choice, where we will use $\hat{\Lambda}$ to construct $\hat{\tau}(x) = e^{-\hat{\Lambda}(t|,1,x)} - e^{-\hat{\Lambda}(t|,0,x)}$. The estimator $\hat{\tau}$ uses $\hat{\Lambda}$, estimated from the entire data, to predict $S(t \mid a, x)$ for $a = 0, 1$ and it is termed the S-learner in the literature. Other methods of estimating $\hat{\tau}$ from initial estimates of Λ of π are possible and an overview of such meta-learners are given in Xu et al. (2023).

For estimation of τ_l, we consider a plug-in estimator utilizing the definition of the parameter:

$$\hat{\tau}_l(x) = \hat{E}_n(\hat{\tau}(X) \mid X_{-l} = x_{-l})$$

where $\hat{E}_n$ is a regression of the predicted $\hat{\tau}(X)$'s onto X_{-l}. Another possibility is to use the meta-learner given by $\hat{\tau}_l(x) = \hat{E}_n(\hat{\varphi}(X) \mid X_{-l} = x_{-l})$, since $\tau_l(x) = \mathrm{E}(\varphi(X) \mid X_{-l} = x_{-l})$. This meta-learner is a version of the DR-learner (Kennedy, 2022b), if $\hat{E}_n$ and $\hat{\varphi}$ are estimated on different samples. Extending the analysis in Kennedy (2022b) to a survival setting might provide theoretical guarantees of $\hat{\tau}_l$ based on the DR-learner, as opposed to the plug-in estimator, but in testing we found that the plug-in estimator performed better. This is in line with the recommendations given in Hines, Diaz-Ordaz, and Vansteelandt, 2022.

Estimation of $\tau_d = \mathrm{E}(\tau(X))$ is a well-studied problem in the survival setting, and it is thus possible to construct an estimator $\hat{\tau}_d$ with further theoretical guarantees compared to $\hat{\tau}_l$. For a thorough analysis of τ_d-estimation see Rytgaard et al. (2023) and Westling et al. (2023). In these articles they construct estimators based on the EIF (3) and derive properties under which such estimators are asymptotically linear with influence function given by the EIF. We will not go into much detail here, but we briefly summarize the construction given in Westling et al. (2023):

For estimators $\hat{\Lambda}, \hat{\Lambda}_c, \hat{\pi}$ following assumption A, we construct the following cross-fitted estimator

$$\hat{\tau}_d^{CF} = \sum_{i=k}^{K} \frac{n_k}{n} \mathbb{P}_n^k \hat{\varphi}_{-k} = \frac{1}{n} \sum_{k=1}^{K} \sum_{i \in \mathcal{T}_k} \hat{\varphi}_{-k}(O_i).$$

Theorem 3 in Westling et al., 2023 gives that $\hat{\tau}_d^{CF}$ is asymptotically linear with influence function given by $\tilde{\psi}_{\tau_d} = \varphi - \tau_d$. Now, let $X_n = \sqrt{n}(\hat{\tau}_d^{CF} - \tau_d)$. Then $X_n \rightsquigarrow X$ with $X \sim \mathcal{N}(0, P\tilde{\psi}_{\tau_d}^2)$ and Prohorov's theorem (van der Vaart, 2000 theorem 2.4) gives $\|X_n\| = O_p(1)$. Hence, $\left\|\hat{\tau}_d^{CF} - \tau_d\right\| = n^{-1/2} \|X_n\| = o_p(n^{-1/2})$, and corollary 2 applies.

Notice, however, that the estimation of $\hat{\Theta}_d^{CF}$ requires estimation of $\hat{\tau}_{d,-k}$ for each split k. Thus, in order to use the convergence rate results above, we need to perform a nested type of cross-fitting, such that $\hat{\tau}_{d,-k}$ is the cross-fitted estimator above but estimated using data in $\mathcal{V}_{-k}$ instead of the entire data. For estimation of Θ_d we have the following procedure:

1. Split the data uniformly at random into K_1 subsamples $\mathcal{V}_k$, $k = 1, \ldots, K_1$, such that $\mathcal{O} = \dot{\cup}_{k=1}^{K_1} \mathcal{V}_k$.

2. for each k estimate $\hat{\Lambda}, \hat{\Lambda}_c, \hat{\pi}$ using data in $\mathcal{V}_{-k}$ and obtain $\hat{\tau}_{-k}$ and $\hat{\varphi}_{-k}$. For $\tau_{d,-k}$-estimation:
   (a) Split $\mathcal{V}_{-k}$ into K_2 subsamples $\mathcal{V}_i^k$, $i = 1, \ldots, K_2$, such that $\mathcal{V}_{-k} = \dot{\cup}_{i=1}^{K_2} \mathcal{V}_i^k$.
   (b) For each $i = 1, \ldots, K_2$ estimate $\hat{\Lambda}, \hat{\Lambda}_c, \hat{\pi}$ using data in $\mathcal{V}_{-i}^k$ and obtain $\hat{\varphi}_{-i}^k$.
   (c) Obtain the $k'th$ estimate $\hat{\tau}_{d,-k}^{CF} = \frac{1}{n_{-k}} \sum_{i=1}^{K_2} \sum_{O_j \in \mathcal{V}_i^k} \hat{\varphi}_{-i}^k(O_j)$, where n_{-k} is the number of observations in $\mathcal{V}_{-k}$.
3. Obtain the estimate

$$\hat{\Theta}_d^{CF} = \frac{1}{n} \sum_{k=1}^{K} \sum_{i \in \mathcal{T}_k} \left\{ (\hat{\varphi}_{-k}(O_i) - \hat{\tau}_{d,-k}^{CF})^2 - (\hat{\varphi}_{-k}(O_i) - \hat{\tau}_{-k}(X_i))^2 \right\}$$

Using this estimating scheme, the convergence rate assumption on $\hat{\tau}_d^{CF}$ is automatically fulfilled by assumption A and corollary 2 applies without further restrictions on $\hat{\tau}_d$.

5 Simulation Study

5.1 Setup

We performed a simulation study to investigate the finite sample properties of the estimators of Ψ_l and Ω_j with and without cross-fitting in the survival setting. We generated data from the following models:

$$\begin{aligned} \Lambda(t \mid A, X) &= 2t^{0.0025} \exp(-x_1 - X_2 - 0.3X_3 + 0.1X_4 - A(2 - 0.5X_1 - 0.3X_2)) \\ \Lambda_c(t \mid A, X) &= 2t^{0.00025} \exp(-0.3X_1) \\ \pi(1 \mid X) &= \frac{1}{1 + exp(-0.3X_1 - 0.3X_2)}, \end{aligned}$$

where $X_j,\ j = 1, \ldots, 4$, are i.i.d standard normally distributed. Note that Λ and Λ_c follow Cox-models with baseline hazards given by Weibull hazards. The time-horizon is chosen as $t = 10$ with the true values Ψ_1 and Ω_1 approximately 0.6907 and -0.1518, respectively, in the survival setting.

The nuisance parameter estimators are chosen in different combinations listed below and the performance of the target parameter estimators is compared between the different choices of nuisance parameter estimators. In the following, we use a naming convention of the nuisance choices in the form ***A-B*** where ***A*** corresponds to the choice of $\hat{\Lambda}$, $\hat{\Lambda}_c$ and $\hat{\pi}$, and ***B*** corresponds to $\hat{E}_n$ (and $\hat{E}_n^1$ in for Ω_1-estimation). When generalized additive models (GAMs) are used, it will be through the implementation given in the R-package `mgcv` with smoothing function given by the default setting, thin plate regression splines. When random forests are used it will be through the implementation given in the R-package `rfsrc` with hyperparameters given by the default settings (which change based on the outcome type, see Ishwaran et al., 2023).

correct-GAM $\hat{\Lambda}$ and $\hat{\Lambda}_c$ are given by Breslow estimates based on correctly specified Cox models, the propensity score is estimated by a correctly specified GLM and $\hat{E}_n$ (and $\hat{E}_n^1$) is given by a GAM including all interactions.

correct-RF $\hat{\Lambda}$ and $\hat{\Lambda}_c$ are given by Breslow estimates based on correctly specified Cox models, the propensity score is estimated by a correctly specified GLM and $\hat{E}_n$ (and $\hat{E}_n^1$) is estimated by a random forest.

RF-RF $\hat{\Lambda}$, $\hat{\Lambda}_c$, $\hat{\pi}$ and $\hat{E}_n$ (and $\hat{E}_n^1$) are all estimated by random (survival) forests.

RF-GAM $\hat{\Lambda}$, $\hat{\Lambda}_c$ and $\hat{\pi}$ are estimated by random (survival) forest and $\hat{E}_n$ (and $\hat{E}_n^1$) is estimated by a GAM.

Each nuisance setting is used to estimate $\hat{\Psi}_1$, $\hat{\Psi}_1^{CF}$, $\hat{\Omega}_1$ and $\hat{\Omega}_1^{CF}$, respectively. We used $K = 10$ folds for the cross-fitted estimators. To separate the cross-fitted and non-cross-fitted estimators we extend the nuisance parameter naming with a suffix ***C*** such that, e.g., ***RF-RF-CF*** corresponds to the estimate $\hat{\Psi}_l^{CF}$ or $\hat{\Omega}_1^{CF}$ using the nuisance estimators ***RF-RF***. For estimation of $\hat{\Psi}_1$, we consider four different sample sizes, $n = 1000, 2000, 3000, 4000$, and for each setting we run $N = 1000$ simulations and calculate the corresponding estimators for each simulation. For estimation of $\hat{\Omega}_1$ we consider four different sample sizes, $n = 250, 500, 750, 1000$. The results are presented in the next subsections.

5.2 Results for Ψ_1

5.2.1 Correctly specified λ, Λ_c and π

In Figure 1a, we see the sampling distributing of the estimators of Ψ_1 under the different choices for the nuisance estimator $\hat{E}_n$ and with and without cross-fitting. All other nuisances estimators are correctly specified according to ***correct-GAM*** and ***correct-RF*** in the previous subsection. From Corollary 2 we know that the estimators based on cross-fitting should be unbiased and asymptotically normal, even when using flexible nuisance estimators, which can not be guaranteed for the corresponding estimators without cross-fitting. Indeed we see that estimators based on GAM seem to follow a normal distribution around the true value, whereas the estimator based on RF is severely biased without cross-fitting but much less so with. The results of the simulations are presented in Table 1a. When GAM is used to estimate $\hat{E}_n$, the estimator seems to perform satisfactory according to Corollary 2 both with and without cross-fitting. When RF is used for $\hat{E}_n$, we get a huge bias without sampling splitting, as we saw in Figure 1a, but when cross-fitting is used, there still seem to be some non-vanishing bias inherent from the RF-estimation. From the standard error of the simulations corresponding to ***correct-RF-CF***, it looks as if the estimator is converging on $\sqrt{n}$-rate, which, with the non-vanishing bias, results coverage decreasing with sample size. Since only $\hat{\tau}_l$ is based on random forest, with all other nuisance estimators being based on correctly specified parametric models, this result suggests that assumption A2 is not fulfilled for $\hat{\tau}_l$, which again can possibly be attributed to the choice of hyperparameters used in the random forest.

5.2.2 λ, Λ_c and π estimated by Random Forest

In figure 1b we see the sampling distribution of the estimators for Ψ_l, where the nuisance parameters λ, Λ_c and π are all estimated flexibly via **RF**. The estimators without cross-fitting are seen to be severely biased as was the case with correctly specified λ, Λ_c and π. Table 1b summarizes the results for the estimators using **RF**. Using **RF** to estimate $\hat{E}_n$ is seen to introduce some non-vanishing bias, as in 1a, resulting in decreasing coverage. In the case of **GAM**-estimation for $\hat{E}_n$, cross-fitting is able correctly de-bias the estimator giving approximately nominal coverage.

5.3 Results for Ω_1

5.3.1 Correctly specified λ, Λ_c and π

Figure 2a presents the sample distribution of the estimators of Ω_1 with correctly specified Λ, Λ_c and π for different choices for estimation of $\hat{E}_n$ and $\hat{E}_n^1$, with and without cross-fitting. Compared to Ψ_1-estimation, there is no severe shift in sample distribution between the estimators. The estimator with **RF** and no cross-fitting is seen to have a slightly wider distribution than the others, for which there is no noticeable difference. Table 2a gives the results of the simulations, where the **RF** without cross-fitting is seen to generally have a slightly larger bias and MSE than the others. Remarkably, compared to the Ψ_1-estimation, far fewer observations are needed for reliable estimation of Ω_1. One thing to note is that Corollary 3 is only guarantied to hold for the cross-fitted estimator, but since the estimator is given as the ratio of two other estimators, it can happen that the bias introduced by employing **RF** without cross-fitting roughly cancels in the ratio, which would explain why the **RF** without cross-fitting is seen to perform reliably with approximately nominal coverage. Indeed this is the case for the simulation study conducted here, as seen in Figure 3a and 4a in the Appendix, where the sample distribution of the estimators of Γ_1 and χ_1 are shown. Hence, we can not recommend **RF** without cross-fitting, as the cancellation of biases in the ratio of Γ_1 and χ_1 are unlikely to occur generally.

5.3.2 λ, Λ_c and π estimated by Random Forest

Figure 2b shows the sample distribution of the estimators of Ω_1 when RF is used for estimation of λ, Λ_c and π. Here, the difference between the cross-fitted and non-cross-fitted estimators are more noticeable. Interestingly, using cross-fitting seem to produce similar distributions, regardless of whether **RF** og **GAM** was used for estimation of $\hat{E}_n$ and $\hat{E}_n^1$. Table 2b presents the results of the simulation study. Generally, the bias seem to vanish with the sample size (again, in the case of **RF-RF**, this might be a coincidence), but the coverage for the non-cross-fitted estimators are far off, whereas cross-fitting seem to provide approximately nominal coverage, even in relatively small samples.

6 Application to HIV data set

We apply the methods described in the previous sections to the AIDS Clinical Trial Group Study 175 (Hammer et al., 1996). The data can be found in the `R`-package `ACTG175` and consists 2139 HIV patients who were randomized to one of four treatments: (1) zidovudine (ZDV)(n=532), (2) zidovudine + didanosine (ZDV+ddI)(n=522), (3) zidovudine + zalcitabine (ZDV+ZAL)(n=524), and (4) didanosine(n=561). Patient were followed from treatment initiation until an event consisting of a decline in CD4 cell count greater than 50%, disease progression to AIDS, or death, or end-of-follow-up. In line with Cui et al. (2023), we define the treatment effect (comparing two treatments) to be given by the RMST at 1000 days after treatment initiation, and we consider 12 baseline covariates for which we will analyse the possible treatment effect heterogeneity explained by each of them. The covariates consist

n	method	bias Ψ_1	coverage	SD	mean SE	MSE
1000	correct-GAM	-0.0161	0.9620	0.0963	0.0958	0.0095
2000		-0.0120	0.9620	0.0640	0.0659	0.0042
3000		-0.0068	0.9340	0.0576	0.0535	0.0034
4000		-0.0048	0.9320	0.0478	0.0464	0.0023
1000	correct-GAM-CF	0.0230	0.9530	0.1030	0.1031	0.0111
2000		0.0107	0.9620	0.0657	0.0681	0.0044
3000		0.0094	0.9260	0.0588	0.0548	0.0035
4000		0.0082	0.9360	0.0484	0.0472	0.0024
1000	correct-RF	-0.3349	0.0030	0.0759	0.0805	0.1179
2000		-0.3382	0.0000	0.0526	0.0548	0.1172
3000		-0.3364	0.0000	0.0422	0.0439	0.1149
4000		-0.3366	0.0000	0.0369	0.0380	0.1146
1000	correct-RF-CF	0.0483	0.9200	0.1068	0.1066	0.0137
2000		0.0381	0.9240	0.0698	0.0716	0.0063
3000		0.0381	0.8840	0.0613	0.0577	0.0052
4000		0.0371	0.8700	0.0518	0.0498	0.0041

(a) Correctly specified Λ, Λ_c and π

n	method	bias Ψ_1	coverage	SD	mean SE	MSE
1000	RF-RF	-0.2438	0.0000	0.0384	0.0332	0.0609
2000		-0.2429	0.0000	0.0256	0.0235	0.0597
3000		-0.2404	0.0000	0.0220	0.0191	0.0583
4000		-0.2377	0.0000	0.0189	0.0165	0.0568
1000	RF-RF-CF	0.0469	0.9140	0.1575	0.1492	0.0270
2000		0.0353	0.9340	0.0967	0.0976	0.0106
3000		0.0304	0.9230	0.0780	0.0765	0.0070
4000		0.0318	0.9050	0.0661	0.0649	0.0054
1000	RF-GAM	0.0503	0.6360	0.0698	0.0453	0.0074
2000		0.0719	0.4260	0.0453	0.0324	0.0072
3000		0.0826	0.2310	0.0391	0.0263	0.0083
4000		0.0901	0.0980	0.0328	0.0227	0.0092
1000	RF-GAM-CF	0.0087	0.9460	0.1601	0.1521	0.0257
2000		-0.0051	0.9500	0.0883	0.0945	0.0078
3000		-0.0088	0.9440	0.0726	0.0739	0.0053
4000		-0.0097	0.9560	0.0601	0.0623	0.0037

(b) Flexible estimation of Λ, Λ_c and π

Table 1: Results of 1000 simulations of $\hat{\Psi}_1$ in the survival function setting with varying nuisance estimators, with and without cross-fitting, across sample sizes $n = 1000, 2000, 3000, 4000$. The abbreviation of the methods should be read as follows: A-B-C, where A corresponds to the nuisance estimators Λ, Λ_c and π, B corresponds to the nuisance estimator $\hat{E}_n$, and C corresponds to whether or not cross-fitting was used. Here, *correct* corresponds to correctly specified Cox and logistic regression, RF corresponds to Random Forest, and GAM corresponds to a generalized additive model. The tables shows the bias, coverage, empirical standard deviation (SD), mean estimated standard error (mean SE), and the mean squared error (MSE).

n	method	bias Ω_1	coverage	SD	mean SE	MSE
250	correct-GAM	- 0.0011	0.934	0.0422	0.0408	0.0018
500		0.0011	0.945	0.0291	0.0283	0.0009
750		0.0009	0.952	0.0230	0.0231	0.0005
1000		0.0021	0.947	0.0203	0.0201	0.0004
250	correct-GAM-CF	-0.0004	0.932	0.0438	0.0419	0.0019
500		0.0006	0.948	0.0291	0.0287	0.0008
750		0.0001	0.950	0.0231	0.0233	0.0005
1000		0.0012	0.948	0.0204	0.0202	0.0004
250	correct-RF	-0.0041	0.946	0.0532	0.0544	0.0028
500		-0.0015	0.953	0.0378	0.0387	0.0014
750		-0.0020	0.955	0.0299	0.0317	0.0009
1000		0.0003	0.956	0.0265	0.0277	0.0007
250	correct-RF-CF	-0.0012	0.936	0.0428	0.0414	0.0018
500		0.0006	0.945	0.0291	0.0283	0.0008
750		-0.0001	0.946	0.0229	0.0229	0.0005
1000		0.0015	0.947	0.0202	0.0199	0.0004

(a) Correctly specified Λ, Λ_c and π

n	method	bias Ω_1	coverage	SD	mean SE	MSE
250	RF-RF	0.0090	0.915	0.0363	0.0324	0.0014
500		0.0031	0.904	0.0265	0.0224	0.0007
750		0.0014	0.891	0.0217	0.0183	0.0005
1000		0.0013	0.897	0.0190	0.0158	0.0004
250	RF-RF-CF	-0.0127	0.935	0.0548	0.0539	0.0032
500		-0.0047	0.946	0.0353	0.0355	0.0013
750		-0.0027	0.948	0.0287	0.0285	0.0008
1000		-0.0001	0.952	0.0242	0.0246	0.0006
250	RF-GAM	0.0244	0.740	0.0331	0.0245	0.0017
500		0.0135	0.747	0.0249	0.0168	0.0008
750		0.0092	0.775	0.0205	0.0137	0.0005
1000		0.0074	0.759	0.0181	0.0118	0.0004
250	RF-GAM-CF	-0.0129	0.941	0.0558	0.0545	0.0033
500		-0.0060	0.949	0.0358	0.0365	0.0013
750		-0.0032	0.952	0.0285	0.0293	0.0008
1000		-0.0011	0.961	0.0241	0.0253	0.0006

(b) Flexible estimation of Λ, Λ_c and π

Table 2: Results of 1000 simulations of $\hat{\Omega}_1$ in the survival function setting with varying nuisance estimators, with and without cross-fitting, across sample sizes $n = 250, 500, 750, 1000$. The abbreveations of the methods are read as follows: A-B-C, where A corresponds to the nuisance estimators Λ, Λ_c and π, B corresponds to the nuisance estimators $\hat{E}_n$ and $\hat{E}_n^j$, and C corresponds to whether or not cross-fitting was used. Here, *correct* corresponds to correctly specified Cox and logistic regression, RF corresponds to Random Forest, and GAM corresponds to a generalized additive model. The tables shows the bias, coverage, empirical standard deviation (SD), mean estimated standard error (mean SE), and the mean squared error (MSE).

of 5 continuous variables, age, CD4 cell count, CD8 cell count, weight (kg), Karnofsky score, and 7 binary variables, gender, race, hemophilia, homosexual activity, antivetroviral history, symptomatic status, intravenous drug use history.

As the aim of the study was to compare monotherapy (ZDV or ddI) with combination therapy (ZDV+ddI or ZDV + ZAL), we consider two comparisons: ddI vs ZDV+ddI and ZDV vs ZDV+ZAL. We applied the cross-fitted TE-VIM and the best partially linear projection estimators, with $K = 10$, described in Section 4, with all nuisance parameters estimated by Random Forests as implemented in the R-package RandomForestSRC. For the TE-VIM, we used the logit-transformed $\hat{\zeta}_l^{CF}$ with an expit-back-transformation to obtain Ψ_l-estimates and confidence intervals that respect the boundary $\Psi_l \in [0, 1)$.

In Table 3 we see the results of estimators applied to the comparison of ddI vs ZDV+ddI. For the TE-VIM (3a) we see that almost all of the covariates are given a TV-VIM measure of exactly or close to 1, but with confidence interval from 0 to 1. In contrast, the results for the best partially linear projection estimates give p-values ranging from 0.068 to 0.98. In both cases we conclude that we can not identify any treatment effect heterogeneity.

In table 4 we see the results based on the comparison of ZDV vs ZDV+ZAl. Here the TE-VIM estimates are ranging from 0.02 to 0.992 with cd8 having the largest estimate, but again, all with a confidence interval ranging from 0 to 1. The results based on the best partially linear projection give p-values in the range 0.007 to 0.953 with 3 significant p-values for cd8, karnof, and cd4, respectively. Both measures ranks CD8 cell count as the most "important" in terms of explaining treatment heterogeneity, but with the TE-VIM having a confidence interval of [0,1]. The results suggest that CD8 cell count, Karnofsky score and CD4 cell count are important in explaining the treatment effect heterogeneity of ZDV vs ZDV+ZAl on RMST at $t = 1000$ days after treatment initiation.

In comparing the results related to Ψ_l and Ω_l, respectively, we see the difference in sample sizes needed for providing meaningful estimates between the two measures, as indicated by the simulation study in Section 5. With confidence intervals of [0,1], the estimates of Ψ_l do not give any inside into the potential heterogeneity in the effect of the treatments considered here, whereas the estimates of Ω_j were able to find significant treatment effect modification for some of the covariates.

7 Discussion

In this paper, we have extended the treatment effect variable importance measures introduced by Hines, Diaz-Ordaz, and Vansteelandt, 2022 to a time-to-event setting allowing for censored data. We have constructed estimators for the TE-VIMs Θ_l and Ψ_l using two different CATE functions and given assumptions under which they are seen to be asymptotically normal and locally efficient. The assumptions require that the nuisance estimators $\hat{\tau}$ and $\hat{\tau}_l$ are both consistent at $n^{-1/4}$-rate, allowing for the use of machine-learning to estimate the nuisance parameters. In the simulation study we saw that the estimators without cross-fitting were heavily biased when using data adaptive nuisance estimators, such as random forest, but that the cross-fitting was mostly able to correct for the bias introduced by the flexible nuisance estimation. Importantly, it seems that the main challenge lies in choosing $\hat{E}_n$ appropriately, since using random forest (with default hyperparameters) was seen to introduce a non-vanishing bias, whereas using GAM for $\hat{E}_n$-estimation gave correct coverage even when other nuisance

covariate	Ψ_j	lower	upper
age	1.000	0.000	1.000
race	1.000	0.000	1.000
cd4	1.000	0.000	1.000
wtkg	1.000	0.000	1.000
hemo	1.000	0.000	1.000
gender	1.000	0.000	1.000
homo	1.000	0.000	1.000
symptom	0.997	0.000	1.000
drugs	0.994	0.000	1.000
karnof	0.993	0.000	1.000
cd8	0.814	0.007	1.000
str2	0.028	0.000	1.000

(a) Heterogeneity explained by Ψ_j

covariate	Ω_j	SE	p-value
wtkg	-2.383	1.308	0.068
cd8	0.045	0.028	0.109
homo	-67.846	42.469	0.110
drugs	55.792	36.719	0.129
gender	85.035	56.927	0.135
hemo	-75.294	58.462	0.198
age	2.157	1.754	0.219
cd4	-0.114	0.113	0.312
karnof	1.472	2.332	0.528
race	8.354	28.432	0.769
symptom	-2.486	40.178	0.951
str2	-0.652	26.378	0.980

(b) Heterogeneity explained by Ω_j

Table 3: Estimation of variable importance on the treatment effect of zidovudine + didanosine (ZDV+ddI) vs didanosine (ddI) on RMST. The data is from the study Hammer et al. (1996) and the outcome is time to an event consisting of a decline in CD4 cell count greater than 50%, disease progression to AIDS, or death. The treatment effect is defined as the difference in RMST at 1000 days after treatment initiation between ZDV+ddI and ZDZ. In table (a), the variable importance is estimated by the *expit*-transformation of $\hat{\zeta}_l^{CF}$, for l ranging over the single covariates, with corresponding confidence intervals. In table (b), the variable importance is estimated by $\hat{\Omega}_j^{CF}$, for j ranging over the single covariates.

parameters were estimated with random forest. One possible avenue to leverage the choice of RF-hyperparameters could be to replace the current cross-fitted one-step estimators with targeted-maximum-likelihood (TMLE). Li et al. (2023) constructed a TMLE for the TE-VIMs of Hines, Diaz-Ordaz, and Vansteelandt (2022), and though the remainder term still calls for initial $\hat{\tau}_l$ estimators that are consistent at $n^{-1/4}$ rate, the estimators are seen to have better finite sample performance compared to the one-step estimator. Thus, one may pursue a TMLE based on the EIF's derived in this paper with the same asymptotic properties as $\hat{\Theta}_l^{CF}$ under assumption A, to possibly achieve better finite sample performance. We leave this for future work.

Furthermore, we have derived a new variable importance measure based on the ideas from Vansteelandt and Dukes (2022a) as a best partially linear projection of the CATE-function. The estimand has the interpretation of the real parameter in a partially linear model of the CATE function, but it continues to serve as a measure of heteoregeneity when the model fails to hold. One consequence, though, is that it could happen that $\Omega_j = 0$ even when X_j explains some of the treatment effect, as seen by plugging $\beta = 0$ into

$$\tau(x) = \beta x_j + w(x_{-j}) + R(x_j, x_{-j}).$$

In contrast to the estimators for Ψ_l, the estimators of Ω_j was seen to perform well in relatively small sample sizes compared to the sample sizes needed for reliable estimation of Ψ_l, even when using Random Forest for all nuisance parameter estimation. This was also evident in the practical example, where the estimates of Ψ_l all had confidence intervals form 0 to 1,

covariate	Ψ_j	lower	upper
cd8	0.992	0.000	1.000
karnof	0.368	0.000	1.000
symptom	0.348	0.000	1.000
gender	0.197	0.001	0.982
str2	0.117	0.000	1.000
age	0.088	0.000	1.000
race	0.052	0.000	1.000
homo	0.037	0.000	1.000
hemo	0.029	0.000	1.000
wtkg	0.026	0.000	1.000
cd4	0.023	0.000	1.000
drugs	0.020	0.000	1.000

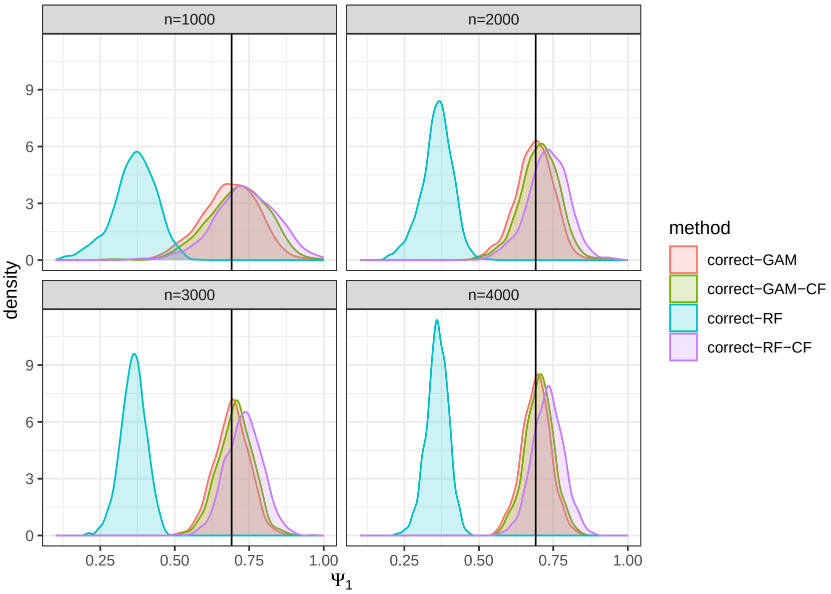
(a) Heterogeneity explained by Ψ_j

covariate	Ω_j	SE	p-value
cd80	0.093	0.035	0.007
karnof	6.446	3.024	0.033
cd40	-0.288	0.136	0.034
symptom	-55.853	47.467	0.239
drugs	48.365	50.376	0.337
wtkg	0.971	1.092	0.374
age	-0.985	1.702	0.563
race	-18.520	37.165	0.618
gender	24.795	54.105	0.647
homo	-19.473	52.341	0.710
hemo	5.463	77.456	0.944
str2	-1.719	29.241	0.953

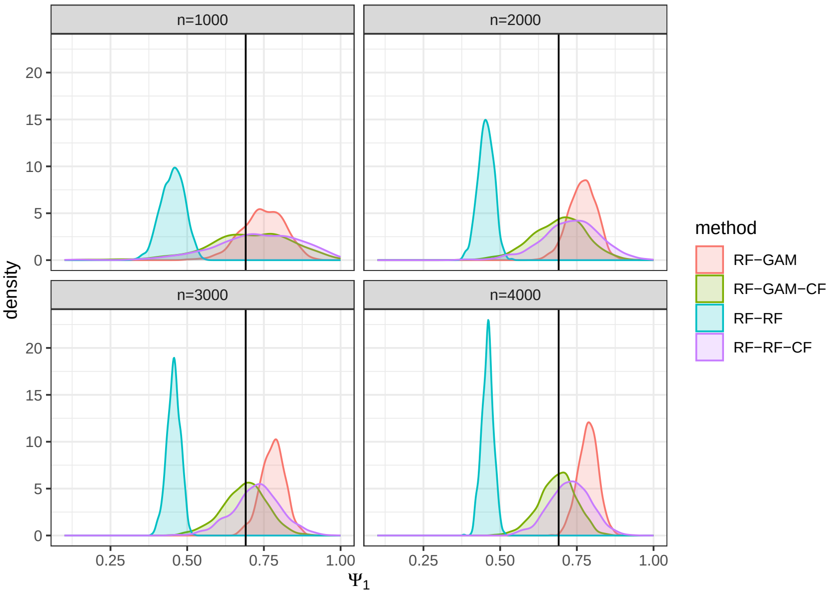
(b) Heterogeneity explained by Ω_j

Table 4: Heterogeneity in the effect of zidovudine + zalcitabine (ZDV+zal) vs zidovudine (ZDV) on RMST. Estimation of variable importance on the treatment effect of ZDV+zal vs ZDV on RMST. The data is from the study Hammer et al. (1996) and the outcome is time to an event consisting of a decline in CD4 cell count greater than 50%, disease progression to AIDS, or death. The treatment effect is defined as the difference in RMST at 1000 days after treatment initiation between ZDV+zal and ZDV. In table (a), the variable importance is estimated by the *expit*-transformation of $\hat{\zeta}_l^{CF}$, for l ranging over the single covariates, with corresponding confidence intervals. In table (b), the variable importance is estimated by $\hat{\Omega}_j^{CF}$, for j ranging over the single covariates.

essentially rendering them useless as measures of variable importance, but where the p-values associated with the hypothesis $H : \Gamma_j = 0$ provided significant findings.

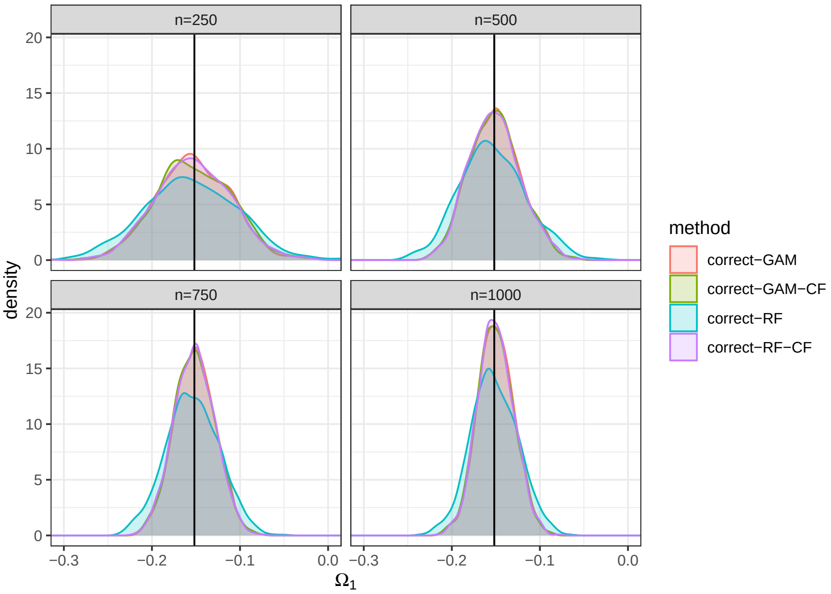


(a) Correctly specified Λ, Λ_c and π

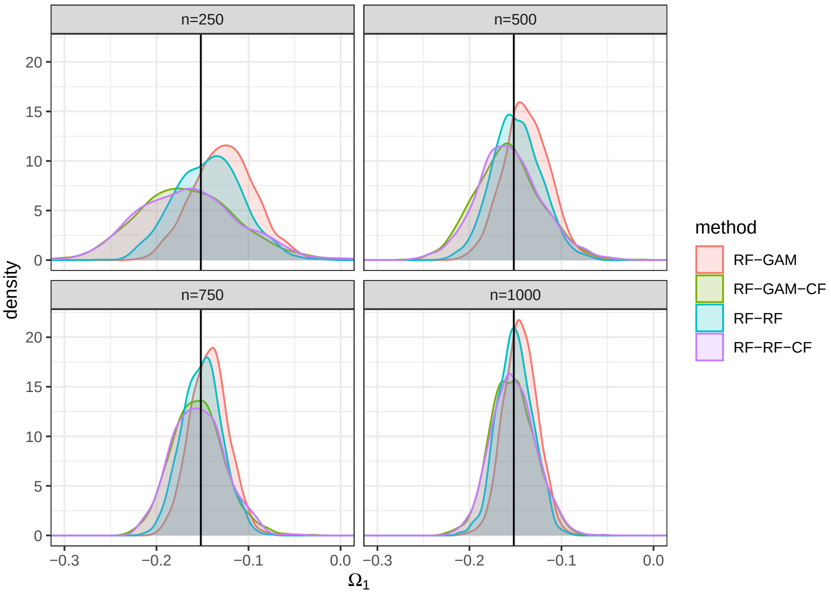


(b) Flexible estimation of Λ, Λ_c and π

Figure 1: Sampling distribution of estimators of Ψ_1 in the survival function setting with varying nuisance estimators, with and without cross-fitting, across sample sizes $n = 1000, 2000, 3000, 4000$. The abbreviation of the methods should be read as follows: A-B-C, where A corresponds to the nuisance estimators Λ, Λ_c and π, B corresponds to the nuisance estimator $\hat{E}_n$, and C corresponds to whether or not cross-fitting was used. Here, *correct* corresponds to correctly specified Cox and logistic regression, RF corresponds to Random Forest, and GAM corresponds to a generalized additive model



(a) Correctly specified Λ, Λ_c and π



(b) Flexible estimation of Λ, Λ_c and π

Figure 2: Sampling distribution of estimators of Ω_1 in the survival function setting with varying nuisance estimators, with and without cross-fitting, across sample sizes $n = 250, 500, 750, 1000$. The abbreveations of the methods are read as follows: A-B-C, where A corresponds to the nuisance estimators Λ, Λ_c and π, B corresponds to the nuisance estimators $\hat{E}_n$ and $\hat{E}_n^j$, and C corresponds to whether or not cross-fitting was used. Here, *correct* corresponds to correctly specified Cox and logistic regression, RF corresponds to Random Forest, and GAM corresponds to a generalized additive model.

References

Boileau, P., Leng, N., Hejazi, N. S., van der Laan, M., & Dudoit, S. (2023). A nonparametric framework for treatment effect modifier discovery in high dimensions. *arXiv preprint arXiv:2304.05323*.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters.

Cui, Y., Kosorok, M. R., Sverdrup, E., Wager, S., & Zhu, R. (2023). Estimating heterogeneous treatment effects with right-censored data via causal survival forests. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *85*(2), 179–211.

Gill, R. D., & Johansen, S. (1990). A survey of product-integration with a view toward application in survival analysis. *The annals of statistics*, *18*(4), 1501–1555.

Hammer, S. M., Katzenstein, D. A., Hughes, M. D., Gundacker, H., Schooley, R. T., Haubrich, R. H., Henry, W. K., Lederman, M. M., Phair, J. P., Niu, M., et al. (1996). A trial comparing nucleoside monotherapy with combination therapy in hiv-infected adults with cd4 cell counts from 200 to 500 per cubic millimeter. *New England Journal of Medicine*, *335*(15), 1081–1090.

Hines, O., Diaz-Ordaz, K., & Vansteelandt, S. (2022). Variable importance measures for heterogeneous causal effects. *arXiv preprint arXiv:2204.06030*.

Hines, O., Dukes, O., Diaz-Ordaz, K., & Vansteelandt, S. (2022). Demystifying statistical learning based on efficient influence functions. *The American Statistician*, *76*(3), 292–304.

Hu, L., Ji, J., & Li, F. (2021). Estimating heterogeneous survival treatment effect in observational data using machine learning. *Statistics in medicine*, *40*(21), 4691–4713.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., & Lauer, M. S. (2008). Random survival forests.

Ishwaran, H., Kogalur, U. B., & Kogalur, M. U. B. (2023). Package 'randomforestsrc'. *breast*, *6*(1), 854.

Kennedy, E. H. (2022a). Semiparametric doubly robust targeted double machine learning: A review. *arXiv preprint arXiv:2203.06469*.

Kennedy, E. H. (2022b). Towards optimal doubly robust estimation of heterogeneous causal effects. *arXiv preprint arXiv:2004.14497*.

Kennedy, E. H., Balakrishnan, S., & G'Sell, M. (2020). Sharp instruments for classifying compliers and generalizing causal effects.

Levy, J., van der Laan, M., Hubbard, A., & Pirracchio, R. (2021). A fundamental measure of treatment effect heterogeneity. *Journal of Causal Inference*, *9*(1), 83–108.

Li, H., Hubbard, A., & van der Laan, M. (2023). Targeted learning on variable importance measure for heterogeneous treatment effect. *arXiv preprint arXiv:2309.13324*.

Martinussen, T., & Stensrud, M. J. (2023). Estimation of separable direct and indirect effects in continuous time. *Biometrics*, *79*(1), 127–139.

Rytgaard, H. C., Eriksson, F., & van der Laan, M. J. (2023). Estimation of time-specific intervention effects on continuously distributed time-to-event outcomes by targeted maximum likelihood estimation. *Biometrics*.

Semenova, V., & Chernozhukov, V. (2021). Debiased machine learning of conditional average treatment effects and other causal functions. *The Econometrics Journal*, *24*(2), 264–289.

Tsiatis, A. A. (2006). *Semiparametric theory and missing data* (Vol. 4). Springer.

Van der Laan, M. J. (2006). Statistical inference for variable importance. *The International Journal of Biostatistics*, *2*(1).

van der Laan, M. J., & Robins, J. M. (2003). *Unified methods for censored longitudinal data and causality*. Springer.

van der Vaart, A. W. (2000). *Asymptotic statistics* (Vol. 3). Cambridge university press.

Vansteelandt, S., & Dukes, O. (2022a). Assumption-lean inference for generalised linear model parameters. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *84*(3), 657–685.

Vansteelandt, S., & Dukes, O. (2022b). Assumption-lean inference for generalised linear model parameters. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *84*(3), 657–685.

Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, *113*(523), 1228–1242.

Wei, W., Petersen, M., van der Laan, M. J., Zheng, Z., Wu, C., & Wang, J. (2023). Efficient targeted learning of heterogeneous treatment effects for multiple subgroups. *Biometrics*, *79*(3), 1934–1946.

Westling, T., Luedtke, A., Gilbert, P. B., & Carone, M. (2023). Inference for treatment-specific survival curves using machine learning. *Journal of the American Statistical Association*, (just-accepted), 1–26.

Xu, Y., Ignatiadis, N., Sverdrup, E., Fleming, S., Wager, S., & Shah, N. (2023). Treatment heterogeneity with survival outcomes. In *Handbook of matching and weighting adjustments for causal inference* (pp. 445–482). Chapman; Hall/CRC.

A On projection parameters

In related research, other authors have studied a variable importance measure that is closely related to our projection parameter, namely the so-called "best linear predictor/projection" (Semenova and Chernozhukov, 2021, Van der Laan, 2006, Cui et al., 2023, Boileau et al., 2023). In the case of continuous outcome, Semenova and Chernozhukov (2021) provide theoretical results via debiased machine-learning (see e.g. Chernozhukov et al. (2018)) and Van der Laan (2006) provide theoretical results based on semiparametric efficiency theory. In both cases they consider a target function/parameter, which is then approximated by a projection of the target function onto a working model indexed by a euclidean parameter. This approach is different from ours in that we seek interpretable summary statistics of our target function (CATE) through a projection, rather than estimating an interpretable target parameter through a projection. The approaches given by Cui et al. (2023) and Boileau et al. (2023) are more akin to ours. In the former, they provide a procedure for estimation of the best linear projection of τ in a survival context but without theoretical results, where the latter considers the best linear projection of τ onto a linear model, in the case where $\operatorname{E} X_i = 0$, $i = 1, \dots, d$, and derives an explicit parameter that is similar to ours, for which they provide an estimation procedure based on semiparametric efficiency theory. All of them consider a projection of τ onto a working model index by a euclidean parameter. In contrast, our projection parameter is defined through a projection of τ onto a subspace indexed by (β, w), where β is a real-valued parameter and w is a measurable function of $d-1$ variables. Since the space indexed by a euclidean parameter is a subspace of the space we consider, the error made from projecting τ onto a subspace is smaller in our setting compared to the best linear projection.

The above discussion will be clarified below. First, we state a result showing that our projection parameter is in fact the desired projection.

Let $\mathcal{H}$ be the Hilbert space of measurable functions with finite variance endowed with the covariance inner product. We have the following result:

Lemma A.1. The projection of $\tau \in \mathcal{H}$ onto the subspace $\mathcal{U} = \{u \in \mathcal{H} : u(x) = \beta x_j + w(x_{-j}),\ \beta \in \mathbb{R},\ Pw^2 < \infty\}$ is given by

$$\Pi(\tau \mid \mathcal{U}) = \beta^* X_j + w^*(X_{-j}),$$

where

$$\beta^* = \Omega_j \quad \text{and} \quad w^*(X_{-j}) = E(\tau(X) \mid X_{-j}) - \Omega_j E(X_j \mid X_{-j}).$$

Proof. We want to find β^* and w^* such that

$$(\beta^*, w^*) = \operatorname*{arg\,min}_{\beta, w} \operatorname{E}\{\tau(X) - \beta X_j - w(X_{-j})\}^2.$$

Observe that for any two measurable functions $a : \mathcal{X} \to \mathbb{R}$ and $b : \mathcal{X}_{-j} \to \mathbb{R}$ with finite variance, we have

$$E([a(X) - b(X_{-j})]^2 \mid X_{-j}) = [E(a(X) \mid X_{-j}) - b(X_{-j})]^2 + \operatorname{var}(a(X) \mid X_{-j}).$$

Hence

$$\begin{aligned}\operatorname{E}\{\tau(X) - \beta X_j - w(x_{-j})\}^2 = & \operatorname{E}\left\{[E(\tau(X) - \beta X_j \mid X_{-j}) - w(X_{-j})]^2\right\} \\ & + \operatorname{E}\{\operatorname{var}(\tau(X) - \beta X_j \mid X_{-j})\}.\end{aligned}$$

The second term on the right hand side of the latter display does not depend on w, and since the integrand in the first term is positive, the expression is minimized in w when the first term is equal to zero. This implies

$$w^*(x_{-j}) = E(\tau(X) - \beta X_j \mid X_{-j} = x_{-j}).$$

For β, observe that

$$\frac{\mathrm{d}}{\mathrm{d}\beta} \mathrm{E}\{\tau(X) - \beta X_j - w(X_{-j})\}^2 = -2\,\mathrm{E}\{X_j[\tau(X) - \beta X_j - w(x_{-j})]\} = 0,$$

together with w^* implies

$$\mathrm{E}\{X_j[\tau(X) - \beta X_j - E(\tau(X) - \beta X_j \mid X_{-j})]\} = 0,$$

and therefore

$$\beta^* = \frac{\mathrm{E}\{\mathrm{cov}(\tau(X), X_j \mid X_{-j})\}}{\mathrm{E}\{\mathrm{var}(X_j \mid X_{-j})\}} = \Omega_j$$

and

$$w^*(X_{-j}) = E(\tau(X) \mid X_{-j}) - \Omega_j E(X_j \mid X_{-j}).$$

□

Next, we consider the best linear projection of τ and contrast it with the best partially linear projection. To that end, we write the CATE function as

$$\tau(x) = \alpha + \gamma^T x + R_{\alpha,\gamma}(x)$$

for $\alpha \in \mathbb{R}$ and $\gamma \in \mathbb{R}^{\mathrm{d}}$ and let

$$(\alpha^*, \gamma^*) = \underset{\alpha,\gamma}{\arg\min}\, \mathrm{E}\{R_{\alpha,\gamma}(X)^2\} = \underset{\alpha,\gamma}{\arg\min}\, \mathrm{E}\{[\tau(X) - \alpha - \gamma^T X]^2\}.$$

We define the remainder corresponding to the best partially linear projection as $R_{\beta,w}(x)$ such that

$$(\beta^*, w^*) = \underset{\beta,w}{\arg\min}\, \mathrm{E}\{R_{\beta,w}(X_j, X_{-j})^2\} = \underset{\beta,w}{\arg\min}\, \mathrm{E}\{[\tau(X) - \beta X_j - w(X_{-j})]^2\}.$$

Since the space of linear functions is a subspace of the space of partially linear functions, $\mathcal{U}$, we have that $\alpha^* + \gamma^{*T} x \in \mathcal{U}$. By the projection theorem for Hilbert spaces (see e.g. Tsiatis, 2006, Theorem 2.1), the distance between τ and the projection onto $\mathcal{U}$ is smaller than the distance between τ and any other function in $\mathcal{U}$. Hence, by Lemma A.1,

$$\|R_{\beta^*,w^*}\| \leq \|R_{\alpha^*,\gamma^*}\|\,.$$

The result shows that the error made from model misspecification is smaller in the best partially linear projection compared to the best linear projection. As the dimension of X grows, the difference in the errors become larger (since the linear restriction of X_{-j} becomes increasingly strict compared to $w(X_{-j})$), and the best partially linear projection is thus better suited as a measure of importance of a single covariate.

B Deriviation of efficient influence functions

Here we derive the efficient influence functions from theorem 1 and 3. We will first derive the Gateaux derivatives of τ, τ_s and $\mathrm{E}^j(x)$, respectively, which are then used in the derivations of the EIF's through the chain rule. In the following we use subscripts P to underline the dependence on P. Let the parametric submodel be given by $P_\epsilon = Q\epsilon + (1-\epsilon)P \in \mathcal{M}$, where Q is the Dirac measure, and define the operator $\partial_\epsilon = \frac{d}{d\epsilon}\big|_{\epsilon=0}$ such that $\partial_\epsilon \psi(P_\epsilon) = \frac{d}{d\epsilon}\psi(P_\epsilon)\big|_{\epsilon=0}$ for some mapping $\psi : \mathcal{M} \to \mathbb{R}$. In the following let

$$g(A, X) = \left(\frac{\mathbb{1}(A=1)}{\pi(1 \mid X)} - \frac{\mathbb{1}(A=0)}{\pi(0 \mid X)}\right)$$

and

$$H(u, t, a, x) = \int_u^t S(u \mid a, x)\,\mathrm{d}u,$$

and we write $\tau = \tau_P$ to denote τ under distribution P.

Lemma B.1. The Gateaux derivative of $\tau(x)$ is given by

$$\partial_\epsilon S_{P_\epsilon}(t \mid 1, x) - S_{P_\epsilon}(t \mid 0, x) = \frac{\mathbb{1}(X=x)}{f(x)} g(A, x) \int_0^t \frac{-S(t \mid A, x)}{S(s \mid A, x) S_c(s \mid A, x)}\,\mathrm{d}M(s \mid A, x)$$

in the survival functions setting and

$$\begin{aligned}&\partial_\epsilon \int_0^{t^*} S_{P_\epsilon}(t \mid 1, x) - \int_0^t S_{P_\epsilon}(t \mid 0, x)\,\mathrm{d}t \\ =&\frac{\mathbb{1}(X=x)}{f(x)} g(A, x) \int_0^t \frac{-H(u, t, a, x)}{S(u \mid a, x) S_c(u \mid a, x)}\,\mathrm{d}M(u \mid a, x),\end{aligned}$$

in the RMST setting.

Proof. We start by calculating the Gateaux derivative of the conditional cumulative hazard function $\Lambda(t \mid a, x)$, which is also given in, e.g, the supplementary material of Martinussen and Stensrud (2023), but included here for completeness. Let $P(\tilde{T} \geq s, a, x) = \sum_{\delta=0,1} \int_s^\infty P(\mathrm{d}s, \delta, a, x)$ and note that $P(\tilde{T} \geq s \mid a, x) = S(s \mid a, x) S_c(s \mid a, x)$ because of independent censoring. Then

$$\begin{aligned}&\partial_\epsilon \Lambda_\epsilon(t \mid a, x) \\ =&\int_0^t \partial_\epsilon \frac{P_\epsilon(\mathrm{d}s, \Delta=1, a, x)}{P_\epsilon(\tilde{T} \geq s, a, x)} \\ =&\int_0^t \frac{Q(\mathrm{d}s, \Delta=1, a, x) - P(\mathrm{d}s, \Delta=1, a, x)}{P(\tilde{T} \geq s, a, x)} \\ &- \int_0^t \left(\sum_{\delta=0,1} \mathbb{1}(\tilde{T} \geq s, \delta, a, x) - P(\tilde{T} \geq s, \delta, a, x)\right) \frac{P(\mathrm{d}s, \Delta=1, a, x)}{P(\tilde{T} \geq s, a, x)^2} \\ =&\frac{\mathbb{1}(A=a)\mathbb{1}(X=x)}{\pi(a \mid x) f(x)} \left\{\int_0^t \frac{1}{P(\tilde{T} \geq s \mid a, x)}\,\mathrm{d}N(s) - \int_0^t \frac{\mathbb{1}(\tilde{T} \geq s)}{P(\tilde{T} \geq s \mid a, x)}\,\mathrm{d}\Lambda(s \mid a, x)\right\} \\ =&\frac{\mathbb{1}(A=a)\mathbb{1}(X=x)}{\pi(a \mid x) f(x)} \int_0^t \frac{1}{S(s \mid a, x) S_c(s \mid a, x)}\,\mathrm{d}M(s \mid a, x).\end{aligned}$$

Consider the survival function setting $\tau(x) = e^{-\Lambda(t|,1,x)} - e^{-\Lambda(t|,0,x)}$. A simple application of the chain rule gives

$$\partial_\epsilon \tau_{P_\epsilon}(x) = \frac{\mathbb{1}(X = x)}{f(x)} g(A, x) \int_0^t \frac{-S(t \mid A, x)}{S(s \mid A, x) S_c(s \mid A, x)} \,\mathrm{d}M(s \mid A, x),$$

which gives the first claim. Now, consider the RMST setting,

$$\tau(x) = \int_0^t S(u \mid 1, x)\, du - \int_0^t S(u \mid 0, x)\, du.$$

Again, the chain rule gives

$$\begin{aligned}&\partial_\epsilon \int_0^t S_{P_\epsilon}(s \mid 1, x)\,\mathrm{d}s - \int_0^t S_{P_\epsilon}(t \mid 0, x)\,\mathrm{d}s \\ =&\frac{\mathbb{1}(X = x)}{f(x)} g(A, x) \int_0^t \frac{-H(u, t, a, x)}{S(u \mid a, x) S_c(u \mid a, x)} \,\mathrm{d}M(u \mid a, x)\end{aligned}$$

where

$$H(u, t, a, x) = \int_u^t S(s \mid a, x)\,\mathrm{d}s,$$

which gives the second claim. □

The next result is from Hines, Diaz-Ordaz, and Vansteelandt (2022). The result is stated in equation (4) in their Appendix.

Lemma B.2 (Hines, Diaz-Ordaz, and Vansteelandt, 2022)**.** Let $g_P(X)$ denote some functional of P. Then

$$\begin{aligned}&\partial_\epsilon \operatorname{E}_{P_\epsilon}(g_{P_\epsilon}(X) \mid X_{-l} = x_{-l}) \\ =&\frac{1(X_{-l} = x_{-l})}{f(x_{-l})} \left[g_P(x) - \operatorname{E}(g_P(X) \mid X_{-l} = x_{-l})\right] + \operatorname{E}_P(\partial_\epsilon g_{P_\epsilon}(X) \mid X_{-l} = x_{-l}).\end{aligned}$$

Lemma B.3. The Gateaux derivative of $\tau_l(x)$ is given by

$$\begin{aligned}&\partial_\epsilon \operatorname{E}_{P_\epsilon}(\tau_{P_\epsilon}(X) \mid X_{-l} = x_{-l}) \\ =&\frac{\mathbb{1}(X_{-l} = x_{-l})}{f_{x_{-l}}(x_{-l})} \left(\tau(x) - \tau_l(x) + g(A, X) \int_0^t \frac{-S(t \mid A, x)}{S(s \mid A, x) S_c(s \mid A, x)} \,\mathrm{d}M(s \mid A, x)\right)\end{aligned}$$

in the survival setting and

$$\begin{aligned}&\partial_\epsilon \operatorname{E}_{P_\epsilon}(\tau_{P_\epsilon}(X) \mid X_{-l} = x_{-l}) \\ =&\frac{\mathbb{1}(X_{-l} = x_{-l})}{f_{x_{-l}}(x_{-l})} \left(\tau(x) - \tau_l(x) + g(A, X) \int_0^t \frac{-H(u, t, a, x)}{S(s \mid A, x) S_c(s \mid A, x)} \,\mathrm{d}M(s \mid A, x)\right)\end{aligned}$$

in the RMST setting.

Proof. We note that for any functional $g_P(X)$ with Gateaux derivative $\frac{\mathbb{1}(X=x)}{f(x)} v(O)$ for some function $v : \mathcal{O} \to \mathbb{R}$ we have

$$\operatorname{E}_P(\partial_\epsilon g_{P_\epsilon}(X) \mid X_{-s} = x_{-s}) = \frac{\mathbb{1}(X_{-s} = x_{-s})}{f(x_{-s})} v(O).$$

Let $g_P(x) = \tau(x)$. An application lemma B.2 followed by an application of lemma B.1 gives the result. □

B.1 Proof of Theorem 1

EIF in survival setting

Consider the survival function setting, $\tau(x) = \exp(-\Lambda(t \mid A = 1, x)) - \exp(-\Lambda(t \mid A = 0, x))$. Lemma B.1 gives the Gateaux derivative of τ from which we can calculate the EIF's of $\operatorname{var}(\tau(X))$ and $\operatorname{var}(\tau_l(X))$ by simple applications of the chain rule.

$$
\begin{aligned}
&\partial_\epsilon \operatorname{var}_{P_\epsilon}(\tau_{P_\epsilon}(X)) \\
=&\int \partial_\epsilon(\tau_{P_\epsilon}(X) - \operatorname{E}\tau_{P_\epsilon}(X))^2 \,\mathrm{d}P_\epsilon \\
=&(\tau_P(X) - \operatorname{E}\tau_P(X))^2 - \int(\tau_P(X) - \operatorname{E}\tau_P(X))^2 \,\mathrm{d}P + \int \partial_\epsilon(\tau_{P_\epsilon}(X) - \operatorname{E}\tau_{P_\epsilon}(X))^2 \,\mathrm{d}P \\
=&(\tau_P(X) - \operatorname{E}\tau_P(X))^2 - \operatorname{var}(\tau(X)) + \int 2(\tau_P(X) - \operatorname{E}\tau_P(X))\partial_\epsilon(\tau_{P_\epsilon}(X) - \operatorname{E}\tau_{P_\epsilon}(X)) \,\mathrm{d}P \\
=&(\tau_P(X) - \operatorname{E}\tau_P(X))^2 - \operatorname{var}(\tau(X)) + \int 2(\tau_P(X) - \operatorname{E}\tau_P(X))\partial_\epsilon\tau_{P_\epsilon}(X) \,\mathrm{d}P \\
=&(\tau_P(X) - \operatorname{E}\tau_P(X))^2 - \operatorname{var}(\tau_P(X)) \\
&+ 2(\tau_P(X) - \operatorname{E}\tau_P(X))g(A, X)\int_0^t \frac{-S(t \mid A, x)}{S(s \mid A, x)S_c(s \mid A, x)} \,\mathrm{d}M(s \mid A, x) \\
=&\ \tilde{\psi}_{\operatorname{var}(\tau(X))}
\end{aligned}
$$

Analogously we find the EIF of $\operatorname{var}(\tau_l(X))$ by use of the Gateaux derivative of τ_l from lemma B.3:

$$
\begin{aligned}
&\partial_\epsilon \operatorname{var}_{P_\epsilon}(\tau_{s,P_\epsilon}(X)) \\
=&\int \partial_\epsilon(\tau_{s,P_\epsilon}(X) - \operatorname{E}\tau_{s,P_\epsilon}(X))^2 \,\mathrm{d}P_\epsilon \\
=&(\tau_l(X) - \operatorname{E}(\tau(X)))^2 - \operatorname{var}(\tau(X)) \\
&+ 2(\tau_l(X) - \operatorname{E}(\tau(X)))\left(\tau(x) - \tau_l(x) + g(A, X)\int_0^t \frac{-S(t \mid A, x)}{S(s \mid A, x)S_c(s \mid A, x)} \,\mathrm{d}M(s \mid A, x)\right) \\
=&\tilde{\psi}_{\operatorname{var}(\tau_l(X))}
\end{aligned}
$$

noting that $\operatorname{E}\tau_l(X) = \operatorname{E}\tau(X)$. From the two EIF's we have that the EIF of Θ_l is given by their difference:

$$
\tilde{\psi}_{\Theta_l} = \tilde{\psi}_{\operatorname{var}(\tau(X))} - \tilde{\psi}_{\operatorname{var}(\tau_l(X))} \tag{16}
$$

and the EIF of Ψ_l is given by

$$
\Phi_l(O) = \frac{1}{\operatorname{var}(\tau(X))}\left(\tilde{\psi}_{\Theta_l}(O) - \Psi_l\tilde{\psi}_{\operatorname{var}(\tau(X))}(O)\right). \tag{17}
$$

EIF in restricted mean setting

Let $\tau(x) = \int_0^{t^*} S(t \mid 1, x)\,\mathrm{d}t - \int_0^{t^*} S(t \mid 0, x)\,\mathrm{d}t$. Since the structure of the Gateaux derivatives of τ and τ_l, from lemma B.1 and B.3, is identical to the survival case with $H(u,t,a,x)$ replacing $S(t \mid a,x)$, the calculations from the survival function setting apply and we have the EIF's of Θ_l and Ψ_l are given by (16) and (17), respectively with

$$\begin{aligned}\tilde{\psi}_{\mathrm{var}(\tau(X))} =&(\tau_P(X) - \mathrm{E}\,\tau_P(X))^2 - \mathrm{var}(\tau_P(X)) \\ &+ 2(\tau_P(X) - \mathrm{E}\,\tau_P(X))g(A,X)\int_0^t \frac{-H(u,t^*,A,x)}{S(u \mid A,x)S_c(u \mid A,x)}\,\mathrm{d}M(u \mid A,x)\end{aligned}$$

and

$$\begin{aligned}&\tilde{\psi}_{\mathrm{var}(\tau_l(X))} \\ =&(\tau_l(X) - \mathrm{E}(\tau_l(X)))^2 - \mathrm{var}(\tau(X)) \\ &+ 2(\tau_l(X) - \mathrm{E}(\tau_l(X)))\left(\tau(x) - \tau_l(x) + g(A,X)\int_0^t \frac{-H(u,t^*,A,x)}{S(u \mid A,x)S_c(u \mid A,x)}\,\mathrm{d}M(u \mid A,x)\right)\end{aligned}$$

B.2 Proof of Theorem 3

Let $g_P(X) = X_j$. Lemma B.2 then gives

$$\partial_\epsilon\, \mathrm{E}_{P_\epsilon}(X_j \mid X_{-j}) = \frac{\mathbb{1}(X_{-j} = x_{-j})}{f(x_{-j})}(X_j - \mathrm{E}(X_j \mid X_{-j})).$$

It follows immediately that the EIF of χ_j is given by

$$\begin{aligned}&\partial_\epsilon \chi_j(P_\epsilon) \\ =&\partial_\epsilon\, \mathrm{E}_{P_\epsilon}\{X_j - \mathrm{E}_{P_\epsilon}(X_j \mid X_{-j})\}^2 \\ =&(X_j - \mathrm{E}(X_j \mid X_{-j}))^2 - \chi_j(P) \\ &- 2\int [x_j - \mathrm{E}(X_j \mid X_{-j} = x_{-j})]\frac{\mathbb{1}(X_{-j} = x_{-j})}{f(x_{-j})}[X_j - \mathrm{E}(X_j \mid X_{-j} = x_{-j})]P_{X_j,X_{-j}}(\mathrm{d}(x_j, x_{-j})) \\ =&(X_j - \mathrm{E}(X_j \mid X_{-j}))^2 - \chi_j(P) \\ &- 2[X_j - \mathrm{E}(X_j \mid X_{-j})]\int [x_j - \mathrm{E}(X_j \mid X_{-j})]P_{X_j \mid X_{-j} = x_{-j}}(\mathrm{d}x_j) \\ =&(X_j - \mathrm{E}(X_j \mid X_{-j}))^2 - \chi_j(P) \\ =&\tilde{\psi}_{\chi_j}.\end{aligned}$$

For the derivation of the EIF of Γ_j we let φ denote the uncentered EIF of the ATE regardless of whether we consider the survival function setting or the RMST setting. Hence, by lemma B.1, we write the Gateaux derivative of the CATE function as

$$\partial_\epsilon \tau_{P_\epsilon}(x) = \frac{\mathbb{1}(X = x)}{f(x)}(\varphi(O) - \tau(x)),$$

and, by lemma B.3, the Gateaux derivative of $\tau_{\{j\}}$ as

$$\partial_\epsilon \tau_{\{j\},P_\epsilon}(x) = \frac{\mathbb{1}(X_{-j} = x_{-j})}{f(x_{-j})}(\varphi(O) - \mathrm{E}(\tau(X) \mid X_{-j})).$$

Observe that

$$\begin{aligned}\Gamma_j =& \mathrm{E}\{\mathrm{cov}(\tau(X), X_j \mid X_{-j})\}\\ =& \mathrm{E}\{E([\tau(X) - \mathrm{E}(\tau(X) \mid X_{-j})][X_j - \mathrm{E}(X_j \mid X_{-j})] \mid X_{-j})\}\\ =& \mathrm{E}\{E(\tau(X)[X_j - \mathrm{E}(X_j \mid X_{-j})] \mid X_{-j}) - \mathrm{E}(\tau(X) \mid X_{-j})\,\mathrm{E}(X_j - \mathrm{E}(X_j \mid X_{-j}) \mid X_{-j})\}\\ =& \mathrm{E}\{\tau(X)X_j\} - \mathrm{E}\{\mathrm{E}(\tau(X) \mid X_{-j})\,\mathrm{E}(X_j \mid X_{-j})\},\end{aligned}$$

and that the EIF of Γ_j is given by

$$\begin{aligned}\partial_\epsilon \Gamma_j(P_\epsilon) =& \partial_\epsilon\big\{\, \mathrm{E}_{P_\epsilon}\{\tau_{P_\epsilon}(X)X_j\} - \mathrm{E}_{P_\epsilon}\{\mathrm{E}_{P_\epsilon}(\tau_{P_\epsilon}(X) \mid X_{-j})\,\mathrm{E}_{P_\epsilon}(X_j \mid X_{-j})\}\big\}\\ =& \tau(X)X_j - \mathrm{E}\{\tau(X)X_j\} + X_j[\varphi(O) - \tau(X)]\\ &- \mathrm{E}(\tau(X) \mid X_{-j})\,\mathrm{E}(X_j \mid X_{-j}) + \mathrm{E}\{\mathrm{E}(\tau(X) \mid X_{-j})\,\mathrm{E}(X_j \mid X_{-j})\}\\ &- \partial_\epsilon\,\mathrm{E}\{\mathrm{E}_{P_\epsilon}(\tau_{P_\epsilon}(X) \mid X_{-j})\,\mathrm{E}(X_j \mid X_{-j})\} - \partial_\epsilon\,\mathrm{E}\{\mathrm{E}(\tau(X) \mid X_{-j})\,\mathrm{E}_{P_\epsilon}(X_j \mid X_{-j})\}\\ =& \tau(X)X_j - \mathrm{E}\{\tau(X)X_j\} + X_j[\varphi(O) - \tau(X)]\\ &- \mathrm{E}(\tau(X) \mid X_{-j})\,\mathrm{E}(X_j \mid X_{-j}) + \mathrm{E}\{\mathrm{E}(\tau(X) \mid X_{-j})\,\mathrm{E}(X_j \mid X_{-j})\}\\ &- [\varphi(O) - \mathrm{E}(\tau(X) \mid X_{-j})]\,\mathrm{E}(X_j \mid X_{-j}) - \mathrm{E}(\tau(X) \mid X_{-j})[X_j - (X_j \mid X_{-j})]\\ =& \varphi(O)[X_j - \mathrm{E}(X_j \mid X_{-j}] - \mathrm{E}(\tau(X) \mid X_{-j})[X_j - \mathrm{E}(X_j \mid X_{-j})] - \Gamma_j\\ =& [\varphi(O) - \mathrm{E}(\tau(X) \mid X_{-j})][X_j - \mathrm{E}(X_j \mid X_{-j}] - \Gamma_j\\ =& \tilde{\psi}_{\Gamma_j}\end{aligned}$$

The EIF of Ω_j follows by an application of the chain rule.

C Proofs of asymptotic results

The proofs of Theorem 2 and 4 follow the same recipe. The strategy is based on an expansion of the target parameter estimator in question as described in Kennedy (2022a), and we will give a short recap of the general idea. In the following, let $\psi(P)$ denote a generic target parameter with EIF given by $\tilde{\psi}_P = \varphi_P - \psi(P)$, i.e., the EIF is linear in $\psi(P)$. Define the corresponding cross-fitted one-step estimator

$$\hat{\psi}^{CF} = \sum_{k=1}^{K} \frac{n_k}{n} \mathbb{P}_n^k \varphi_{\hat{P}_k},$$

where $\hat{P}_{-k}$ is an estimate of P obtained from $\mathcal{V}_{-k}$. Consider the following expansion of $\mathbb{P}_n^k \varphi_{\hat{P}_{-k}}$.

$$\mathbb{P}_n^k \varphi_{\hat{P}_{-k}} = \mathbb{P}_n^k \tilde{\psi} + (\mathbb{P}_n^k - P)(\varphi_{\hat{P}_{-k}} - \varphi_P) + P\varphi_{\hat{P}_{-k}} - \psi(P).$$

Given the above expansion, we obtain the decomposition

$$\hat{\psi}^{CF} - \psi(P) = \mathbb{P}_n\tilde{\psi} + \underbrace{\sum_{k=1}^{K}\frac{n_k}{n}\mathbb{P}_n^k(\mathbb{P}_n^k - P)(\varphi_{\hat{P}_{-k}} - \varphi_P)}_{\text{empirical process term}} + \underbrace{\sum_{k=1}^{K}\frac{n_k}{n}P(\varphi_{\hat{P}_{-k}} - \psi(P))}_{\text{remainder term}}.$$

By Lemma 2 in the supplementary material in Kennedy et al. (2020), the empirical process term is $o_p(n^{-1/2})$ if $\left\|\varphi_{\hat{P}_{-k}} - \varphi_P\right\| = o_p(1)$ for each k. The remainder term is $o_p(n^{-1/2})$ if $P\varphi_{\hat{P}_{-k}} - \psi(P) = o_p(n^{-1/2})$ for each k by the continuous mapping theorem, since $\frac{n_k}{n} \overset{P}{\to} \frac{1}{K}$. This is essentially Proposition 2 in Kennedy (2022a). Hence, the statements in Theorem 2 and Theorem 4 follow, if we can show that $\left\|\varphi_{\hat{P}_{-k}} - \varphi_P\right\| = o_p(1)$ and that $P(\varphi_{\hat{P}_{-k}}) - \psi(P) = o_p(n^{-1/2})$ for the corresponding estimators. In the following, we drop the dependence on k to ease notation.

We start by stating two results related to the empirical process term and remainder term for the ATE, which will come in handy in the proofs of Theorem 2 and Theorem 4. The results are essentially found in Westling et al. (2023) for the survival function setting (albeit, stated slightly differently), but we repeat them here for completeness and extend them to the RMST setting.

Lemma C.1. Let φ be given as in 4 for the survival function setting and as in 5 in the RMST setting. Under assumption A1 and A4, $P\{\varphi(\hat{\nu}) - \tau\} = o_p(n^{-1/2})$.

Proof. The result for the survival function setting is proved in Westling et al. (2023) and Rytgaard et al. (2023). We include the computations for completeness and extend it to the RMST case.

Survival Case

Let $\tau_a(x) = S(t \mid A = a, X = x)$ such that

$$\varphi(\hat{\nu}) - \tau = \varphi_1(\hat{\nu}) - \tau_1 - (\varphi_0(\hat{\nu}) - \tau_0),$$

where $\varphi_a(\hat{\nu})$ is given in (4). Thus, to bound $P\{\varphi(\hat{\nu}) - \tau\}$, we only need to derive a bound for

$P\{\varphi_a(\hat{\nu}) - \tau_a\}$. In the following we consider the nuisance estimates fixed.

$$
\begin{aligned}
&\mathrm{E}\{\varphi_a(\hat{\nu})(O) - \tau_a(X)\} \\
=&\mathrm{E}\{E(\varphi_a(\hat{\nu})(O) - \tau_a(X) \mid A = a, X)\} \\
=&\mathrm{E}\left\{ \hat{S}(t \mid, A = a, X) - S(t \mid, A = a, X) - \frac{\mathbb{1}(A = a)\hat{S}(t \mid A, X)}{\hat{\pi}(a \mid X)} \right. \\
&\left. \times \left(\int_0^t \frac{\mathrm{E}(\mathrm{d}N(s) \mid A = a, X)}{\hat{S}(s \mid A, X)\hat{S}_c(s \mid A, X)} - \int_0^t \frac{\mathrm{E}(\mathbb{1}(\tilde{T} \geq t) \mid A = a, X)\, \mathrm{d}\hat{\Lambda}(s \mid A, X)}{\hat{S}(s \mid A, X)\hat{S}_c(s \mid A, X)} \right) \right\} \\
=&\mathrm{E}\left\{ \hat{S}(t \mid, A = a, X) - S(t \mid, A = a, X) \right. \\
&\left. - \frac{\mathbb{1}(A = a)\hat{S}(t \mid A, X)}{\hat{\pi}(a \mid X)} \int_0^t \frac{S(s \mid A, X)S_c(s \mid A, X)}{\hat{S}(s \mid A, X)\hat{S}_c(s \mid A, X)}\, \mathrm{d}\left[\Lambda(s \mid A, X) - \hat{\Lambda}(s \mid A, X)\right] \right\}.
\end{aligned}
\tag{18}
$$

Now, consider the survival function difference above. Using Duhamel's equation (Gill and Johansen, 1990) we have

$$
\hat{S}(t \mid a, x) - S(t \mid a, x) = \int_0^t \frac{S(s \mid a, x)}{\hat{S}(s \mid a, x)}\, \mathrm{d}\left[\Lambda(s \mid a, x) - \hat{\Lambda}(s \mid a, x)\right] \hat{S}(t \mid a, x).
$$

Plugging this into (18) yields

$$
\begin{aligned}
&\mathrm{E}\{\hat{\varphi}_a(O) - \tau_a(X)\} \\
=&\mathrm{E}\left\{ \int_0^t \left(1 - \frac{\pi(a \mid X)S_c(s \mid a, X)}{\hat{\pi}(a \mid X)\hat{S}_c(s \mid a, X)} \right) \frac{S(s \mid a, X)}{\hat{S}(s \mid a, X)} \hat{S}(t \mid a, X)\, \mathrm{d}\left[\Lambda(s \mid a, X) - \hat{\Lambda}(s \mid a, X)\right] \right\} \\
=&o_p(n^{-1/2})
\end{aligned}
$$

by assumption A4.

RMST case

Now consider the case where $\tau(x) = \int_0^{t^*} S(t \mid 1, x)\, \mathrm{d}t - \int_0^{t^*} S(t \mid 0, x)\, \mathrm{d}t$. As in the survival setting we define

$$
\tau_a(x) = \int_0^{t^*} S(t \mid a, x)\, \mathrm{d}t
$$

and

$$
\varphi_a(O) = \tau_a(X) - \frac{\mathbb{1}(A = a)}{\pi(a \mid x)} \int_0^{t^*} \frac{H(u, t, A, X)}{S(s- \mid A, X)S_C(s- \mid A, X)} dM(s \mid A, X).
$$

By derivations analogous to (18) we have

$$
\begin{aligned}
&\mathrm{E}\{\hat{\varphi}_a(O)-\tau(X)\}\\
&=\mathrm{E}\{E(\hat{\varphi}_a(O)-\tau(X)\mid A=a,X)\}\\
&=\mathrm{E}\left\{\int_0^t \hat{S}(s\mid a,X)-S(s\mid a,X)\,\mathrm{d}s\right.\\
&-\left.\frac{\pi(a\mid X)}{\hat{\pi}(a\mid X)}\int_0^t \frac{\hat{H}(s,t\ a,x)S(s\mid a,X)S_c(s\mid a,X)}{\hat{S}(s\mid a,X)\hat{S}_c(s\mid a,X)}\,\mathrm{d}\left[\Lambda(s\mid a,X)-\hat{\Lambda}(s\mid a,X)\right]\right\}\\
&=\mathrm{E}\left\{\int_0^t\int_0^s \frac{S(u\mid a,X)}{\hat{S}(u\mid a,X)}\,\mathrm{d}\left[\Lambda(u\mid a,X)-\hat{\Lambda}(u\mid a,X)\right]\hat{S}(s\mid a,X)\,\mathrm{d}s\right.\\
&-\left.\frac{\pi(a\mid X)}{\hat{\pi}(a\mid X)}\int_0^t \frac{\hat{H}(s,t\ a,x)S(s\mid a,X)S_c(s\mid a,X)}{\hat{S}(s\mid a,X)\hat{S}_c(s\mid a,X)}\,\mathrm{d}\left[\Lambda(s\mid a,X)-\hat{\Lambda}(s\mid a,X)\right]\right\}\\
&=\mathrm{E}\left\{\int_0^t \frac{\hat{H}(u,t\mid a,X)S(u\mid a,X)}{\hat{S}(u\mid a,X)}\,\mathrm{d}\left[\Lambda(u\mid a,X)-\hat{\Lambda}(u\mid a,X)\right]\right.\\
&-\left.\frac{\pi(a\mid X)}{\hat{\pi}(a\mid X)}\int_0^t \frac{\hat{H}(s,t\ a,x)S(s\mid a,X)S_c(s\mid a,X)}{\hat{S}(s\mid a,X)\hat{S}_c(s\mid a,X)}\,\mathrm{d}\left[\Lambda(s\mid a,X)-\hat{\Lambda}(s\mid a,X)\right]\right\}\\
&=\mathrm{E}\left\{\int_0^t \frac{\hat{H}(s,t\mid a,X)S(s\mid a,X)}{\hat{S}(s\mid a,X)}\left(1-\frac{S_c(s\mid a,X)\pi(a\mid X)}{\hat{S}_c(s\mid a,X)\hat{\pi}(a\mid X)}\right)\mathrm{d}\left[\Lambda(u\mid a,X)-\hat{\Lambda}(u\mid a,X)\right]\right\}\\
&=o_p(n^{-1/2})
\end{aligned}
$$

by assumption A4. The third equality follows from Duhamel's equation. □

Next we have a lemma, which is essentially given in Westling et al. (2023) (lemma 3 in their supplementary material) in the survival function setting, though our assumptions are stated slightly different. We include the proof for completeness and extend the result to the RMST setting.

Lemma C.2. Let φ be given as in (4) for the survival function setting and as in (5) for the RMST setting. Under assumption A1, A2 and A5 it holds that $\|\varphi(\hat{\nu})-\varphi(\nu)\|=o_p(1)$.

Proof. Observe that

$$\|\hat{\varphi}-\varphi\|\leq\|\hat{\varphi}_1-\varphi_1\|+\|\hat{\varphi}_0-\varphi_0\|$$

so that we only need to focus on $\|\hat{\varphi}_a-\varphi_a\|$. We start deriving a bound in the survival setting and then proceed to the RMST setting.

Survival function setting

Consider the decomposition

$$
\begin{aligned}
&\hat{\varphi}_a(O) - \varphi_a(O) \\
=&(\hat{\tau}(X) - \tau(X)) - \left(\frac{\mathbb{1}(A=a)}{\hat{\pi}(a \mid X)} \int_0^t \frac{\hat{S}(t \mid A, X)}{\hat{S}(s \mid A, X)\hat{S}_c(s \mid A, X)} \, \mathrm{d}\hat{M}(s \mid A, X) \right. \\
&\qquad \left. - \frac{\mathbb{1}(A=a)}{\pi(a \mid X)} \int_0^t \frac{S(t \mid A, X)}{S(s \mid A, X)S_c(s \mid A, X)} \, \mathrm{d}M(s \mid A, X) \right) \\
=&(\hat{\tau}(X) - \tau(X)) \\
&- \mathbb{1}(A=a) \left(\int_0^t \frac{\hat{S}(t \mid a, X)}{\hat{S}(s \mid a, X)\hat{g}(s \mid a, X)} - \frac{S(t \mid a, X)}{S(s \mid a, X)g(s \mid a, X)} \, \mathrm{d}N(s) \right) \\
&- \mathbb{1}(A=a) \left(\int_0^t \frac{\hat{S}(t \mid a, X)\mathbb{1}(\tilde{T} \geq s)}{\hat{S}(s \mid a, X)\hat{g}(s \mid a, X)} \hat{\Lambda}(\mathrm{d}s \mid a, X) - \int_0^t \frac{S(t \mid a, X)\mathbb{1}(\tilde{T} \geq s)}{S(s \mid a, X)g(s \mid a, X)} \Lambda(\mathrm{d}s \mid a, X) \right).
\end{aligned}
\tag{19}
$$

We need to bound each term in the above expression (separated by parentheses) individually. For the first term, A2 gives that $\|\hat{\tau} - \tau\| = o_p(1)$. For the second term we have for almost all x

$$
\begin{aligned}
&\mathbb{1}(A=a) \left(\int_0^t \frac{\hat{S}(t \mid a, x)}{\hat{S}(s \mid a, x)} \left(\frac{1}{\hat{g}(s \mid a, x)} - \frac{1}{g(s \mid a, x)} \right) \right. \\
&\qquad\qquad \left. - \frac{1}{g(s \mid a, x)} \left(\frac{\hat{S}(t \mid a, x)}{\hat{S}(s \mid a, x)} - \frac{S(t \mid a, x)}{S(s \mid a, x)} \right) \mathrm{d}N(s) \right)^2 \\
\leq& \ 2 \left(\int_0^t \frac{1}{\hat{g}(s \mid a, x)} - \frac{1}{g(s \mid a, x)} \mathrm{d}N(s) \right)^2 + 2\eta^{-2} \left(\int_0^t \frac{\hat{S}(t \mid a, x)}{\hat{S}(s \mid a, x)} - \frac{S(t \mid a, x)}{S(s \mid a, x)} \mathrm{d}N(s) \right)^2 \\
\leq& \ 2\eta^{-4} \left\{ \sup_{s \leq t} |\hat{g}(s \mid a, x) - g(s \mid a, x)| \right\}^2 + 4\eta^{-2} \left(\int_0^t \frac{1}{\hat{S}(s \mid a, x)} \left(\hat{S}(t \mid a, x) - S(t \mid a, x) \right) \mathrm{d}N(s) \right)^2 \\
&\qquad\qquad + 4\eta^{-2} \left(\int_0^t S(t \mid a, x) \left(\frac{1}{\hat{S}(s \mid a, x)} - \frac{1}{S(s \mid a, x)} \right) \mathrm{d}N(s) \right)^2 \\
\leq& \ 2\eta^{-4} \left\{ \sup_{s \leq t} |\hat{g}(s \mid a, x) - g(s \mid a, x)| \right\}^2 + 4\eta^{-4} \left(\hat{S}(t \mid a, x) - S(t \mid a, x) \right)^2 \\
&+ 4\eta^{-4} \left\{ \sup_{s \leq t} \left| \hat{S}(s \mid a, x) - S(s \mid a, x) \right| \right\}^2 \\
\leq& \ 2\eta^{-4} \left\{ \sup_{s \leq t} |\hat{g}(s \mid a, x) - g(s \mid a, x)| \right\}^2 + 8\eta^{-4} \left\{ \sup_{s \leq t} \left| \hat{S}(s \mid a, x) - S(s \mid a, x) \right| \right\}^2.
\end{aligned}
$$

which, together with A5, shows that the $L_2(P)$ norm of the second term converges in probability to zero. For the third term in (19), we use the same technique as described in the proof of lemma 3 in Westling et al. (2023). It is included here for completeness, and extended to the RMST setting in the following part of the proof. let $K_1(s, t \mid a, x) = \frac{S(t|a,x)}{S(s|a,x)}$ and let $\hat{K}_1$ be defined accordingly with $\hat{S}$ in place of S. The backwards equation (Gill and Johansen, 1990, Theorem 5) gives that for almost all x

$$K_1(s,t \mid a,x) = 1 - \int_s^t \frac{S(t \mid a,x)}{S(w \mid a,x)} \Lambda(\mathrm{d}w).$$

Hence, $K_1(\mathrm{d}s,t \mid a,x) = \frac{S(t|a,x)}{S(s|a,x)}\Lambda(\mathrm{d}s)$ and similarly, $\hat{K}_1(\mathrm{d}s,t \mid a,x) = \frac{\hat{S}(t|a,x)}{\hat{S}(s|a,x)}\hat{\Lambda}(\mathrm{d}s)$. This result allows us to write the third term in (19) for almost all x as (dropping the indicator $\mathbb{1}(A=a)$ since it disappears in the bound anyway)

$$\begin{aligned}
&\left(\int_0^t \frac{\hat{S}(t \mid a,X)\mathbb{1}(\tilde{T} \geq s)}{\hat{S}(s \mid a,X)\hat{g}(s \mid a,X)}\hat{\Lambda}(\mathrm{d}s \mid a,X) - \int_0^t \frac{S(t \mid a,X)\mathbb{1}(\tilde{T} \geq s)}{S(s \mid a,X)g(s \mid a,X)}\Lambda(\mathrm{d}s \mid a,X)\right)^2 \\
&= \left(\int_0^{t\wedge\tilde{T}} \frac{1}{\hat{g}(s \mid a,X)}\hat{K}_1(\mathrm{d}s \mid a,X) - \int_0^{t\wedge\tilde{T}} \frac{1}{g(s \mid a,X)}K_1(\mathrm{d}s \mid a,X)\right)^2 \\
&= \left(\int_0^{t\wedge\tilde{T}} \left(\frac{1}{\hat{g}(s \mid a,X)} - \frac{1}{g(s \mid a,X)}\right)\hat{K}_1(\mathrm{d}s \mid a,X)\right. \\
&\qquad\qquad \left. + \int_0^{t\wedge\tilde{T}} \frac{1}{g(s \mid a,X)}\left[\hat{K}_1(\mathrm{d}s \mid a,X) - K_1(\mathrm{d}s \mid a,X)\right]\right)^2. \qquad (20)
\end{aligned}$$

Thus, if we can show that the two integrals in the above display are consistent in $L_2(P)$-norm, it follows that (19) is is $o_p(1)$, which completes the proof. For the first integral in (20), we have for almost all x

$$\begin{aligned}
&\int_0^{t\wedge\tilde{T}} \left(\frac{1}{\hat{g}(s \mid a,X)} - \frac{1}{g(s \mid a,X)}\right)\hat{K}_1(\mathrm{d}s \mid a,X) \\
&= \int_0^{t\wedge\tilde{T}} \left(\frac{1}{\hat{g}(s \mid a,X)} - \frac{1}{g(s \mid a,X)}\right)\frac{\hat{S}(t \mid a,X)}{\hat{S}(s \mid a,X)}\hat{\Lambda}(\mathrm{d}s \mid a,X) \\
&\leq \sup_{s\leq t}\left|\frac{\hat{S}(t \mid a,x)}{\hat{S}(s \mid a,x)}\right| \sup_{s\leq t}\left|\frac{1}{\hat{g}(s \mid a,X)} - \frac{1}{g(s \mid a,X)}\right|\hat{\Lambda}(t) \\
&\leq \left|\log\eta\right|\eta^{-2}\sup_{s\leq t}\left|\hat{g}(s \mid a,x) - g(s \mid a,x)\right|,
\end{aligned}$$

where we have used $\frac{\hat{S}(t|a,x)}{\hat{S}(s|a,x)} \leq 1$ together with assumption A1. Assumption A5 then shows that the first integral in (20) is $o_p(1)$ in $L_2(P)$-norm. Using integration by parts, we can bound the second integral in (20). For almost all x we have

$$\begin{aligned}
&\int_0^{t\wedge\tilde{T}} \frac{1}{g(s\mid a,x)}\left[\hat{K}_1(\mathrm{d}s\mid a,x)-K_1(\mathrm{d}s\mid a,x)\right]\\
=&\frac{1}{g(t\mid a,x)}\left[\hat{K}_1(t\wedge\tilde{T},t\mid a,x)-K_1(t\wedge\tilde{T},t\mid a,x)\right]-\frac{1}{g(0\mid a,x)}\left[\hat{K}_1(0\mid a,x)-K_1(0,t\mid a,x)\right]\\
&\qquad\qquad-\int_0^{t\wedge\tilde{T}}\left[\hat{K}_1(s\mid a,x)-K_1(s,t\mid a,x)\right]\left(\frac{1}{g}\right)(\mathrm{d}s\mid a,x)\\
\leq&3\eta^{-1}\sup_{s\leq t}\left|\hat{K}_1(s\mid a,x)-K_1(s\mid a,x)\right|\\
\leq&C\sup_{s\leq t}\left|\hat{\Lambda}(s\mid a,x)-\Lambda(s\mid a,x)\right|
\end{aligned}$$

for some $C>0$, where the last inequality follows from the mean value theorem. By A5, the above expression is $o_p(1)$ in $L_2(P)$-norm and it follows that (20) is $o_p(1)$ in $L_2(P)$-norm, which completes the proof for the survival function setting.

RMST setting

Consider the decomposition

$$\begin{aligned}
&\hat{\varphi}_a(O)-\varphi_a(O)\\
=&(\hat{\tau}(X)-\tau(X))\\
&-\left(\frac{\mathbb{1}(A=a)}{\hat{\pi}(a\mid X)}\int_0^t\frac{\hat{H}(s,t\mid A,X)}{\hat{S}(s\mid A,X)\hat{S}_c(s\mid A,X)}\,\mathrm{d}\hat{M}(s\mid A,X)\right.\\
&\qquad\left.-\frac{\mathbb{1}(A=a)}{\pi(a\mid X)}\int_0^t\frac{H(s,t\mid A,x)}{S(s\mid A,X)S_c(s\mid A,X)}\,\mathrm{d}M(s\mid A,X)\right)\\
=&\left(\int_0^t\hat{S}(s\mid a,X)-S(s\mid a,X)\,\mathrm{d}u\right)\\
&-\mathbb{1}(A=a)\left(\int_0^t\frac{\hat{H}(s,t\mid a,X)}{\hat{S}(s\mid a,X)\hat{g}(s\mid a,X)}-\frac{H(s,t\mid a,X)}{S(s\mid a,X)g(s\mid a,X)}\,\mathrm{d}N(s)\right)\\
&-\mathbb{1}(A=a)\left(\int_0^t\frac{\hat{H}(s,t\mid a,X)\mathbb{1}(\tilde{T}\geq s)}{\hat{S}(s\mid a,X)\hat{g}(s\mid a,X)}\hat{\Lambda}(\mathrm{d}s\mid a,X)-\int_0^t\frac{H(s,t\mid a,X)\mathbb{1}(\tilde{T}\geq s)}{S(s\mid a,X)g(s\mid a,X)}\Lambda(\mathrm{d}s\mid a,X)\right).
\end{aligned}\tag{21}$$

Since the structure is similar to the survival function setting, the arguments will be similar too. We will again consider each term in turn. For first term we have

$$\int_0^t\hat{S}(s\mid a,X)-S(s\mid a,X)\,\mathrm{d}u\;\leq\;t\sup_{s<t}\left|\hat{S}(s\mid a,X)-S(s\mid a,X)\right|,$$

which is $o_p(1)$ in $L_2(P)$ by assumption A5. For the second term, note that $H(s,t\mid a,X)\leq t$, and

$$\begin{aligned}\left|\hat{H}(s,t \mid a,X) - H(s,t \mid a,X)\right| = & \int_s^t \hat{S}(s \mid a,X) - S(s \mid a,X)\,\mathrm{d}u \\ \leq & t \sup_{s<t} \left|\hat{S}(s \mid a,X) - S(s \mid a,X)\right|,\end{aligned}$$

and

$$\frac{H(s,t \mid a,X)}{S(s \mid a,X)} = \int_s^t \frac{S(u \mid a,X)}{S(s \mid a,X)}\,\mathrm{d}u \ \leq \ \int_s^t \mathrm{d}u \ \leq \ t.$$

Hence, replacing $S(t \mid a,X)$ and $\hat{S}(t \mid a,X)$ in the derivations from the survival function setting with $H(s,t \mid a,X)$ and $\hat{H}(s,t \mid a,X)$, gives that the second term in (21) is $o_p(1)$ in $L_2(P)$ by assumption A1 and A5 by similar arguments as in the survival function setting. For the third term in (21), we use the same strategy as in the survival function setting as described in Westling et al. (2023), but now extended to the RMST setting. Define

$$K_2(s,t \mid a,x) = \frac{H(s,t \mid, a,x)}{S(s \mid a,x)} - \int_s^t \mathrm{d}u,$$

and let $\hat{K}_2$ be defined accordingly, with $\hat{S}$ in place of S and $\hat{H} = \int \hat{S}\,\mathrm{d}s$. Then, by the backward equation (Gill and Johansen, 1990, Theorem 5)

$$\begin{aligned}K_2(s,t \mid a,x) &= \int_s^t \frac{S(u \mid a,x)}{S(s \mid a,x)}\,\mathrm{d}u - \int_s^t \mathrm{d}u \\ &= \int_s^t \left(1 - \int_s^u \frac{S(u \mid a,x)}{S(w \mid a,x)}\Lambda(\mathrm{d}w \mid a,x)\right) - \int_s^t \mathrm{d}u \\ &= -\int_s^t \int_s^u \frac{S(u \mid a,x)}{S(w \mid a,x)}\Lambda(\mathrm{d}w \mid a,x)\,\mathrm{d}u \\ &= -\int_s^t \int_w^t \frac{S(u \mid a,x)}{S(w \mid a,x)}\,\mathrm{d}u\Lambda(\mathrm{d}w \mid a,x) \\ &= -\int_s^t \frac{H(w,t \mid a,x)}{S(w \mid a,x)}\Lambda(\mathrm{d}w \mid a,x).\end{aligned}$$

Hence $K_2(\mathrm{d}s,t \mid a,x) = \frac{H(s,t|a,x)}{S(s|a,x)}\Lambda(\mathrm{d}s \mid a,x)$, and similarly for $\hat{K}_2$. Now, we can write the third term in (21) as (again, dropping $\mathbb{1}(A=a)$)

$$\begin{aligned}&\left(\int_0^t \frac{\hat{H}(s,t \mid a,X)\mathbb{1}(\tilde{T} \geq s)}{\hat{S}(s \mid a,X)\hat{g}(s \mid a,X)}\hat{\Lambda}(\mathrm{d}s \mid a,X) - \int_0^t \frac{H(s,t \mid a,X)\mathbb{1}(\tilde{T} \geq s)}{S(s \mid a,X)g(s \mid a,X)}\Lambda(\mathrm{d}s \mid a,X)\right)^2 \\ =& \left(\int_0^{t\wedge\tilde{T}} \frac{1}{\hat{g}(s \mid a,X)}\hat{K}_2(\mathrm{d}s \mid a,X) - \int_0^{t\wedge\tilde{T}} \frac{1}{g(s \mid a,X)}K_2(\mathrm{d}s \mid a,X)\right)^2. \\ =& \left(\int_0^{t\wedge\tilde{T}} \left(\frac{1}{\hat{g}(s \mid a,X)} - \frac{1}{g(s \mid a,X)}\right)\hat{K}_2(\mathrm{d}s \mid a,X)\right. \\ &\left.+ \int_0^{t\wedge\tilde{T}} \frac{1}{g(s \mid a,X)}\left[\hat{K}_2(\mathrm{d}s \mid a,X) - K_2(\mathrm{d}s \mid a,X)\right]\right)^2.\end{aligned}$$

Hence, by the same arguments used for in the survival function setting, it follows the third term in (21) is $o_p(1)$ in $L_2(P)$-norm, which concludes the proof. □

C.1 Proof of Theorem 2

C.1.1 Remainder term

We will use a decomposition of the remainder term $P\phi_{\Theta_l}(\hat{\nu}) - \Theta_l$ from Hines, Diaz-Ordaz, and Vansteelandt (2022). Observe that $\Theta_l = P\{\tau - \tau_l\}^2$. Then

$$\begin{aligned}
P\phi(\hat{\nu}) - \Theta_l &= P\{(\hat{\tau}_l - \varphi(\hat{\nu}))^2 - (\hat{\tau} - \varphi(\hat{\nu}))^2 - (\tau - \tau_l)^2\} \\
&= P\{(\hat{\tau}_l - \varphi(\hat{\nu}))^2 - (\hat{\tau} - \varphi(\hat{\nu}))^2 - \tau^2 - \tau_l^2 + 2\tau_l^2\} \\
&= P\{(\hat{\tau}_l - \tau_l)^2 - (\hat{\tau} - \tau)^2 + 2\hat{\tau}_l\tau_l - 2\hat{\tau}\tau + 2(\hat{\tau} - \hat{\tau}_l)\hat{\varphi}\} \\
&= \|\hat{\tau}_l - \tau_l\|^2 - \|\hat{\tau} - \tau\|^2 + P\{2\hat{\tau}_l\tau_l - 2\hat{\tau}_l\tau + 2(\hat{\tau} - \hat{\tau}_l)(\hat{\varphi} - \tau)\} \\
&= \|\hat{\tau}_l - \tau_l\|^2 - \|\hat{\tau} - \tau\|^2 + 2P\{(\hat{\tau} - \hat{\tau}_l)(\hat{\varphi} - \tau)\} \\
&\leq o_p(n^{-1/2}) + 2KP(\varphi(\hat{\nu}) - \tau) \\
&\leq o_p(n^{-1/2})
\end{aligned}$$

for some $K \geq 0$, where the second and fifth equality is due to iterated expectation, the first inequality follows from assumption A2 and A3 the fact that $\hat{\tau}(X) - \hat{\tau}_l(X)$ is bounded almost surely and the second inequality follows from lemma C.1.

C.1.2 Empirical process term

We need to show that $\|\phi_{\Theta_l}(\hat{\nu}) - \phi_{\Theta_l}(\nu)\| = o_p(1)$. Consider the decomposition given in Hines, Diaz-Ordaz, and Vansteelandt (2022)

$$\begin{aligned}
\phi_{\Theta_l}(\hat{\nu}) - \phi_{\Theta_l}(\nu) &= (\hat{\tau}_l - \tau_l)^2 - (\hat{\tau} - \tau)^2 + 2(\varphi - \tau_l)(\tau_l - \hat{\tau}_l) - 2(\varphi - \tau)(\tau - \hat{\tau}) + 2(\hat{\varphi} - \varphi)(\hat{\tau} - \hat{\tau}_l) \\
&= \sum_{i=1}^{5} a_i
\end{aligned}$$

so that $\|\phi(\hat{\nu}) - \phi(\nu)\| \leq \sum_{i=1}^{5} \|a_i\|$. We will treat each term separately.

(a_1): From A2 we have that $\left\|(\hat{\tau}_l - \tau_l)^2\right\| = o_p(1)$ since $x \mapsto x^2$ is continuous.

(a_2): Same argument as in (a_1).

(a_3): Consider the survival case. Then the following bound holds almost surely:

$$\begin{aligned}
&(\varphi(O) - \tau_l(x))^2 \\
=&\left(\tau(x) - \tau_l(x) - \left(\frac{\mathbb{1}(A=1)}{\pi(1 \mid x)} - \frac{\mathbb{1}(A=0)}{\pi(0 \mid x)}\right)\int_0^t \frac{S(t \mid A, x)}{S(s \mid A, x) S_c(s \mid A, x)} \,\mathrm{d}M(s \mid A, x)\right)^2 \\
\leq& 2(\tau(x) - \tau_l(x))^2 + 2\left(\eta^{-1}\int_0^t \frac{S(t \mid A, x)}{S(s \mid A, x)} \,\mathrm{d}N(s) - \eta^{-1}\int_0^t \frac{S(t \mid A, x)}{S(s \mid A, x)} \mathbb{1}(\tilde{T} \geq s) \,\mathrm{d}\Lambda(s \mid A, x)\right)^2 \\
\leq& 2K^2 + 4(\eta^{-1})^2 + 4\left(\eta^{-1}\int_0^t \frac{S(t \mid A, x)}{S(s \mid A, x)} \,\mathrm{d}\Lambda(s \mid A, x)\right)^2 \\
=& 2K^2 + 4(\eta^{-1})^2 + 4(\eta^{-1})^2 \left(1 - S(t \mid a, x)\right)^2 \\
\leq& 2K^2 + 8\eta^{-2}
\end{aligned}$$

where the first inequality comes A1, the second inequality from $\tau(x) - \tau_l(x)$ being bounded almost surely and that $\frac{S(t|a,x)}{S(s|a,x)} \leq 1$, $s \leq t$. The second equality is due the backward equation (theorem 5, Gill and Johansen, 1990), realising that $\frac{S(t|a,x)}{S(s|a,x)} = \prod_{]s,t]}(1 - d\Lambda(u \mid a, x))$.

Now consider the RMST setting. Start by observing that

$$\frac{H(s, t \mid a, x)}{S(s \mid a, x)} = \int_s^t \frac{S(u \mid a, x)}{S(s \mid a, x)} \,\mathrm{d}u \leq t - s \leq t$$

and

$$\int_0^t \frac{H(s, t \mid a, x)}{S(s \mid a, x)} \,\mathrm{d}\Lambda(s \mid, a, x) = \int_0^t \int_0^u \frac{S(u \mid a, x)}{S(s \mid a, x)} \,\mathrm{d}\Lambda(s \mid, a, x) \,\mathrm{d}u = \int_0^t S(u \mid a, x) - 1 \,\mathrm{d}u$$

by the backward equation. Then by calculations similar to the once from the survival setting we have

$$\begin{aligned}
&(\varphi(O) - \tau_l(x))^2 \\
\leq& 2(\tau(x) - \tau_l(x))^2 + 2\left(\eta^{-1}\int_0^t \frac{H(s, t \mid A, x)}{S(s \mid A, x)} \,\mathrm{d}N(s) - \eta^{-1}\int_0^t \frac{H(s, t \mid A, x)}{S(s \mid A, x)} \mathbb{1}(\tilde{T} \geq s) \,\mathrm{d}\Lambda(s \mid A, x)\right)^2 \\
\leq& 2K^2 + 4(\eta^{-1}t)^2 + 4\left(\eta^{-1}\int_0^t S(u \mid a, x) - 1 \,\mathrm{d}u\right)^2 \\
=& 2K^2 + 4\eta^{-2}t^2 + 4\eta^{-2}(t^2 + t^2) \\
=& 2K^2 + 12\eta^{-2}t^2.
\end{aligned}$$

Letting $C(\eta, t) = \max\{2K^2 + 8\eta^{-2}, 2K^2 + 12\eta^{-2}t^2\}$, A2 gives that

$$\|a_3\| \leq \sqrt{C(\eta, t)}\, \|\tau_l - \hat{\tau}_l\| = o_p(1).$$

(a_4) : Noticing that

$$(\varphi(O) - \tau(x))^2 = \left(-\left(\frac{\mathbb{1}(A=1)}{\pi(1 \mid x)} - \frac{\mathbb{1}(A=0)}{\pi(0 \mid x)}\right)\int_0^t \frac{S(t \mid A, x)}{S(s \mid A, x) S_c(s \mid A, x)} \,\mathrm{d}M(s \mid A, x)\right)^2$$

in the survival setting and

$$(\varphi(O) - \tau(x))^2 = \left(- \left(\frac{\mathbb{1}(A=1)}{\pi(1 \mid x)} - \frac{\mathbb{1}(A=0)}{\pi(0 \mid x)} \right) \int_0^t \frac{H(s, t \mid A, x)}{S(s \mid A, x) S_c(s \mid A, x)} \mathrm{d}M(s \mid A, x) \right)^2$$

in the RMST setting, the calculations from (a_3) gives that

$$\|a_4\| \leq \sqrt{C(\eta, t)}\, \|\tau - \hat{\tau}\| = o_p(1)$$

with $C(\eta, t) = \max\{4\eta^{-2}, 8\eta^{-2}t^2\}$.

(a_5) : By lemma C.2 and assumption A6, $\|a_5\| = o_p(1)$.

C.2 Proof of Theorem 4

C.2.1 Remainder term related to $\hat{\Gamma}_j^{CF}$

Consider the decomposition

$$\begin{aligned}
& P\{\phi_{\Gamma_j}(\hat{\nu}) - \phi_{\Gamma_j}(\nu)\} \\
=& \mathrm{E}\left\{ [\varphi(\hat{\nu})(O) - \hat{\tau}_j(X)][X_j - \hat{E}_n^j(X_{-j})] - [\varphi(\nu)(O) - \tau_j(X)][X_j - E(X_j \mid X_{-j})] \right\} \\
=& \mathrm{E}\left\{ [\varphi(\hat{\nu})(O) - \varphi(\nu)(O)]X_j + [\tau_j(X) - \hat{\tau}_j(X)]X_j \right. \\
& \left. - \varphi(\hat{\nu})(O)\hat{E}_n^j(X_{-j}) + \hat{\tau}_j(X)\hat{E}_n^j(X_{-j}) + \varphi(\nu)(O)E(X_j \mid X_{-j}) - \tau_j(X)E(X_j \mid X_{-j}) \right\} \\
=& E\left\{ [\hat{\varphi}(\nu)(O) - \varphi(\nu)(O)][X_j - \hat{E}_n^j(X_{-j})] \right. \\
& - [\hat{\tau}_j(X) - \tau_j(X)][X_j - \hat{E}_n^j(X_{-j})] \\
& \left. - [\hat{E}_n^j(X_{-j}) - E(X_j \mid X_{-j})][\varphi(\nu)(O) - \tau_j(X)] \right\}.
\end{aligned}$$

We note that

$$\mathrm{E}\{\varphi(\nu)(O) - \tau_j(X)\} = \mathrm{E}\{E(\tau(X) \mid X_{-j}) + E(M \mid A, X) - \tau_j(X)\} = 0$$

by iterated expectation, where M is the martingale integral in the expression of φ, which is itself a martingale conditional on A and X. Thus, by iterated expectation and assumption B, the remainder term is given by

$$\begin{aligned}
& E\left\{ [\hat{\varphi}(\nu)(O) - \varphi(\nu)(O)][X_j - \hat{E}_n^j(X_{-j})] - [\hat{\tau}_j(X) - \tau_j(X)][X_j - \hat{E}_n^j(X_{-j})] \right\} \\
\leq & \sqrt{\delta} P\{\varphi(\hat{\nu}) - \tau\} - \|\hat{\tau}_j - \tau_j\| \left\| \hat{E}_n^j - E(\cdot \mid X_{-j}) \right\| \\
= & o_p(n^{-1/2})
\end{aligned}$$

Here, the inequality is given by Cauchy-Schwarz together with assumption B2 and the equality is given by assumption A3, B1 and lemma C.1.

C.2.2 Empirical process term related to $\hat{\Gamma}_j^{CF}$

Consider again the decomposition from the remainder term:

$$\begin{aligned}\phi_{\Gamma_j}(\hat{\nu}) - \phi_{\Gamma_j}(\nu) =& [\hat{\varphi}(\nu)(O) - \varphi(\nu)(O)][X_j - \hat{E}_n^j(X_{-j})] \\ & - [\hat{\tau}_j(X) - \tau_j(X)][X_j - \hat{E}_n^j(X_{-j})] \\ & - [\hat{E}_n^j(X_{-j}) - E(X_j \mid X_{-j})][\varphi(\nu)(O) - \tau_j(X)].\end{aligned}$$

Thus, we need to bound each term in $L_2(P)$. For the first term we have

$$\mathrm{E}\left\{[\hat{\varphi}(\nu)(O) - \varphi(\nu)(O)]^2[X_j - \hat{E}_n^j(X_{-j})]^2\right\} \leq \delta^2 \left\|\varphi(\hat{\nu}) - \varphi(\nu)\right\|^2 = o_p(1)$$

by lemma C.2 and assumption B2. Consistency of the second term follows from the consistency of $\hat{\tau}$ together with assumption B2, and consistency of the third term follows from consistency of $\hat{E}_n^j$ together with the bound of $P\{\varphi - \tau_j\}^2$ calculated in $(a3)$ in the empirical process section of the proof of Theorem 2.

C.2.3 Remainder term related to $\hat{\chi}_j^{CF}$

We note that

$$\begin{aligned}& \mathrm{E}\left\{[X_j - E(X_j \mid X_{-j})]^2\right\} \\ =\ & \mathrm{E}\left\{X_j^2 + E(X_j \mid X_{-j})^2\right\} - 2\,\mathrm{E}\left\{X_j E(X_j \mid X_{-j})\right\} \\ =\ & \mathrm{E}\left\{X_j^2 + E(X_j \mid X_{-j})^2\right\} - 2\,\mathrm{E}\left\{E(X_j \mid X_{-j})^2\right\} \\ =\ & \mathrm{E}\left\{X_j^2 - E(X_j \mid X_{-j})^2\right\}\end{aligned}$$

by iterated expectation. Hence

$$\begin{aligned}& P\left\{\phi_{\chi_j}(\hat{\nu}) - \phi_{\chi_j}(\nu)\right\} \\ =& \mathrm{E}\left\{\phi_{\chi_j}(\hat{\nu}) - X_j^2 + E(X_j \mid X_{-j})^2\right\} \\ =& \mathrm{E}\left\{X_j^2 + \hat{E}_n^{j2}(X_{-j}) - 2X_j\hat{E}_n^{j2}(X_{-j}) - X_j^2 + E(X_j \mid X_{-j})^2\right\} \\ =& \mathrm{E}\left\{\left[\hat{E}_n^j(X_{-j}) - E(X_j \mid X_{-j})\right]^2 + 2\hat{E}_n^j(X_{-j})E(X_j \mid X_{-j}) - 2X_j\hat{E}_n^{j2}(X_{-j})\right\} \\ =& \left\|\hat{E}_n^j - E(\cdot \mid X_{-j})\right\|^2 \\ =& o_p(n^{-1/2}),\end{aligned}$$

where the fourth equality is due to iterated expectation and the last equality is given by assumption B1.

C.2.4 Empirical process term related to $\hat{\chi}_j^{CF}$

Consider

$$\begin{aligned}& \phi_{\chi_j}(\hat{\nu}) - \phi_{\chi_j}(\nu) \\ =& \left[\hat{E}_n^j(X_{-j}) - E(X_j \mid X_{-j})\right]^2 + 2\left[E(X_j \mid X_{-j}) - X_j\right]\left[\hat{E}_n^j(X_{-j}) - E(X_j \mid X_{-j})\right],\end{aligned}$$

where the consistency of $\hat{E}_n^j$ gives consistency in $L_2(P)$ of the first term. For the second term, note that

$$E([E(X_j \mid X_{-j}) - X_j]^2 \mid X_{-j}) = \operatorname{var}(X_j \mid X_{-j})$$

and hence

$$\begin{aligned}
&E\left\{4\left[E(X_j \mid X_{-j}) - X_j\right]^2\left[\hat{E}_n^j(X_{-j}) - E(X_j \mid X_{-j})\right]^2\right\}\\
=&4E\left\{\operatorname{var}(X_j \mid X_{-j})\left[\hat{E}_n^j(X_{-j}) - E(X_j \mid X_{-j})\right]^2\right\}\\
\leq&4K\operatorname{var}(X_j \mid X_{-j})\left\|\hat{E}_n^j - E(\cdot \mid X_{-j})\right\|^2\\
=&o_p(1)
\end{aligned}$$

for some $K > 0$, by assumption B1 and boundedness $\operatorname{var}(X_j \mid X_{-j})$, which gives the result.

C.3 Consistency of cross-fitted variance estimators

Proof of Lemma 1. The first claim, (13), is given by the functional delta method (van der Vaart, 2000, ch. 25.7) and hence $\sqrt{n}(\hat{\Psi}^{CF} - \Psi) \xrightarrow{D} \mathcal{N}(0, P\tilde{\psi}^2)$. By Prohorov's theorem (van der Vaart, 2000 theorem 2.4) $\sqrt{n}\left\|\hat{\Psi}^{CF} - \Psi\right\| = O_p(1)$ and hence $\left\|\hat{\Psi}^{CF} - \Psi\right\| = o_p(n^{-1/2})$. By the same argument we have $\left\|\hat{\psi}_i^{CF} - \psi_i\right\| = o_p(n^{-1/2})$, $i = 1, 2$. For the claim (14) we note that it is suffices to show that

$$\mathbb{P}_n^k\tilde{\psi}(\hat{\psi}_i^{CF}, \hat{\nu}_{i,-k})^2 \xrightarrow{P} P\tilde{\psi}(\psi_i, \nu_i)^2$$

for each k, since K is assumed finite and not depending on n. We note that

$$\mathbb{P}_n^k\tilde{\psi}(\hat{\psi}_i^{CF}, \hat{\nu}_{i,-k})^2 = \hat{\psi}_i^{2,CF} + \mathbb{P}_n^k\varphi_i(\hat{\nu}_{i,-k})^2 - 2\hat{\psi}_i^{CF}\mathbb{P}_n^k\varphi_i(\hat{\nu}_{i,-k}) \tag{22}$$

and that $\hat{\psi}_i^{2,CF} \xrightarrow{P} \psi_i^2$ by the continuous mapping theorem. Hence, by the continuous mapping theorem, the result follows if each of the $\mathbb{P}_n^k$-sums in the above display are consistent. Observe that

$$\mathbb{P}_n^k\varphi_i(\hat{\nu}_{i,-k}) - P\varphi(\nu) = (\mathbb{P}_n^k - P)(\varphi_i(\hat{\nu}_{i,-k}) - \varphi(\nu)) + P(\varphi_i(\hat{\nu}_{i,-k} - \varphi(\nu)) + (\mathbb{P}_n^k - P)\varphi_i(\nu).$$

The first term above is $o_p(n^{-1/2})$ by lemma 2 in the supplementary material of Kennedy et al. (2020). The second term is $o_p(1)$ since $P(\varphi_i(\hat{\nu}_{i,-k}) - \varphi(\nu)) \leq \|\varphi_i(\hat{\nu}_{i,-k}) - \varphi(\nu)\| = o_p(1)$ by assumption and the third term is $o_p(1)$ by the law of large numbers, since $\mathrm{E}\{\varphi_i(\nu_i)(O_i)\} = \psi_i < \infty$. Thus $\mathbb{P}_n^k\varphi_i(\hat{\nu}_{i,-k})$ converges to $P\varphi_i(\nu_i)$ in probability. For the sum $\mathbb{P}_n^k\varphi_i(\hat{\nu}_{i,-k})^2$ in (22), we note that since $\|\varphi_i(\hat{\nu}_{i,-k}) - \varphi_i(\nu_i)\| = o_p(1)$, by assumption, the continuous mapping theorem (for metric spaces, see e.g. van der Vaart, 2000, Theorem 18.11) gives that $\left\|\varphi_i(\hat{\nu}_{i,-k})^2 - \varphi_i(\nu_i)^2\right\| = o_p(1)$. Combined with the fact that $\mathrm{E}\{\varphi_i(\nu_i)(O_i)^2\} < \infty$, we can use the same arguments given for the consistency of $\mathbb{P}_n^k\varphi_i(\hat{\nu}_{i,-k})$, to show that $\mathbb{P}_n^k\varphi_i(\hat{\nu}_{i,-k})^2$ converges in probability to $P\varphi_i(\nu_i)^2$. Collecting the results, the continuous mapping theorem now gives the following convergence for (22):

$$\hat{\psi}_i^{2,CF} + \mathbb{P}_n^k\varphi_i(\hat{\nu}_{i,-k})^2 - 2\hat{\psi}_i^{CF}\mathbb{P}_n^k\varphi_i(\hat{\nu}_{i,-k}) \xrightarrow{P} \psi_i^2 + \mathbb{P}_n^k\varphi_i(\nu_i)^2 - 2\psi_i\mathbb{P}_n^k\varphi_i(\nu_i) = P\tilde{\psi}_i(\psi_i, \nu_i)^2,$$

and hence, the claim, (14), follows.

As for the second claim, the last claim, (15), follows if

$$\mathbb{P}_n^k\tilde{\psi}(\hat{\psi}_1^{CF},\hat{\psi}_2^{CF},\hat{\nu}_{1,-k},\hat{\nu}_{2,-k})^2 \xrightarrow{P} P\tilde{\psi}(\psi_1,\psi_2,\nu_1,\nu_2)^2$$

for each k, since K is assumed finite and not depending on n. Observe that

$$\begin{aligned}
&\mathbb{P}_n^k\tilde{\psi}(\hat{\psi}_1^{CF},\hat{\psi}_2^{CF},\hat{\nu}_{1,-k},\hat{\nu}_{2,-k})^2\\
=&\mathbb{P}_n^k\frac{1}{\left(\hat{\psi}_2^{CF}\right)^2}\left(\varphi_1(\hat{\nu}_{1,-k})-\hat{\psi}_1^{CF}-\frac{\hat{\psi}_1^{CF}}{\hat{\psi}_2^{CF}}\left(\varphi_2(\hat{\nu}_{2,-k})-\hat{\psi}_2^{CF}\right)\right)^2\\
=&\frac{1}{\left(\hat{\psi}_2^{CF}\right)^2}\mathbb{P}_n^k\left(\varphi_1(\hat{\nu}_{1,-k})-\hat{\psi}_1^{CF}\right)^2+\left(\hat{\Psi}^{CF}\right)^2\mathbb{P}_n^k\left(\varphi_2(\hat{\nu}_{2,-k})-\hat{\psi}_2^{CF}\right)^2\\
&\qquad-2\hat{\Psi}^{CF}\mathbb{P}_n^k\left(\varphi_1(\hat{\nu}_{1,-k})-\hat{\psi}_1^{CF}\right)\left(\varphi_2(\hat{\nu}_{2,-k})-\hat{\psi}_2^{CF}\right).
\end{aligned}$$

We will consider each term in the above display separately. In the proof of (14), we showed consistency of the $\mathbb{P}_n^k$-sums in the first two terms, and the continuous mapping theorem gives that $\frac{1}{\left(\hat{\psi}_2^{CF}\right)^2}\xrightarrow{P}\frac{1}{\psi_2^2}$ and $\left(\hat{\Psi}^{CF}\right)^2\xrightarrow{P}\Psi^2$. The continuous mapping theorem then gives consistency of the first two terms. For the last term, we use the decomposition

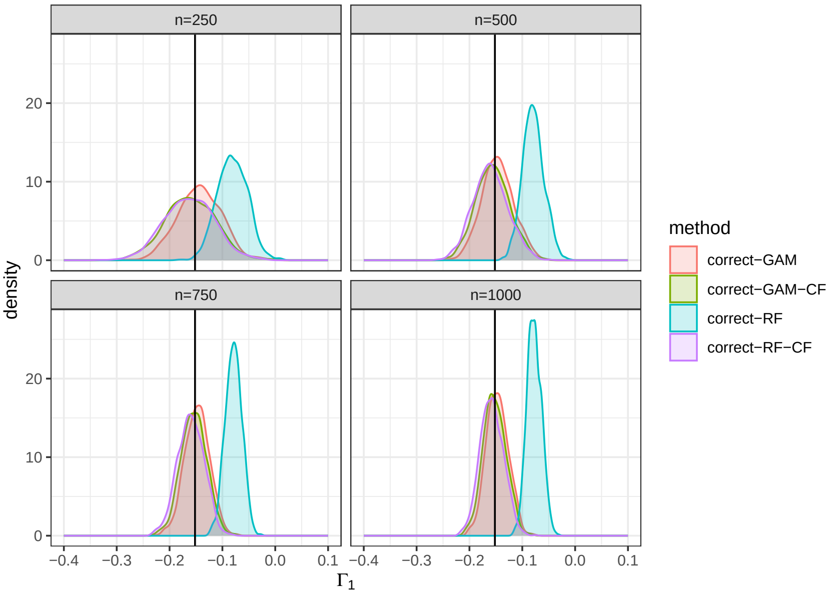
$$\begin{aligned}
&2\hat{\Psi}^{CF}\mathbb{P}_n^k\left(\varphi_1(\hat{\nu}_{1,-k})-\hat{\psi}_1^{CF}\right)\left(\varphi_2(\hat{\nu}_{2,-k})-\hat{\psi}_2^{CF}\right)\\
=&2\hat{\Psi}^{CF}\left(\hat{\psi}_1^{CF}\hat{\psi}_2^{CF}-\hat{\psi}_1^{CF}\mathbb{P}_n^k\varphi_2(\hat{\nu}_{2,-k})-\hat{\psi}_2^{CF}\mathbb{P}_n^k\varphi_1(\hat{\nu}_{1,-k})+\mathbb{P}_n^k\varphi_1(\hat{\nu}_{1,-k})\varphi_2(\hat{\nu}_{2,-k})\right).
\end{aligned}$$

Consistency of the three first terms inside the parenthesis are shown in the in the proof of (14) and hence, we only need to show consistency of the last term. Here, it suffices to show consistency of $\varphi_1(\hat{\nu}_{1,-k})\varphi_2(\hat{\nu}_{2,-k})$ together with $P\varphi_1(\nu_1)\varphi_2(\nu_2)<\infty$, from which the result follows by the same arguments used to show consistency of $\mathbb{P}_n^k\varphi_i(\hat{\nu}_{i,-k})$. For the latter Cauchy-Schwarz and the triangle inequality gives

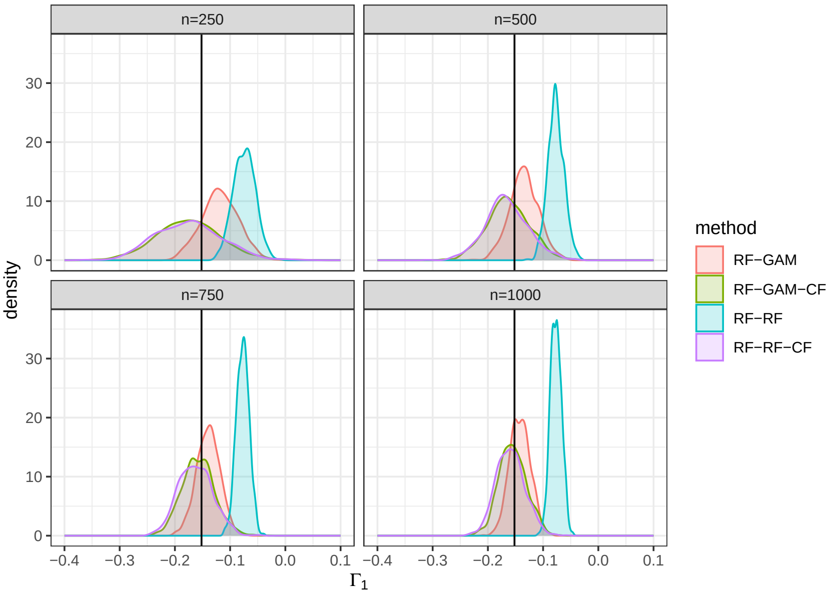
$$P\varphi_1(\nu_1)\varphi_2(\nu_2)\le\left\|\tilde{\psi}_1+\psi_1\right\|\left\|\tilde{\psi}_2+\psi_2\right\|\le\left(\left\|\tilde{\psi}_1\right\|+\|\psi_1\|\right)\left(\left\|\tilde{\psi}_2\right\|+\|\psi_2\|\right)<\infty.$$

By Theorem 18.10 in van der Vaart (2000) $(\varphi_1(\hat{\nu}_{1,-k}),\varphi_2(\hat{\nu}_{2,-k}))\xrightarrow{P}(\varphi_1(\nu_1),\varphi_2(\nu_2))$ since $\varphi_i(\hat{\nu}_{i,-k})\xrightarrow{P}\varphi_i(\nu_i)$ by assumption. Then, the continuous mapping theorem (18.11 in van der Vaart, 2000) gives that $\varphi_1(\hat{\nu}_{1,-k})\varphi_2(\hat{\nu}_{2,-k})\xrightarrow{P}\varphi_1(\nu_1)\varphi(\nu_2)$, and the result follows. □

D Additional simulation results for Ω_1

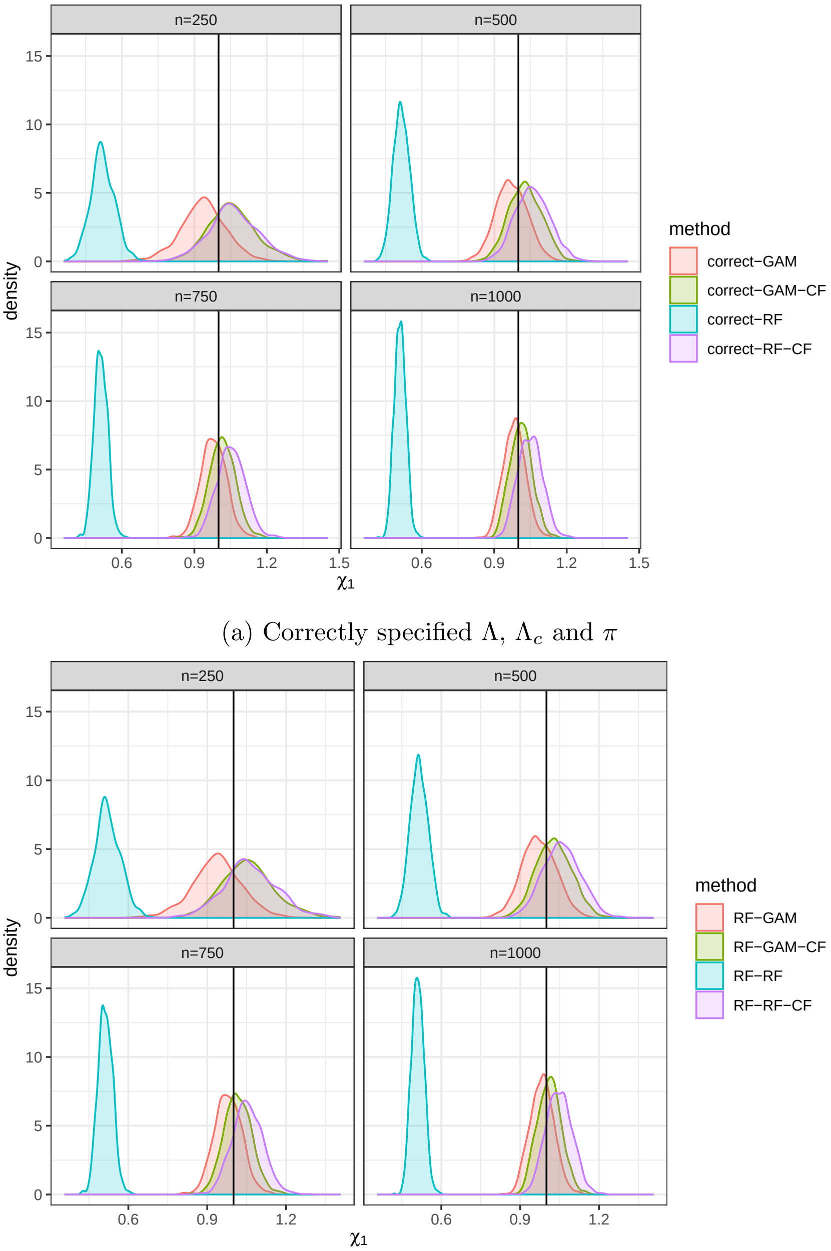


(a) Correctly specified Λ, Λ_c and π



(b) Flexible estimation of Λ, Λ_c and π

Figure 3: Sampling distribution of estimators of Γ_1 in the survival function setting with varying nuisance estimators, with and without cross-fitting, across sample sizes $n = 250, 500, 750, 1000$. The abbreveations of the methods are read as follows: A-B-C, where A corresponds to the nuisance estimators Λ, Λ_c and π, B corresponds to the nuisance estimators $\hat{E}_n$ and $\hat{E}_n^j$, and C corresponds to whether or not cross-fitting was used. Here, *correct* corresponds to correctly specified Cox and logistic regression, RF corresponds to Random Forest, and GAM corresponds to a generalized additive model.



(a) Correctly specified Λ, Λ_c and π

(b) Flexible estimation of Λ, Λ_c and π

Figure 4: Sampling distribution of estimators of χ_1 in the survival function setting with varying nuisance estimators, with and without cross-fitting, across sample sizes $n = 250, 500, 750, 1000$. The abbreveations of the methods are read as follows: A-B-C, where A corresponds to the nuisance estimators Λ, Λ_c and π, B corresponds to the nuisance estimators $\hat{E}_n$ and $\hat{E}_n^j$, and C corresponds to whether or not cross-fitting was used. Here, *correct* corresponds to correctly specified Cox and logistic regression, RF corresponds to Random Forest, and GAM corresponds to a generalized additive model.

Manuscript III

Causal effect on the number of life-years lost due to a specific event - ATE and variable importance

Simon Christoffer Ziersen & Torben Martinussen

Details: In preparation

Causal effect on the number of life years lost due to a specific event: Average treatment effect and variable importance

Simon Christoffer Ziersen & Torben Martinussen

Section of Biostatistics, University of Copenhagen

Abstract

Competing risk is a common phenomenon when dealing with time-to-event outcomes in biostatistical applications. The event of interest may be a certain cause of death and other causes then constitute competing risks, or the event of interest may a be an event different from death, in which case any other absorbing event is a competing event. As the overall survival function now depends on both the hazard of the event of interest and the competing event, it is common to base ones analysis on the cumulative incidence function. Accordingly, in the causal inference literature on time-to-event analysis with competing risk, the difference of cumulative incidence functions is often chosen as a measure of potential treatment effect on the event of interest. Another, estimand of interest when dealing with competing risk is the "number of life-years lost due to a specific cause of death", first described in Andersen (2013). It provides a direct interpretation on the time-scale on which the data is observed. In this paper, we introduce the causal effect of the number of life years lost due to a specific event, and we give assumptions under which the average treatment effect (ATE) and the conditional average treatment effect (CATE) are identified from the observed data. We provide an estimator of the ATE together with an estimator of the best partially linear projection of the CATE as a variable importance measure. The estimators are based on semiparametric efficiency theory and they are agnostic to any model-specifications, thus enabling the use machine learning for nuisance parameter estimation. We present assumptions under which the estimators are asymptotically normal, and their performance are investigated in a simulation study. Lastly, the methods are implemented in a study of the response to different antidepressants using data from the Danish national registers.

1 Introduction

The ATE is a well studied parameter in the causal inference literature. It is typically defined as $\mathrm{E}\{Y^1 - Y^0\}$, where Y^a is the counterfactual outcome under treatment a, that is, the outcome one would have observed had the patient, possibly contrary to the fact, received treatment a. The ATE is often used to asses the effect of an intervention trial. With observational data, the goal is to estimate the ATE under structural causal assumptions from which it can be identified from the observed data as $\mathrm{E}\{E(Y \mid A = 1, X) - E(Y \mid A = 0, X)\}$, where A and X are the observed treatment and covariates, resepctively. The structural assumptions relates to the fact that we do not get to observe the counterfactuals for each individual, but rather a *coarsened* version through the observed outcome and treatment, (Y, A) (van der Laan and Robins, 2003). There is a rich literature on estimation of the ATE, allowing for data-adaptive nuisance estimators, when Y is continuous or binary (see e.g. van der Laan and Rose, 2011, Kennedy, 2022, Chernozhukov et al., 2018) which builds on results from semiparametric efficiency theory (Bickel et al., 1993, van der Vaart, 2000 ch. 25).

In many biostatistical applications, individuals are followed over time and one is interested in the effect of treatment on the time to a given event, say T. It is often the case that only a censored version of the underlying time T is available together with information of whether T was observed. That is, for a censoring time C and an event indicator Δ, the observed outcome is given by $(T \wedge C, \Delta)$. The censoring acts as another form of coarsening in addition to not observing the counterfactuals, and methods from survival analysis are combined with the aforementioned causal inference methodology to define a target parameter that is identifiable from the observed data. This is achieved by setting $Y(t) = f(T;t)$ for some known function f for which identification results for the ATE based on $Y(t)$ exist. Some examples include $f(T;t) = \mathbb{1}(T > t)$, for which the ATE corresponds to the mean difference in survival probability, and $f(T;t) = t \wedge T$, for which the ATE corresponds to the mean difference in the $t-$restricted mean lifetime. Estimation of the ATE in the presence of censoring with data-adaptive nuisance estimators has been studied by several authors (see e.g. Rytgaard et al., 2022, Rytgaard et al., 2023 and Westling et al., 2023).

In some applications, patients may experience a competing event that prevents observation of the event of interest. This is the case when the event of interest is different from some naturally absorbing event, such as death. In that case, patients may die before experiencing the event of interest and death is then a competing event. Letting T_j be the time to the j'th event, we note that the time to the event of interest may not be fully observed since $P(T_j = \infty) > 0$, and direct applications of survival analysis methods are not sufficient. Instead we denote T as the time to the first event and Δ the event indicator, and with censoring, the observed data is now comprised of observations on the form $(\tilde{T}, \tilde{\Delta}, A, X) = (T \wedge C, \mathbb{1}(T \leq C)\Delta, A, X)$. Letting $Y_j^a(t) = \mathbb{1}(T^a \leq t, \Delta^a = j)$ denote the counterfactual outcome, the ATE is now identified in the observed data via the cumulative incidence function, $F_j(t \mid a, x) = P(T \leq t, \Delta = j \mid A = a, X = x)$. The ATE is then interpreted as the mean difference in the absolute risk of the event of interest within a certain time period. Examples of estimation of the ATE based on this identification can be found in e.g. Ozenne et al. (2020), Rytgaard et al. (2023) and Rytgaard and van der Laan (2024). In the three papers, estimation is based on semiparametric efficiency theory, where the first paper considers (semi)parametric working models in a certain type of inverse probability weighted estimator, and the second and third consider estimators based on the efficient influence function with nonparametric nuisance estimators in form of highly adaptive lasso, thus relaxing assumptions on the data-generating distribution.

Without the presence of competing risk, the restricted mean lifetime provides an alternative interpretation to survival probabilities. The ATE estimates are presented on the timescale inherent in the data, which is attractive, as it provides results that are easier to communicate. Another advantage is that the ATE can be written as the mean difference of the integrated survival function over the time-horizon of interest. Thus, if the difference in survival probabilities between the two treatment groups is large in some interval in the time-horizon $[0, t]$ but close to zero at time t, that difference will be reflected in the ATE based on the restricted mean lifetime but not in the ATE based on survival probabilities. An analog of the restricted mean lifetime, that can be used in a competing risk setting, is presented in Andersen (2013) in the form of the *number of life years lost due to a specific cause of death.* It is shown that the number of life years lost due to the j'th event before time t can be written as the integrated cumulative incidence function over the time-horizon $[0, t]$.

In this paper we introduce the ATE based on the number of life years lost due to a specific event. As in the survival setting, the estimand has an interpretation directly on the timescale inherent in the data and it captures possible "early" effects of treatment. We provide an estimator of the ATE based on semiparametric efficiency theory, allowing for the use of machine learning methods for nuisance parameter estimation without relying on model-specifications. We provide assumptions on the nuisance estimators under which the ATE estimator is asymptotically normal and nonparametric locally efficient. The development of the estimator and the assumptions required for inference are akin to those presented in Westling et al. (2023) in the survival setting. Furthermore, we extend the treatment effect variable importance measure given in Ziersen and Martinussen (2024) as a best partially linear projection of the conditional average treatment effect. The projection parameter provides a measure of treatment effect heterogeneity through a given covariate and we provide an estimator analogous to the one presented in Ziersen and Martinussen (2024) in the survival setting. Under additional assumptions on the nuisance parameters, the estimator admits an asymptotic normal distribution which defines a test of treatment effect heterogeneity of a given covariate. The finite sample performance of the proposed estimators are investigated in a simulation study and the estimators are applied to a study on treatment response to difference antidepressants based on data from the Danish national registers (Kessing et al., 2024).

2 Notation and setup

We consider a time-to-event setting with competing risks. Let T be the time to event and $\Delta \in \{1, 2\}$ the event indicator for two competing events. Let X be a d-dimensional vector of covariates, and let A denote the baseline treatment indicator. We enforce censoring through a censoring time C, such that the observed event time is $\tilde{T} = T \wedge C$ and the observed event indicator is $\tilde{\Delta} = \mathbb{1}(C \geq T)\Delta$. Our observed data, $\mathcal{O}$, is given by n i.i.d. copies of $O = (\tilde{T}, \tilde{\Delta}, A, X) \sim P_0$, where $P_0 \in \mathcal{M}$, with $\mathcal{M}$ being a non-parametric model.

We introduce the conditional cause-specific hazard functions, $\lambda_{0,j}(t \mid a, x)$, for the j'th cause, $j = 1, 2$, and let $\lambda_{0,c}(t \mid a, x)$ denote the censoring hazard function. We let $\Lambda_{0,j}(t \mid a, x) = \int_0^t \lambda_{0,j}(s \mid a, x)\,\mathrm{d}s$ and $\Lambda_{0,C}(t \mid a, x) = \int_0^t \lambda_{0,C}(s \mid a, x)\,\mathrm{d}s$ denote the corresponding cumulative hazard functions. We denote $S(t \mid a, x) = \exp\{-\Lambda_1(t \mid a, x) - \Lambda_2(t \mid a, x)\}$ the survival function, and $\pi_0(a \mid x) = P_0(A = a \mid X = x)$ the conditional distribution of A given X. Furthermore, we introduce the cause-specific event times T_j, $j = 1, 2$, and let T_j^a, $a = 0, 1$, denote the counterfactual time corresponding to the j'th cause.

We use the notation $Pf = \int f\,\mathrm{d}P$ and $\mathbb{P}_n f = \sum_{i=1}^n f(O_i)$, and $E_0\{f(O)\} = \int f\,\mathrm{d}P_0$ is the expectation of $f(O)$ under the true data generating distribution. Throughout, all expectations, $P\hat{f}$, considers the function $\hat{f}$ fixed, even when it is estimated from the data, unless otherwise specified. Finally, $\|\cdot\|$ denotes the $L_2(P)$-norm, such that $\|f\| = \left(\int f^2\,\mathrm{d}P\right)^{1/2}$.

3 Causal estimand and nuisance parameters

Inspired by Andersen (2013), we introduce

$$L_0(0, t^*|a, x) = t^* - \int_0^{t^*} S(u|a, x)du$$

for a given time-horizon $[0, t^*]$, which can be interpreted as the expected number of years lost before time t^* in strata (a, x). As Andersen (2013) shows, this quantity can be decomposed naturally into

$$L_0(0, t^*|a, x) = L_1(0, t^*|a, x) + L_2(0, t^*|a, x)$$

where

$$L_j(0, t^*|a, x) = \int_0^{t^*} F_j(u|a, x)du, \quad j = 1, 2,$$

can be interpreted as number of years lost "due to cause j" (Andersen, 2013), with F_j being the jth cumulative incidence function given $A = a$ and $X = x$, i.e. $F_j(t \mid a, x) = \int_0^t S(s \mid a, x) \, d\Lambda_j(s \mid a, x)$. To introduce the counterfactual number of life years lost due to a specific event, we first remark on an observation given in Andersen (2013). The random variable T_j is improper because $P(T_j = \infty) > 0$, but the random variable $T_j \wedge t^*$ is proper with expectation given by

$$E\{T_j \wedge t^*\} = t^* - \int_0^{t^*} F_j(s) \, ds,$$

and hence

$$E(T_j \wedge t^* \mid a, x) = t^* - \int_0^{t^*} F_j(s \mid a, x) \, ds.$$

We now introduce the counterfactual $Y_j^a(t^*) = t^* - T_j^a \wedge t^*$ for $a = 0, 1$, which is the number of life-years lost due to event j before time t^* under treatment a. We define the j'th-specific ATE as

$$E_0\{Y_j^1(t^*) - Y_j^0(t^*)\}$$

and the CATE is

$$E_0(Y_j^1(t^*) - Y_j^0(t^*) \mid X = x).$$

In order to identify the ATE and CATE from the observed data we need the following assumptions:

Assumption A (Identification)**.**

A1 (Consistency) $Y_j(t^*) = t^* - T_j \wedge t^* = AY_j^1(t^*) + (1 - A)Y_j^0(t^*)$ *conditional on* A.

A2 (Exchangeability) $Y_j^a(t^*) \perp\!\!\!\perp A \mid X, \; A = 0, 1$.

A3 (Positivity) $\pi(a \mid X = x)P(C > t \mid a, x)P(T > t \mid a, x) > \eta > 0, \; \forall(t, x) \in [0, t^*] \times \mathcal{X}, \; a = 0, 1$.

A4 (Independent censoring) $T \perp\!\!\!\perp C \mid A, X$.

Define

$$\tau_j(x;t^*) \equiv L_j(0,t^*|1,x) - L_j(0,t^*|0,x), \quad j=1,2.$$

In Appendix A, we show that the CATE function is identified in the observed data because

$$\tau_j(x;t^*) = E_0(Y_j^1(t^*) - Y_j^0(t^*) \mid X = x)$$

and the average treatment effect as

$$E_0\{\tau(X;t^*)\} = E_0\{Y_j^1(t^*) - Y_j^0(t^*)\}$$

under assumption A. Going forward we drop the dependence of t^* and write $\tau(x) = \tau(x;t^*)$ to ease notation. Based on the identification results, we define two target parameters as mappings from the model $\mathcal{M}$ on the observed data to the reals. The first parameter is the j-specific average treatment effect, defined as the mapping $\psi_j : \mathcal{M} \to \mathbb{R}$, where

$$\psi_j(P) = E\{L_j(0,t^*|1,X) - L_j(0,t^*|0,X)\}.$$

The second parameter is defined as a variable importance measure of the $l'th$ covariate on $\tau_j(x)$ based on the best partially linear projection given in Ziersen and Martinussen (2024). It is defined as the mapping $\Omega_j^l : \mathcal{M} \to \mathbb{R}$ with

$$\Omega_j^l(P) = \frac{E\{\mathrm{cov}(X_l, \tau_j(X) \mid X_{-l})\}}{E\{\mathrm{var}(X_l \mid X_{-l})\}},$$

where X_{-l} denotes the covariates indexed by $\{1,\ldots,d\} \setminus \{l\}$. The parameter Ω_j^l can be viewed as a weighted average of the conditional covariance of the CATE and the covariate X_l given the rest of the covariates. The parameter is sensitive to the scale of the covariate in question, and when assessing the importance of different covariates it is not the estimate of the parameter that determines the ranking of variable importance, but rather the p-value associated with test $H : \Omega_j^l = 0$, since Ω_j^l is zero if there is no heterogeneity explained by X_l. The parameter can be derived as the least-squares projection of the CATE onto the partially linear model. For more details and discussion, see Ziersen and Martinussen (2024).

4 Estimation and inference

We base the estimation of the two target parameters, $\psi_j(P)$ and $\Omega_j^l(P)$, on semiparametric efficiency theory (Bickel et al., 1993, van der Vaart, 2000 ch. 25, van der Laan and Rose, 2011).

For a general target parameter ψ, an estimator $\hat{\psi}$ is said to be asymptotically linear if it can be written on the form $\hat{\psi} - \psi = \mathbb{P}_n\mathbb{IF} + o_p(n^{-1/2})$ with $P\mathbb{IF} = 0$. The function $\mathbb{IF}$ is called the influence function of the estimator $\hat{\psi}$, and it characterizes the asymptotic distribution of the estimator. This can be seen by applying the central limit theory together with Slutsky's lemma, from which $\sqrt{n}(\hat{\psi} - \psi) \xrightarrow{D} \mathcal{N}(0, P\mathbb{IF}^2)$. If the target parameter is differentiable at P as a map $\psi : \mathcal{M} \to \mathbb{R}$, there exist a unique function, say $\tilde{\psi}$, associated to the pathwise derivative of the parameter, which characterizes the information bound of any regular estimator. The function $\tilde{\psi}$ is called the efficient influence function (EIF), and any estimator is regular and asymptotically efficient, if it is asymptotically linear with influence function given by the EIF (van der Vaart, 2000 ch. 25.3).

Since the EIF is uniquely determined by the target parameter it can be calculated without reference to any estimator. Once it is known, several techniques exist for constructing estimators that are asymptotically linear with the EIF as their influence function (van der Laan and Rose, 2011, Chernozhukov et al., 2018, Kennedy, 2022, Hines, Dukes, et al., 2022). We focus on the so-called *one-step estimator*, which is defined as

$$\hat{\psi}^{OS} = \psi(\hat{P}) + \mathbb{P}_n\tilde{\psi}(\cdot;\hat{P}),$$

where $\hat{P}$ is obtained from some (possibly) data-adaptive estimators. In order to show that the one-step estimator is asymptotically linear, one typically relies on a certain decomposition involving an empirical process term and a remainder term, which are both required to be $o_p(n^{-1/2})$. The former can be obtained if $\hat{P}$ is assumed to belong to a Donsker class, but this requirement has been shown to be too restrictive for some data-adaptive estimators, and a certain type of sample splitting, termed *cross-fitting*, has to be applied to the one-step estimator in order to obtain the required convergence rate (Chernozhukov et al., 2018, Kennedy, 2022).

4.1 Average treatment effect

To derive an estimator for the ATE we first derive its EIF. We define the nuisance parameter $\nu = (\Lambda_1, \Lambda_2, \Lambda_c, \pi)$. The EIF is then given in the following lemma.

Lemma 1. *The efficient influence function of $\psi_j(P)$ is given by*

$$\tilde{\psi}_{\psi_j}(O;\nu) = \varphi_j(\nu)(O) - \psi_j(P),$$

where $\varphi_j(\nu)$ is a real-valued function defined on the sample space of O at a given value of ν with

$$\varphi_j(\nu)(O) = \tau_j(X) + \left(\frac{\mathbb{1}(A=1)}{\pi(1 \mid X)} - \frac{\mathbb{1}(A=0)}{\pi(0 \mid X)}\right)\left\{\sum_{i=1,2}\int_0^{t^*}\frac{H_{ij}(s,t^* \mid A,X)}{S_C(s \mid A,X)}\,\mathrm{d}M_i(s \mid A,X)\right\} \tag{1}$$

where

$$H_{ij}(s,t \mid a,x) = \int_s^t \mathbb{1}(i=j) + \frac{F_j(s \mid a,x) - F_j(u \mid a,x)}{S(s \mid a,x)}\,\mathrm{d}u. \tag{2}$$

Proof. See Appendix B. □

The function $\varphi_j(\nu)$ is the uncentered EIF of the ATE and will also appear in the development of an estimator for the best partially linear projection in Section 4.2. As the EIF of the ATE is linear in the target parameter, the one-step estimator for ψ_j reduces to

$$\hat{\psi}_j^{OS} = \mathbb{P}_n\varphi_j(\hat{\nu}).$$

As noted earlier, the one-step estimator may fail to be asymptotically linear when using data-adaptive estimators for $\hat{\nu}$ and we further have to use a cross-fitted version of the estimator in order to obtain the desired asymptotic properties.

To define the sample splitting involved in constructing the cross-fitted estimator, let $\boldsymbol{i} = (i_1, i_2, \dots, i_n)$ be an index vector drawn from an n-dimensional multinomial distribution with K events with probability $p_k = \frac{1}{K}$, $k = 1, \dots, K$, for the k'th event. Define the index sets $\mathcal{T}_k = \{j : i_j = k\}$ for $k = 1 \dots K$ such that $\{1, \dots, n\} = \dot{\cup}_{k=1}^{K} \mathcal{T}_k$, where $\dot{\cup}$ denotes the disjoint union. Corresponding to the index sets, we define K disjoint data splits by $\mathcal{V}_k = \{O_j : i_j = k\}$ such that $\mathcal{O} = \dot{\cup}_{k=1}^{K} \mathcal{V}_k$, and we define K leave-out data splits by $\mathcal{V}_{-k} = \dot{\cup}_{j \neq k} \mathcal{V}_j$.

To construct a cross-fitted one-step estimator of ψ_j based on the EIF, let $\hat{\nu} = (\hat{\Lambda}_1, \hat{\Lambda}_2, \hat{\Lambda}_c, \hat{\pi})$ denote the estimated nuisance parameter, and let $\hat{\nu}_{-k}$ be the estimated nuisance parameter based on data in the k'th leave out sample, $\mathcal{V}_{-k}$, and let $\mathbb{P}_n^k$ be the empirical measure of $O \in \mathcal{V}_k$. We define the K-fold cross-fitted estimator of ψ_j as

$$\hat{\psi}_j^{CF} = \sum_{k=1}^{K} \frac{n_k}{n} \mathbb{P}_n^k \varphi(\hat{\nu}_{-k}) = \frac{1}{n} \sum_{k=1}^{K} \sum_{i \in \mathcal{T}_k} \varphi_j(\hat{\nu}_{-k})(O_i), \tag{3}$$

where n_k is the number of observations in the k'th data split. The construction of the cross-fitted estimator detailed here is quite general, see Kennedy (2022) for a nice discussion and comparison to the one-step estimator without cross-fitting.

In order to derive asymptotic results for $\hat{\psi}_j^{CF}$ we need a set of assumptions on the nuisance estimators. The assumptions below are stated for a general nuisance estimator $\hat{\nu}$, but in applications to cross-fitted estimators, they are assumed to hold for each leave-out sample $\mathcal{V}_{-k}$. We will be explicit about this when stating results on the obtained estimators, but it is left out of assumption B for notational convenience.

Assumption B. *The nuisance estimator $\hat{\nu}$ satisfy the following conditions*

B1 There exist a real-valued parameter $\eta > 0$ such that $\hat{S}(t \mid a, x) > \eta$, $S(t \mid a, x) > \eta$, $\hat{S}_C(t \mid a, x) > \eta$, $S_C(t \mid a, x) > \eta$, $\hat{\pi}(a \mid x) > \eta$, $\pi(a \mid x) > \eta$ for all $(t, a, x) \in [0, t^*] \times \{0, 1\} \times \mathcal{X}$.

B2 For $a = 0, 1$, it holds that

$$\mathrm{E}\left[\sup_{s \le t^*} \left|\hat{\Lambda}_1(s \mid a, X) - \Lambda_{1,0}(s \mid a, X)\right|\right]^2 = o_p(1)$$
$$\mathrm{E}\left[\sup_{s \le t^*} \left|\hat{\Lambda}_2(s \mid a, X) - \Lambda_{2,0}(s \mid a, X)\right|\right]^2 = o_p(1)$$
$$\mathrm{E}\left[\sup_{s \le t^*} \left|\hat{\Lambda}_c(s \mid a, X) - \Lambda_{c,0}(s \mid a, X)\right|\right]^2 = o_p(1)$$
$$\mathrm{E}\left[\hat{\pi}(a \mid X) - \pi_0(a \mid X)\right]^2 = o_p(1)$$

B3 For $a = 0, 1$, it holds that

$$\mathrm{E}\left\{\sum_{i=1,2} \int_0^{t^*} S(s \mid a, X) \hat{H}_{ij}(s, t^* \mid a, X) \right.$$
$$\left. \times \left(1 - \frac{\pi(a \mid X) S_C(s \mid a, X)}{\hat{\pi}(a \mid X) \hat{S}_C(s \mid a, X)}\right) \mathrm{d}\left[\hat{\Lambda}_i(s \mid a, X) - \Lambda_i(s \mid a, X)\right]\right\} = o_p(n^{-1/2}).$$

Assumption B1 is the usual positivity assumption found through out the causal inference literature (for examples with censored data, see e.g. Westling et al., 2023, Rytgaard et al., 2023, Rytgaard and van der Laan, 2024). Whereas assumption A3 relates to the true data generating mechanism, assumption B1 extends to the estimators as well. We note that employing cross-fitting in relatively small sample sizes can sometimes lead to practical positivity-violations, when a "rare" covariate lies in $\mathcal{V}_k$ but not in $\mathcal{V}_{-k}$. For B2 to hold, all nuisance estimators must be consistent, and for the hazard estimators this amounts to uniform consistency. This requirement suggest the use of flexible learners for nuisance estimation. Assumption B3 reflects the dobule robustness that is common for one-step estimators. In lemma C.1 in the Appendix it is shown that the so-called remainder term, coming from the aforementioned decomposition of the estimator, takes this form. It is sometimes referred to as a second order remainder term, when it can be shown to hold if each of the nuisance estimators converge on $n^{-1/4}$-rate in $L_2(P)$-norm. If one considers cumulative hazard estimators that are absolute continuous, the result can be obtained from a simple application of the Cauchy-Schwarz inequality (for examples involving Highly Adaptive Lasso, see Munch et al., 2024, Rytgaard et al., 2023, Rytgaard et al., 2022), but this exclude many commonly used cumulative hazard estimators such as any Breslow-type estimators. We expect nonetheless that the double robustness is obtained for most reasonable estimators, and in the later simulation studies, this will be exemplified by the use of random survival forests (Ishwaran et al., 2008).

Next follows our main result for the cross-fitted ATE estimator:

Theorem 1. *Assume that the nuisance estimators $\hat{\nu}_{-k}$, $k = 1 \dots K$ follow assumption B for each k. Then the cross-fitted estimator is asymptotically linear with influence function given by $\tilde{\psi}_{\psi_j}$ and hence*

$$\sqrt{n}\left(\hat{\psi}_j^{CF} - \psi_j\right) \xrightarrow{d} \mathcal{N}(0, P\tilde{\psi}_{\psi_j}(\cdot, \nu_0)^2).$$

Proof. See Appendix C. □

In practise, the variance of the influence function is estimated by the cross-fitted estimator:

$$\hat{\sigma}_{\psi_j}^{2,CF} = \sum_{k=1}^{K} \frac{n_k}{n} \mathbb{P}_n^k \left(\varphi(\hat{\nu}_{-k}) - \hat{\psi}_j^{CF}\right)^2,$$

and the standard error of $\hat{\psi}_j^{CF}$ is given by $\sqrt{\hat{\sigma}_{\psi_j}^{2,CF}/n}$. We note that $\hat{\sigma}_{\psi_j}^{2,CF}$ is a type of plug-in estimator, and contrary to the one-step estimator, it is not debiased via its influence function. Hence, even though the estimator is consistent (by Lemma 1 in Ziersen and Martinussen, 2024), for consistent nuisance estimators, it is not generally asymptotically linear for data-adaptive nuisance estimators, which may result in biased standard error estimates in finite samples. In the simulation studies in Section 5, this will be explored by contrasting the use of (semi)parametric and data-adaptive nuisance estimators, with the latter obtained from random forests.

4.2 Best partially linear projection

Estimation of $\Omega_j^l(P)$ follows the same overall strategy, i.e., construct an asymptotically linear estimator using semiparametric theory. The difference now is that the target parameter is a ratio of two parameters which can both be written as a map from $\mathcal{M}$ to the reals. If we construct asymptotically linear estimators for both parameters in the ratio, separately, then the ratio of the estimators will also be asymptotically linear. Furthermore, if each estimator in the ratio has their respective EIF as their influence function, the ratio of the estimators will have its EIF as its influence, by the functional delta method (van der Vaart, 2000 ch. 25.7). In the following we will extend the approach of Ziersen and Martinussen (2024) to the CATE function defined by the number of life years lost, $\tau_j(x)$, for a given time-horizon $[0, t^*]$, where each of the parameters in the ratio of Ω_j^l is estimated separately. We start by calculating the EIF for the relavant parameters:

Lemma 2. *Define the mappings* $\Gamma_j^l : \mathcal{M} \to \mathbb{R}$ *and* $\chi_j^l : \mathcal{M} \to \mathbb{R}$ *as*

$$\Gamma_j^l(P) = E\{\text{cov}(X_l, \tau_j(X) \mid X_{-l})\}$$

and

$$\chi^l(P) = E\{\text{var}(X_l \mid X_{-l})\}$$

such that $\Omega_j^l(P) = \frac{\Gamma_j^l(P)}{\chi^l(P)}$. *The efficient influence functions of* $\Gamma_j^l(P)$, $\chi^l(P)$ *and* $\Omega_j^l(P)$, *respectively, are given by*

$$\tilde{\psi}_{\Gamma_j^l}(O; P) = [\varphi_j(\nu)(O) - \mathrm{E}(\tau_j(X) \mid X_{-l})][X_l - \mathrm{E}(X_l \mid X_{-l})] - \Gamma_j^l(P), \tag{4}$$

$$\tilde{\psi}_{\chi^l}(O; P) = [X_l - \mathrm{E}(X_l \mid X_{-l})]^2 - \chi^l(P), \tag{5}$$

$$\tilde{\psi}_{\Omega_j^l}(O; P) = \frac{1}{\chi^l(P)} \left(\tilde{\psi}_{\Gamma_j^l}(O; P) - \Omega_j^l(P)\tilde{\psi}_{\chi^l}(O; P) \right). \tag{6}$$

Proof. See Appendix B □

The EIF's above depend explicitly on the conditional distribution of X_l given X_{-l} through $E(X_l \mid X_{-l})$ and $E(\tau_j(X) \mid X_{-l})$, so to express them as mappings of the nuisance parameter, we extend the notion of ν. Let $\tau_j^l(x_{-l}) = E(\tau_j(X) \mid X_{-l} = x_{-l})$ and $E^l(x_{-l}) = E(X_l \mid X_{-l} = x_{-l})$, and define $\nu_l^1 = E^l$ and $\nu_l^2 = (\Lambda_1, \Lambda_2, \Lambda_c, \pi, \tau_j^l, E^l)$. We define the uncentered EIF's corresponding to the EIF's $\tilde{\psi}_{\Gamma_j^l}$ and $\tilde{\psi}_{\chi_j^l}$ as

$$\phi_{\Gamma_j^l}(O; \nu_l^2) = [\varphi_j(\nu)(O) - \tau_j^l(X_{-l})][X_l - E^l(X_{-l})] \tag{7}$$

$$\phi_{\chi^l}(O; \nu_l^1) = [X_l - E^l(X_{-l})]^2. \tag{8}$$

The construction of the estimators for Γ_j^l and χ^l now follow similar to the procedure in the ATE setting. The estimation of χ^l is given in Ziersen and Martinussen (2024), but is included here as well for completeness. Let $\hat{\nu}_l^1$ and $\hat{\nu}_l^2$ denote the estimated nuisance parameters. As in the ATE-setting, we define $\hat{\nu}_{l,-k}^1$ and $\hat{\nu}_{l,-k}^2$ as the nuisance estimators based on data in $\mathcal{V}_{-k}$. The cross-fitted estimators are defined as

$$\hat{\Gamma}_j^{l,CF} = \frac{1}{n} \sum_{k=1}^{K} \sum_{i \in \mathcal{T}_k} \phi_{\Gamma_j^l}(O_i; \hat{\nu}_{l,-k}^2), \quad \hat{\chi}^{l,CF} = \frac{1}{n} \sum_{k=1}^{K} \sum_{i \in \mathcal{T}_k} \phi_{\chi^l}(O_i; \hat{\nu}_{l,-k}^1), \quad \hat{\Omega}_j^{l,CF} = \frac{\hat{\Gamma}_j^{l,CF}}{\hat{\chi}^{l,CF}}$$

Since the above estimators depend on the extended nuisance estimators, we have to make additional assumption in order to derive the desired asymptotic linearity. Accordingly, we have the following result:

Theorem 2. *Assume that for each fold $k = 1, \dots, K$ it holds that*

(i) $\left(X_l - \hat{E}^l\right)^2 \leq M,$ *a.s for all n and some $M > 0$.*

(ii) $\left\|\hat{\tau}^l_j - \tau^l_j\right\| = o_p(n^{-1/4})$.

(iii) $\left\|\hat{E}^l - E^l\right\| = o_p(n^{-1/4})$.

Then, if assumption B holds for each k, it follows that $\hat{\Omega}^{l,CF}_j$ is asymptotically linear with influence function given by $\tilde{\psi}_{\Omega^l_j}$ and hence

$$\sqrt{n}(\hat{\Omega}^{l,CF}_j - \Omega^l_j) \xrightarrow{d} \mathcal{N}(0, P\tilde{\psi}^2_{\Omega^l_j}).$$

Proof. See Appendix C. □

Assumption $(i) - (iii)$ refer to the nuisance estimators related to the conditional distribution of X_l given X_{-l}. Regarding assumption B3 for ATE estimation, we discussed the double robustness properties of the cross-fitted estimator in relation to the convergence rates of the nuisance estimators, and we can add to that discussion the rates given in (ii) and (iii). We see that the estimator for our target parameter achieves parametric rates (asymptotic linearity) if the nuisance estimators related to $X_l|X_{-l}$ are estimated at $n^{-1/4}$-rate, adding to the notion of "double robustness". The rate in assumption (iii) is known for many estimators, as it is an assumption on a typical regression estimator. Whether it is fulfilled depends on the type of the estimator used and possibly on the dimension d in relation to n, but we note that the assumption is to be considered rather mild, allowing for many types of data-adaptive estimators (see e.g. the discussion in Kennedy, 2022, Section 4.3). For estimation of $\hat{\tau}^l_j$, we regress the CATE estimates $(\hat{\tau}_j(X_i))^n_{i=1}$ onto $X_{-l} = (X_{i,-l})^n_{i=1}$ in line with the approach suggested in Hines, Diaz-Ordaz, and Vansteelandt (2022), and Ziersen and Martinussen (2024). This approach constitutes a certain type of meta-learning and convergence rates related to (ii) are generally less known compared to the regression in assumption (iii). We refer to Hines, Diaz-Ordaz, and Vansteelandt (2022) for a discussion of a specific meta-learner termed the DR-learner (Kennedy, 2023) for estimation of $\hat{\tau}^l_j$ (their analogy is termed $\hat{\tau}_s$) and convergence rates analogous to (ii).

As in the ATE setting, the variance, $P\tilde{\psi}^2_{\Omega^l_j}$, is estimated by the cross-fitted plugin estimator:

$$\hat{\sigma}^{2,CF}_{\Omega^l_j} = \sum_{k=1}^{K} \frac{n_k}{n} \mathbb{P}^k_n \tilde{\psi}_{\Omega^l_j}(\hat{\nu}^2_{l,-k})^2,$$

where (with some abuse of notation) we define

$$\tilde{\psi}_{\Omega^l_j}(\hat{\nu}^2_{l,-k})(O) = \frac{1}{\hat{\chi}^{l,CF}} \left(\phi_{\Gamma^l_j}(O; \hat{\nu}^2_{l,-k}) - \hat{\Gamma}^{l,CF}_j - \hat{\Omega}^{l,CF}_j \left(\phi_{\chi^l}(O; \hat{\nu}^1_{l,-k}) - \hat{\chi}^{l,CF}\right)\right).$$

Because of scale sensitivity, the estimate of Ω_j^l may be of less interest than testing the null-hypothesis $H_0 : \Omega_j^l = 0$. A test statistic for H_0 can be defined by

$$\text{TST}_1^l \equiv \frac{\hat{\Omega}_j^{l,CF}}{\sqrt{\hat{\sigma}^{2,CF}_{\Omega_j^l}/n}}$$

that is asymptotically standard normal distributed. Lemma 1 in Ziersen and Martinussen (2024) shows that the cross-fitted variance estimators considered in this paper, i.e., $\hat{\sigma}^{2,CF}_{\Omega_j^l}$ and $\hat{\sigma}^{2,CF}_{\psi_j}$ are consistent. The following corollary (analogous to Corollary 4 in Ziersen and Martinussen, 2024) gives the desired asymptotic properties of our test-statistic:

Corollary 1. *Under the same setup as in Theorem 2, we have under the null-hypothesis, $H_0 : \Omega_j^l = 0$, that*

$$TST_1^l \xrightarrow{D} \mathcal{N}(0,1).$$

Proof. Since the variance estimator $\hat{\sigma}^{2,CF}_{\Omega_j^l}$ is consistent, Theorem 2 together with Slutsky's theorem and an application of the delta method gives the result. □

5 Simulation study

We conduct simulation studies to investigate the proposed asymptotic properties of the estimators $\hat{\psi}_j^{CF}$ and $\hat{\Omega}_j^{l,CF}$ under two different nuisance estimator settings with and without cross-fitting (i.e. setting $K = 1$). For all cross-fitted estimators, we set $K = 10$. In one nuisance estimator setting we consider correctly specified (semi)parametric nuisance estimators and in the other we use completely nonparametric estimators via random forest. The parametric nuisance estimators adhere to assumption B and so we would expect the target parameter estimators to perform according to theory both with and without cross-fitting. In case of nonparametric estimators, survival random forest are shown to adhere to assumption B2 in Cui et al. (2022), but it is unclear to what extend they admit rates corresponding to B3. Furthermore, the nonparametric estimators do not in general belong to a Donsker class, and we therefore expect the cross-fitted estimators to perform more in line with the theory compared to the non-cross-fitted version (see Chernozhukov et al., 2018 and Kennedy, 2022 for a discussion on cross-fitted one-step estimators).

We consider data generated from the following models:

- $X_l \sim \text{Unif}[-1,1],\ l = 1,\dots,4$
- $\pi(1 \mid X) = \text{expit}(0.5X_1 + 0.5X_2)$
- $\lambda_1(t \mid A, X) = 0.0025 \cdot 2t^{2-1}\exp(-X_1 - X_2 - 0.2X_3 + A(0.5X_1 - 0.3X_2 - 2))$
- $\lambda_2(t \mid A, X) = 0.00025 \cdot 2t^{2-1}\exp(-X_1 - X_2 - 0.2X_3 + A)$
- $\lambda_c(t \mid A, X) = 0.00025 \cdot 2t^{2-1}\exp(-0.5X_1)$.

Note that the above hazard functions correspond to Cox models with baseline hazards given by a Weibull hazard, bkt^{k-1}, where b and k are the scale and shape parameter, respectively. We consider four sample size settings of $n = 250, 500, 750, 1000$, and for each setting we run 1000 simulations. For each simulation we generate data according to the models above and estimate the target parameters according to specifications given below.

5.1 Average treatment effect

For estimation of $\hat{\psi}_1^{CF}$ we choose the time-horsizon $t^* = 30$. The true ATE is approximately $\psi_1(P_0) = -9.6135$. We consider two nuisance setting for estimation of $\hat{\nu}$ (here dropping k from the notation). In one setting $\hat{\nu}$ consists of correctly specified Cox models with corresponding Breslow estimators for the cumulative hazards, and a correctly specified logistic regression for the propensity score model. This setting will be abbreviated **cor** in tables and figures going forward. In the second setting we estimate the cumulative hazards by random survival forests (Ishwaran et al., 2008) as implemented in the R-package `randomForestSRC` with default tuning parameters (see Ishwaran et al., 2023 for documentation), and we estimate the propensity score by random forest, again with the implementation and tuning parameters given by `randomForestSRC`. This setting will be abreviated **RF** in tables and figures going forward. Furthermore, each setting will be given the suffix **CF** if cross-fitting is used.

n	method	bias	SD	mean SE	coverage
250	**cor**	-0.1521	0.9424	0.9746	0.9550
	corCF	-0.0494	0.9701	1.0340	0.9620
	RF	-0.3501	0.9020	0.5644	0.7520
	RFCF	0.0953	1.2012	1.3429	0.9710
500	**cor**	-0.0980	0.6946	0.6893	0.9570
	corCF	-0.0538	0.7040	0.7084	0.9550
	RF	-0.3576	0.6750	0.3957	0.6800
	RFCF	-0.0383	0.7912	0.8799	0.9680
750	**cor**	-0.0665	0.5650	0.5633	0.9480
	corCF	-0.0377	0.5684	0.5735	0.9500
	RF	-0.3269	0.5538	0.3237	0.6750
	RFCF	-0.0678	0.6394	0.6970	0.9670
1000	**cor**	-0.0265	0.4724	0.4884	0.9540
	corCF	-0.0046	0.4754	0.4949	0.9570
	RF	-0.2800	0.4636	0.2808	0.6950
	RFCF	-0.0319	0.5383	0.5943	0.9710

Table 1: Results of 1000 simulations of estimators of Ψ_1 with varying nuisance estimators, and with and without cross-fitting for sample sizes $n = 250, 500, 750, 1000$. The abbreveations of the methods are read as follows: **cor** corresponds to the nuisance parameters $\Lambda_1, \Lambda_2, \Lambda_c, \pi$ estimated by correctly specified Cox and logistic regressions, and **RF** corresponds to the same parameters estimated by Random Forest. A suffix **CF** indicates that cross-fitting was employed in estimation of Ψ_j. The tables gives the bias, empirical standard deviation (SD), mean of the estimated standard error (mean SE), and coverage.

Table 1 gives the results for ATE-estimation. For correctly specified nuisance estimators the bias decreases with n and standard deviations decrease at an approximately $n^{1/2}$-rate. With a coverage around 0.95, even for relatively small n, it looks as if the estimator follows the asymptotic distribution from Theorem 1. Surprisingly though, cross-fitting seems to decrease the bias of the estimator with correctly specified nuisance parameters even further. Overall we find that the estimator performs as expected for correctly specified (semi)parametric nuisance estimators.

For the nuisance estimators using random forests, we see a non-vanishing bias for the non-cross-fitted estimators. Furthermore, the standard errors are underestimated compared to the empirical standard deviation of the estimators which, together with the bias result in undercoverage. When using cross-fitting together with random forests, the bias disappears on roughly the same order as the correctly specified parametric estimators (without cross-fitting), and the standard deviation of the estimator seem to be on roughly $\sqrt{n}$-rate, as with the correctly specified nuisance parameters. The standard errors seem to be slighty overestimated, though, resulting in a slight overcoverage. This might be due to the hyperparameter choices for the random forests.

5.2 Best partially linear projection

For estimation of $\hat{\Omega}_1^{l,CF}$, we set $t^* = 30$, as for ATE-estimation. The true value of the target parameters are approximately $(\Omega_1^1, \Omega_1^2, \Omega_1^3, \Omega_1^4) = (4.949, 3.137, 0.737, 0)$. We also need estimates of $\hat{\tau}^l$ and $\hat{E}^l$, for which we will consider two settings. In the first, both $\hat{\tau}^l$ and $\hat{E}^l$ are estimated with a generalised additive model (GAM) including spline smoothing of each term but without interactions as implemented in the `R`-package `mgcv`, and in the second, each of the nuisance parameters are estimated with random forest (again with default tuning parameters from `randomForestSRC`). The GAM setting will be added to the correctly specified setting, **cor**, from earlier in tables and figures going forward, and the random forest setting will be added to the random forest setting, **RF**, from earlier.

In Figure 1, we see the absolute bias for estimation of $\hat{\Omega}_j^l$, $l = 1, \ldots, 4$, for the different nuisance settings. In general, we see that **cor** and **corCF** perform similarly across all sample sizes and across all l, with a bias converging to zero. The **RFCF**-setting performs slightly worse than **cor** and **corCF** for small n, but has approximately similar performance for large n, whereas **RF** has a large bias for large enough values of Ω_1^l. Generally, the estimators seem to perform as we would expect according to Theorem 2 in terms of bias. The coverage of the estimators are presented in Figure 2. The settings **cor**, **corCF** and **RFCF** all exhibit approximately nominal coverage across n and l, whereas **RF** has poor coverage. Again, this is in line with our expectations. Figure 3 gives the estimated probability of rejecting $H : \Omega_1^4 = 0$, i.e. the type-1 error (since $\Omega_1^4 = 0$ in our data generating mechanism), together with Monte Carlo confidence intervals. The type-1 error is approximately 0.05 for all n, expect for **RF** where the type 1-error increases with n.

Lastly, Figure 4 shows the probability of rejecting $H : \Omega_1^l = 0$, $l = 1, 2, 3$, which correspond the power of the test. Interestingly, using data-adaptive estimation of the nuisance parameters seem to decrease the power of the test TST_1^l.

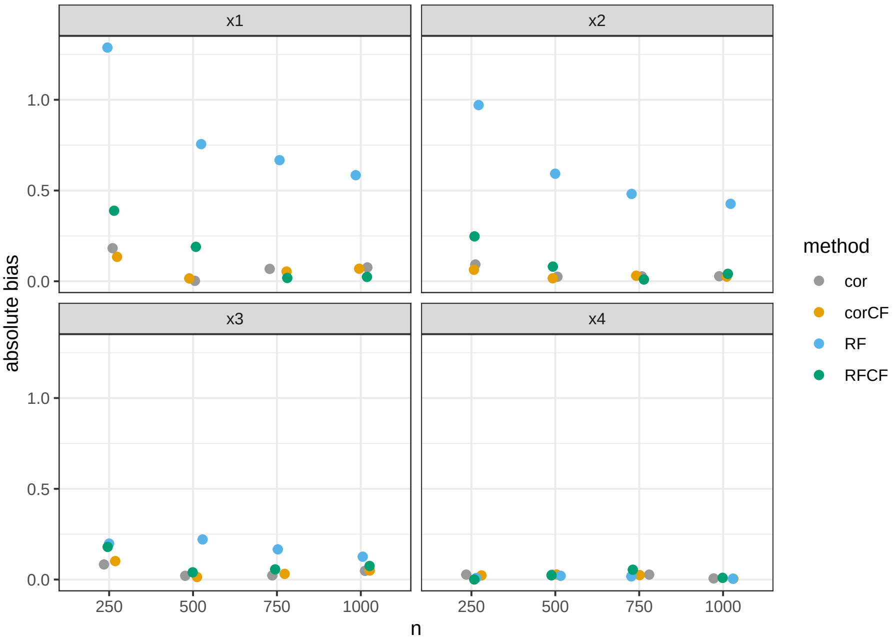


Figure 1: Results based on 1000 simulations of the estimators of Ω_1^l with for $l = 1, \ldots, 4$ with varying nuisance estimators and across sample sizes $n = 250, 500, 750, 1000$. The plot shows the absolute bias of the estimators, where **cor** corresponds to the nuisance parameters $\Lambda_1, \Lambda_2, \Lambda_c, \pi$ estimated by correctly specified Cox and logistic regressions, and **RF** corresponds to the same parameters estimated by Random Forest. A suffix **CF** indicates that cross-fitting was employed. The true values are $(\Omega_1^1, \Omega_1^2, \Omega_1^3, \Omega_1^4) = (4.949, 3.137, 0.737, 0)$.

6 Application

To demonstrate the methods outlined in the previous sections, we consider the study by Kessing et al. (2024). In this study, the non-response to 17 different antidepressants are compared based on data from the Danish national registers. Patients enter the study at their first diagnosis with major depressive disorder from a psychiatric hospital. Their treatment, in terms of a specific antidepressant, is defined as the first purchase of an antidepressant after discharge from the hospital, which also determines the index date. The main outcome was time to non-response, defined as a switch to or add-on of another antidepressant or antipsychotic medicine or readmission to psychiatric hospital with a major depressive disorder. Competing risk for the time to non-response was admission to a psychiatric hospital with a higher order psychiatric diagnosis (bipolar disorder, schizophrenia or organic mental disorder) or death. The 17 antidepressants are categorised into six groups (SSRI, NARI, SNRI, NaSSA, TCA, and others) and within each group a reference drug is chosen to which the other drugs in that group are compared. The estimand for each comparison is the average treatment effect on the risk of non-response within two years after index date, i.e., it is defined as $\mathrm{E}\{F_1(t \mid A = 1, X) - F_1(t \mid A = 0, X)\}$, where F_1 is the conditional cumulative incidence function for non-response and $A = 0$ denotes the reference drug and $A = 1$ is the comparitor. The study includes patients from 1995-2018 and not all of the antidepressants considered were

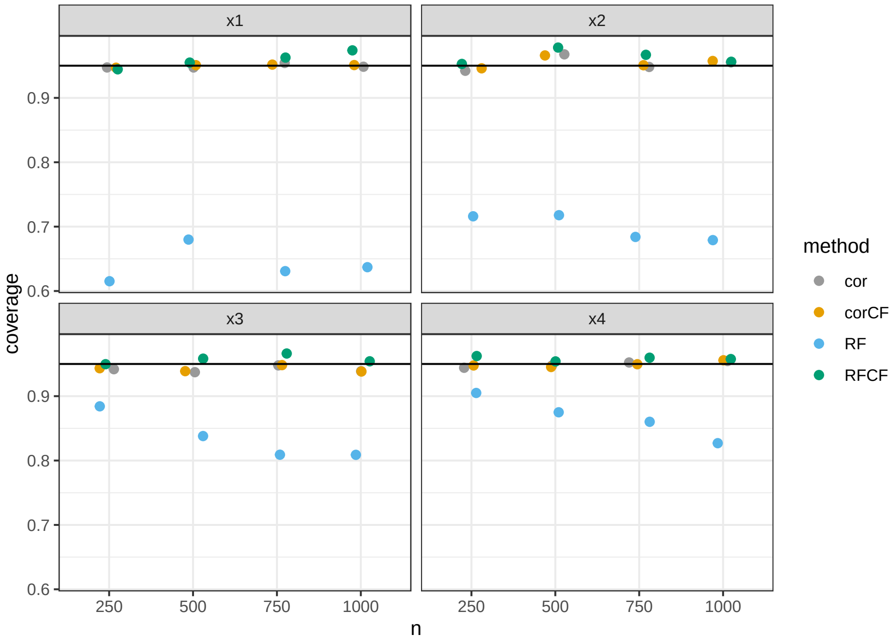


Figure 2: Results based on 1000 simulations of the estimators of Ω_1^l with for $l = 1, \ldots, 4$ with varying nuisance estimators and across sample sizes $n = 250, 500, 750, 1000$. The plot shows coverage of the estimators, where **cor** corresponds to the nuisance parameters $\Lambda_1, \Lambda_2, \Lambda_c, \pi$ estimated by correctly specified Cox and logistic regressions, and **RF** corresponds to the same parameters estimated by Random Forest. A suffix **CF** indicates that cross-fitting was employed. The black line indicates a coverage of 0.95.

available on the marked in the entire period. Accordingly, for each comparison, a minimum date is set for which both drugs in comparison were available and all patients with index dates prior to the minimum date are excluded.

For the sake of illustration we constrict ourselves to the comparison of Setraline (reference drug, $n = 14416$) and Escitalopram (comparitor, $n = 7508$). Kessing et al. (2024) used the G-formula with F_1 estimated by cause-specific Cox regressions (Ozenne et al., 2020) to estimate the average treatment effect. To control for confounding, the Cox-regressions were adjusted for the covariates in Table 2 and the ATE was estimated to be 0.10 (0.09, 0.12), that is, the probability of non-response was 0.1 higher amongst patients treated with Escitalopram within two years after treatment initiation. For comparison, instead of defining the treatment effect through the cumulative incidence function, F_1, we consider estimation of the ATE and the best partially linear projection based on the number of life-years-lost estimands defined in Section 3. That is, we consider estimation of ψ_j and Ω_j^l with $t^* = 730.5$ *days* (2 years). We include the same confounders as in Kessing et al. (2024) with the exception that age is included as a numeric variable instead of a categorised version. The target parameters are estimated with the cross-fitted estimators described in Section 4 with all nuisance parameters estimated by random forests (as described in the simulation study) and $K = 10$ folds.

The ATE is estimated to 48.96 (40.02, 57.90). The interpretation here is that patients on Setraline on average lost 49 ”healthy” days less before two years after treatment initiation due

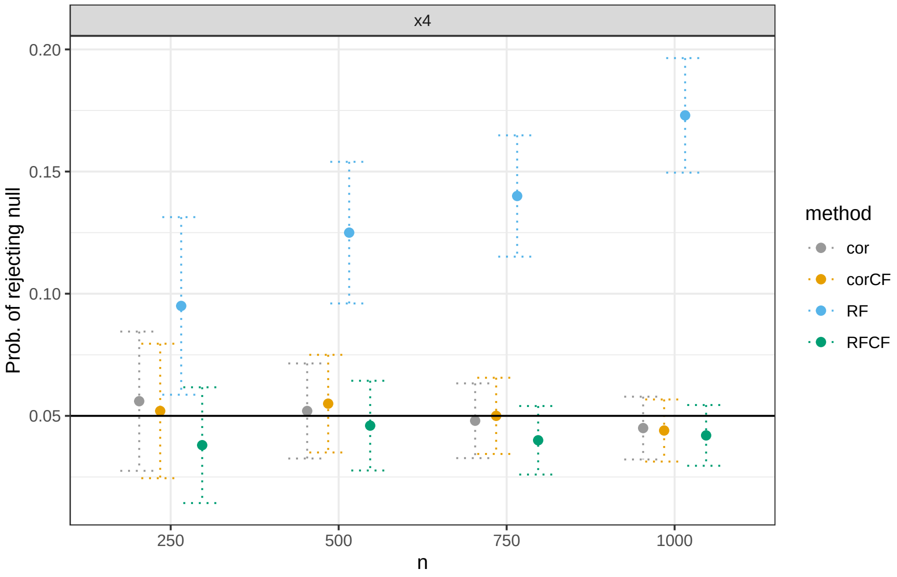


Figure 3: Results based on 1000 simulations of the test statistic corresponding to the test $H_0 : \Omega_1^4 = 0$ with varying nuisance estimators and across sample sizes $n = 250, 500, 750, 1000$. The plot shows probability of rejecting H_0, which equals the type-1 error as $\Omega_1^4 = 0$. **cor** corresponds to the nuisance parameters $\Lambda_1, \Lambda_2, \Lambda_c, \pi$ estimated by correctly specified Cox and logistic regressions, and **RF** corresponds to the same parameters estimated by Random Forest. A suffix **CF** indicates that cross-fitting was employed.

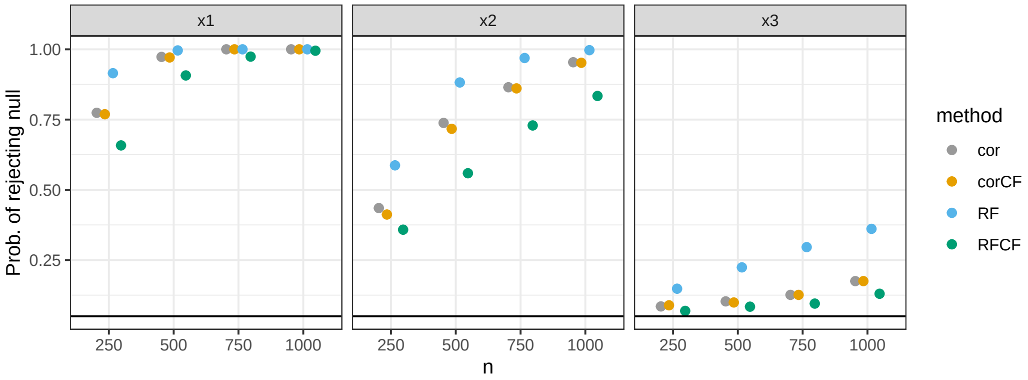


Figure 4: Results based on 1000 simulations of the test statistic corresponding to the test $H_0 : \Omega_1^l = 0$, for $l = 1, 2, 3$, with varying nuisance estimators and across sample sizes $n = 250, 500, 750, 1000$. The plot shows probability of rejecting H_0, which corresponds to the power of the test as $\Omega_1^l > 0$ for $l = 1, 2, 3$. **cor** corresponds to the nuisance parameters $\Lambda_1, \Lambda_2, \Lambda_c, \pi$ estimated by correctly specified Cox and logistic regressions, and **RF** corresponds to the same parameters estimated by Random Forest. A suffix **CF** indicates that cross-fitting was employed.

Annotation	**Explanation**
age	age-group: $(< 30, 30–50, 50–70, > 70)$ at index-date
sex	female, male
Secondary diagnosis from the psychiatric hospital at inclusion. The annotations reflect the ICD-10 codes used in the definition:	
F10-19	other psychiatric disorders
F40-48	neurotic, stress-related and somatoform disorders
F60-69	personalitiy disorders
Diagnosis with somatic disorder within 10 years prior to index date. The annotation are given in form of the corresponding ICD-10 chapter:	
I	infections
II	neoplasms
III	blood diseases
IV+IX+X	endocrine, nutritional, and metabolic diseases and diseases of the circulatory or respiratory system
VI-VIII	diseases of the nervous system, eye and ear
XI	diseases of the digestive system
XII	diseases of the skin and subcutaneous tissue
XIII	diseases if the musculoskeletal system
XIX	physical lesions and poisoning)

Table 2: Confounders in Kessing et al. (2024)

to non-response compared to patients who start on Escitalopram, where ”healthy” is meant as time without a non-response event or a competing event. Table 3 shows the estimates $\hat{\Omega}_1^{l,CF}$, for l given by each confounder in Table 2 (with age included as numeric), together with a p-value associated with the test statistic TST_1^l.

	age	sex	XIII	XI	F60-69	II	VI-VIII	IV+IX+X	F40-48	III	XIX	XII	I	F10-19
$\hat{\Omega}_1^{l,CF}$	-0.85	18.7	-28.3	-16.7	15.1	17.4	8.88	4.38	5.21	-4.80	-2.85	-2.00	-0.41	0.23
p-value	0.03	0.04	0.17	0.27	0.30	0.36	0.40	0.73	0.76	0.82	0.83	0.88	0.98	0.98

Table 3: Estimates of $\hat{\Omega}_1^{l,CF}$ ranked according to the p-value associated with the test of $H : \Omega_1^l = 0$, for l ranging over different covariates. Random Forest was used for all nuisance parameter estimators. The data comes from the study Kessing et al. (2024), and the outcome is time to non-response, which is defined as a switch in psychiatric treatment or rehospitalization at psychiatric ward. The treatment effect was defined as the difference in the number of healthy days lost (days without switch of treatment or re-hospitalization) due to non-response before two years after treatment initiation between Escitalopram and Setraline.

The p-values indicate that potential treatment effect heterogeneity can be attributed to sex and age, while the treatment effect does not seem to vary across any of the other variables. Specifically, since Ω_j is defined as the projection of the CATE function onto the partially linear model, the estimates related to sex and age can be interpreted as regression coefficients. The

CATE function is defined as the difference in number of healthy days lost due to non-response between Escitalopram and Setraline, and the estimate $\hat{\Omega}_1^{sex,CF} = 18.69$ then corresponds to the treatment effect being larger among women compared to men, i.e., the difference in number of healthy days lost due to non-response between Escitalopram and Setraline was larger among women. This interpretation is of course relying on the partially linear model to hold for the CATE function, but as the parameter still measures the association between τ and sex, when the partially linear model does not hold, we would still conclude, that the treatment effect is larger among women.

7 Discussion

In this paper, we have introduced the causal effect of a treatment on the number of life-years lost due to a specific event. We have shown that the ATE and the CATE are identifiable from the observed data under common assumptions found throughout the causal inference and survival literature. Different measures of treatment effect in the presence of competing risk are available in causal inference (Rytgaard and van der Laan, 2024, Rytgaard et al., 2023, Ozenne et al., 2020, Martinussen and Stensrud, 2023) and the treatment effect studied in this paper adds a new interpretability compared to existing variants. One advantage is that the treatment effect is defined on the time scale of the study and it thus provides a quantity that is easy to communicate to non-statisticians, whereas treatment effects based on the cumulative incidence function are harder to communicate. Furthermore, as the treatment effect can be written as a difference of integrated cumulative incidence functions, it is not as sensitive to the choice of time horizon in terms of detecting an effect of treatment. As is common when assessing the treatment effect in the presence of competing risk, the effect of a treatment on the number of life years lost due to a specific event depends on the effect of the treatment on both the hazard of the event of interest and on the competing event. As such, one can in principle conclude that there is an effect of treatment, even when all of the effect is driven by the effect on the competing event. Martinussen and Stensrud (2023) provide a measure of separable treatment effects based on the cumulative incidence function, which allow one to estimate the effect of treatment only driven by the intensity of the event of interest under additional causal assumptions. An interesting avenue for future research is to extend their method to the number of life years lost due to a specific event.

We have provided an estimator of the ATE based on semi-parametric efficiency theory, which allows for data-adaptive estimation of the nuisance parameters. The estimator is efficient in the non-parametric model with variance given by the efficient influence function. One of the assumptions needed to ensure the asymptotic results relies on convergence of a remainder term on $n^{-1/2}$-rate (assumption B3), which is reminiscent of similar assumption in the causal inference literature with censored data (Westling et al., 2023, Rytgaard et al., 2023). Without assuming absolute continuity of the hazard estimators, it is difficult to obtain an equivalent double rate-robustness property as is seen in the literature on uncensored data (e.g. Kennedy, 2022, Hines, Dukes, et al., 2022, Chernozhukov et al., 2018, van der Laan and Rose, 2011). Accordingly, we conducted a simulation study, where the nuisance parameters were estimated by variants of random forests, which confirmed that the estimator performed according to asymptotic results when using data-adaptive nuisance estimators.

Lastly, we extended a measure of treatment effect heterogeneity, termed the best partially linear projection (Ziersen and Martinussen, 2024), to the CATE-function defined on the number of life-years lost due to a specific event. The measure asserts the importance of a given covariate on the treatment effect, but with competing risk the interpretation is more delicate compared to the survival setting. When the effect of treatment on the competing event is large, one can imagine scenarios where the importance of one covariate on the CATE is driven by the effect on the competing event, and a ranking of importance (shown in Section 6) based on the event of interest might be misleading. In such scenarios, one can switch the event of interest and competing event and make a separate ranking of the covariates to get a full picture.

References

Andersen, P. K., Borgan, O., Gill, R. D., & Keiding, N. (1993). *Statistical models based on counting processes*. Springer Science & Business Media.

Andersen, P. K. (2013). Decomposition of number of life years lost according to causes of death. *Statistics in medicine*, *32*(30), 5278–5285.

Bickel, P. J., Klaassen, C. A., Bickel, P. J., Ritov, Y., Klaassen, J., Wellner, J. A., & Ritov, Y. (1993). *Efficient and adaptive estimation for semiparametric models* (Vol. 4). Springer.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters.

Cui, Y., Kosorok, M. R., Sverdrup, E., Wager, S., & Zhu, R. (2023). Estimating heterogeneous treatment effects with right-censored data via causal survival forests. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *85*(2), 179–211.

Cui, Y., Zhu, R., Zhou, M., & Kosorok, M. (2022). Consistency of survival tree and forest models: Splitting bias and correction. *Statistica Sinica*, *32*(3), 1245–1267.

Gill, R. D., & Johansen, S. (1990). A survey of product-integration with a view toward application in survival analysis. *The annals of statistics*, *18*(4), 1501–1555.

Gill, R. D., Van Der Laan, M. J., & Robins, J. M. (1997). Coarsening at random: Characterizations, conjectures, counter-examples. *Proceedings of the First Seattle Symposium in Biostatistics: Survival Analysis*, 255–294.

Hines, O., Diaz-Ordaz, K., & Vansteelandt, S. (2022). Variable importance measures for heterogeneous causal effects. *arXiv preprint arXiv:2204.06030*.

Hines, O., Dukes, O., Diaz-Ordaz, K., & Vansteelandt, S. (2022). Demystifying statistical learning based on efficient influence functions. *The American Statistician*, *76*(3), 292–304.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., & Lauer, M. S. (2008). Random survival forests.

Ishwaran, H., Kogalur, U. B., & Kogalur, M. U. B. (2023). Package ‘randomforestsrc’. *breast*, *6*(1), 854.

Kennedy, E. H. (2022). Semiparametric doubly robust targeted double machine learning: A review. *arXiv preprint arXiv:2203.06469*.

Kennedy, E. H. (2023). Towards optimal doubly robust estimation of heterogeneous causal effects. *Electronic Journal of Statistics*, *17*(2), 3008–3049.

Kennedy, E. H., Balakrishnan, S., & G’Sell, M. (2020). Sharp instruments for classifying compliers and generalizing causal effects.

Kessing, L. V., Ziersen, S. C., Andersen, F. M., Gerds, T., & Budtz-Jørgensen, E. (2024). Comparative responses to 17 different antidepressants in major depressive disorder: Results from a 2-year long-term nation-wide population-based study emulating a randomized trial. *Acta Psychiatrica Scandinavica*.

Martinussen, T., & Scheike, T. H. (2006). *Dynamic regression models for survival data* (Vol. 1). Springer.

Martinussen, T., & Stensrud, M. J. (2023). Estimation of separable direct and indirect effects in continuous time. *Biometrics*, *79*(1), 127–139.

Munch, A., Gerds, T. A., van der Laan, M. J., & Rytgaard, H. C. (2024). Estimating conditional hazard functions and densities with the highly-adaptive lasso. *arXiv preprint arXiv:2404.11083*.

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

Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical modelling*, *7*(9-12), 1393–1512.

Rytgaard, H. C., Eriksson, F., & van der Laan, M. J. (2023). Estimation of time-specific intervention effects on continuously distributed time-to-event outcomes by targeted maximum likelihood estimation. *Biometrics*.

Rytgaard, H. C., Gerds, T. A., & van der Laan, M. J. (2022). Continuous-time targeted minimum loss-based estimation of intervention-specific mean outcomes. *The Annals of Statistics*, *50*(5), 2469–2491.

Rytgaard, H. C., & van der Laan, M. J. (2024). Targeted maximum likelihood estimation for causal inference in survival and competing risks analysis. *Lifetime Data Analysis*, *30*(1), 4–33.

van der Laan, M. J., & Robins, J. M. (2003). *Unified methods for censored longitudinal data and causality*. Springer.

van der Laan, M. J., & Rose, S. (2011). *Targeted learning*. Springer.

van der Vaart, A. W. (2000). *Asymptotic statistics* (Vol. 3). Cambridge university press.

Westling, T., Luedtke, A., Gilbert, P. B., & Carone, M. (2023). Inference for treatment-specific survival curves using machine learning. *Journal of the American Statistical Association*, (just-accepted), 1–26.

Ziersen, S. C., & Martinussen, T. (2024). Variable importance measure for heterogeneous treatment effect with survival outcome. *unpublished manuscript*.

A Identification of causal estimands

We present an argument for the identification results in the main text. The argument follows usual steps in the causal inference literature on censored data, and it is based of a combination of the G-formula (Robins, 1986) and identification results from the literature on survival analysis (see e.g. Andersen et al., 1993 and Martinussen and Scheike, 2006).

As noted in Andersen (2013), T_j is improper due to $P(T_j = \infty) > 0$, but the random variable $T_j \wedge t^*$ is proper with an expectation given by

$$E\{T_j \wedge t^*\} = t^* - \int_0^{t^*} F_j(s)\,\mathrm{d}s.$$

Hence

$$E(T_j \wedge t^* \mid a, x) = t^* - \int_0^{t^*} F_j(s \mid a, x)\,\mathrm{d}s.$$

which is identified in the observed data under assumption A4 and A3 (Andersen et al., 1993). For the ATE, this allows us to write

$$\begin{aligned}
&E_0\{Y_j^1(t^*) - Y_j^0(t^*)\}\\
&= E_0\{\mathrm{E}(Y_j^1(t^*) - Y_j^0(t^*) \mid X)\}\\
&\overset{ass.A2}{=} E_0\{\mathrm{E}(Y_j^1(t^*) \mid A = 1, X) - \mathrm{E}(Y_j^1(t^*) \mid A = 0, X)\}\\
&\overset{ass.A1}{=} E_0\{\mathrm{E}(Y_j(t^*) \mid A = 1, X) - \mathrm{E}(Y_j(t^*) \mid A = 0, X)\}\\
&\overset{ass.A4}{=} E_0\{L_j(0, t^* \mid A = 1, X) - L_j(0, t^* \mid A = 0, X)\}
\end{aligned}$$

where assumption A3 ensures that all conditional distributions are well defined. We note that CATE is identified by the same arguments.

B Derivation of influence functions

In the following we consider the parametric submodel $P_\epsilon = \epsilon Q + (1-\epsilon)P$, where Q is the Dirac measure with pointmass in the observation $O = (\tilde{T}, \tilde{\Delta}, A, X)$, and we define the operator ∂_ϵ with $\partial_\epsilon f_\epsilon = \frac{d}{d\epsilon}|_{\epsilon=0} f_\epsilon$.

Lemma B.1. *The Gateaux derivative of $\Lambda_j(ds \mid a, x)$ is given by*

$$\partial_\epsilon \Lambda_{j,\epsilon}(ds \mid a, x) = \frac{1}{P(\tilde{T} \geq s, a, x)} \Big(Q(ds, \Delta = j, a, x) - \mathbb{1}(\tilde{T} \geq s, a, x)\Lambda_j(ds \mid a, x) \Big) \tag{9}$$

where

$$P(\tilde{T} \geq s, a, x) = \sum_{\delta=0}^{2} \int_s^\infty P(ds, \delta, a, x).$$

Proof. Observe

$$\begin{aligned}
&\partial_\epsilon \Lambda_{j,\epsilon}(ds \mid a,x) \\
=&\partial_\epsilon \frac{P_\epsilon(ds, \Delta = j, a, x)}{P_\epsilon(\tilde{T} \geq s, a, x)} \\
=&\frac{Q(ds, \Delta = j, a, x) - P(ds, \Delta = j, a, x)}{P(\tilde{T} \geq s, a, x)} - \partial_\epsilon P_\epsilon(\tilde{T} \geq s, a, x) \frac{P(ds, \Delta = j, a, x)}{P(\tilde{T} \geq s, a, x)^2} \\
=&\frac{Q(ds, \Delta = j, a, x) - P(ds, \Delta = j, a, x)}{P(\tilde{T} \geq s, a, x)} \\
&- \sum_{\delta=0}^{2} \Big(\mathbb{1}(\tilde{T} \geq s, \delta, a, x) - P(\tilde{T} \geq s, \delta, a, x) \Big) \frac{P(ds, \Delta = j, a, x)}{P(\tilde{T} \geq s, a, x)^2} \\
=&\frac{1}{P(\tilde{T} \geq s, a, x)} \Big(Q(ds, \Delta = j, a, x) - \mathbb{1}(\tilde{T} \geq s, a, x) \Lambda_j(ds \mid a, x) \Big)
\end{aligned}$$

□

Lemma B.2. *The Gateaux derivative of $S(s \mid a, x)$ is given by*

$$\partial_\epsilon S_\epsilon(s \mid a, x) = -S(s \mid a, x) \frac{\mathbb{1}(A = a, X = x)}{\pi(a \mid x), f(x)} \int_0^s \frac{\mathrm{d}M_1(u \mid a, x) + \mathrm{d}M_2(u \mid a, x)}{P(\tilde{T} \geq s, a, x)} \tag{10}$$

Proof. First we note that

$$\begin{aligned}
\partial_\epsilon \Lambda_{j,\epsilon}(s \mid a, x) &= \int_0^s \partial_\epsilon \Lambda_{j,\epsilon}(ds \mid a, x) \\
&= \frac{\mathbb{1}(A = a, X = x)}{\pi(a \mid x) f(x)} \int_0^s \frac{\mathrm{d}M_j(u \mid a, x)}{P(\tilde{T} \geq s \mid a, x)}
\end{aligned} \tag{11}$$

by lemma B.1. Then

$$\begin{aligned}
\partial_\epsilon S(s \mid a, x) &= \partial_\epsilon \exp(-[\Lambda_{1,\epsilon}(s \mid a, x) + \Lambda_{2,\epsilon}(s \mid a, x)]) \\
&= -S(s \mid a, x) \partial_\epsilon [\Lambda_{1,\epsilon}(s \mid a, x) + \Lambda_{2,\epsilon}(s \mid a, x)].
\end{aligned}$$

Applying (11) gives the result. □

Lemma B.3. *The Gateaux derivative of $L_j(0, t^* \mid a, x)$ is given by*

$$\partial_\epsilon L_{j,\epsilon}(0, t^* \mid a, x) = \frac{\mathbb{1}(A = a, X = x)}{\pi(a \mid x) f(x)} \sum_{i=1,2} \int_0^{t^*} \frac{H_{ij}(s, t^* \mid a, x)}{S_C(s \mid a, x)} \mathrm{d}M_i(s \mid a, x) \tag{12}$$

where

$$H_{ij}(s, t \mid a, x) = \int_s^t \mathbb{1}(i = j) + \frac{F_j(s \mid a, x) - F_j(u \mid a, x)}{S(s \mid a, x)} \mathrm{d}u.$$

Proof. First we note that

$$\partial_\epsilon F_{j,\epsilon}(t \mid a, x) = \int_0^t \partial_\epsilon S_\epsilon(s \mid a, x) \Lambda_j(ds \mid a, x) + \int_0^t \partial_\epsilon S(s \mid a, x) \Lambda_{j,\epsilon}(ds \mid a, x). \tag{13}$$

For the first term in (13), observe

$$\begin{aligned}
&\int_0^t \partial_\epsilon S_\epsilon(s \mid a,x)\Lambda_j(ds \mid a,x) \qquad (14)\\
=&\frac{\mathbb{1}(A=a,X=x)}{\pi(a \mid x)f(x)}\left[\int_0^t\int_0^s \frac{-S(s \mid a,x)}{P(\tilde{T}\geq u \mid a,x)}\,\mathrm{d}M_1(u \mid a,x)\Lambda_j(s \mid a,x)\right.\\
&\qquad\left.+\int_0^t\int_0^s \frac{-S(s \mid a,x)}{P(\tilde{T}\geq u \mid a,x)}\,\mathrm{d}M_2(u \mid a,x)\Lambda_j(s \mid a,x)\right]
\end{aligned}$$

by lemma B.2, and for the second term

$$\begin{aligned}
&\int_0^t \partial_\epsilon S(s \mid a,x)\Lambda_{j,\epsilon}(ds \mid a,x) \qquad (15)\\
=&\int_0^t \frac{S(s \mid a,x)}{P(\tilde{T}\geq s,a,x)}\Big(Q(ds,\Delta=j,a,x)-\mathbb{1}(\tilde{T}\geq s,a,x)\Lambda_j(ds \mid a,x)\Big)\\
=&\frac{\mathbb{1}(A=a,X=x)}{\pi(a \mid x)f(x)}\int_0^t \frac{\mathrm{d}N_j(s)}{S_C(s \mid a,x)}-\frac{\mathbb{1}(\tilde{T}\geq s)\Lambda_j(ds \mid a,x)}{S_C(s \mid a,x)}\\
=&\frac{\mathbb{1}(A=a,X=x)}{\pi(a \mid x)f(x)}\int_0^t \frac{\mathrm{d}M_j(s \mid a,x)}{S_C(s \mid a,x)}
\end{aligned}$$

by lemma B.1. Plugging into (13) gives

$$\begin{aligned}
&\partial_\epsilon F_{j,\epsilon}(t \mid a,x)\\
=&\frac{\mathbb{1}(A=a,X=x)}{\pi(a \mid x)f(x)}\left[\int_0^t \frac{\int_u^t -S(s \mid a,x)\Lambda_j(ds \mid a,x)}{P(\tilde{T}\geq u \mid a,x)}\,\mathrm{d}M_1(u \mid a,x)\right.\\
&\qquad\left.+\int_0^t \frac{\int_u^t -S(s \mid a,x)\Lambda_j(ds \mid a,x)}{P(\tilde{T}\geq u \mid a,x)}\,\mathrm{d}M_2(u \mid a,x)+\int_0^t \frac{\mathrm{d}M_j(s \mid a,x)}{S_C(s \mid a,x)}\right]\\
=&\frac{\mathbb{1}(A=a,X=x)}{\pi(a \mid x)f(x)}\left[\int_0^t \frac{F_j(u \mid a,x)-F_j(t \mid a,x)}{S(u \mid a,x)S_C(u \mid a,x)}\,\mathrm{d}M_1(u \mid a,x)\right.\\
&\qquad\left.+\int_0^t \frac{F_j(u \mid a,x)-F_j(t \mid a,x)}{S(u \mid a,x)S_C(u \mid a,x)}\,\mathrm{d}M_2(u \mid a,x)+\int_0^t \frac{\mathrm{d}M_j(u \mid a,x)}{S_C(u \mid a,x)}\right]\\
=&\frac{\mathbb{1}(A=a,X=x)}{\pi(a \mid x)f(x)}\sum_{i=1,2}\int_0^t \frac{1}{S_C(u \mid a,x)}\left(\frac{F_j(u \mid a,x)-F_j(t \mid a,x)}{S(u \mid a,x)}+\mathbb{1}(i=j)\right)\mathrm{d}M_i(u \mid a,x).
\end{aligned} \qquad (16)$$

Finally, by (16), we have

$$
\begin{aligned}
&\partial_\epsilon L_{j,\epsilon}(0, t^* \mid a, x) \qquad (17)\\
&= \int_0^{t^*} \partial_\epsilon F_{j,\epsilon}(t \mid a, x)\,\mathrm{d}t \\
&= \int_0^{t^*} \sum_{i=1,2} \int_0^t \frac{1}{S_C(u \mid a, x)} \left(\frac{F_j(u \mid a, x) - F_j(t \mid a, x)}{S(u \mid a, x)} + \mathbb{1}(i = j) \right) \mathrm{d}M_i(u \mid a, x)\,\mathrm{d}t \\
&\quad \times \frac{\mathbb{1}(A = a, X = x)}{\pi(a \mid x) f(x)} \\
&= \int_0^{t^*} \sum_{i=1,2} \int_u^{t^*} \left(\frac{F_j(u \mid a, x) - F_j(t \mid a, x)}{S(u \mid a, x)} + \mathbb{1}(i = j) \right) \mathrm{d}t \frac{1}{S_C(u \mid a, x)}\,\mathrm{d}M_i(u \mid a, x) \\
&\quad \times \frac{\mathbb{1}(A = a, X = x)}{\pi(a \mid x) f(x)} \\
&= \frac{\mathbb{1}(A = a, X = x)}{\pi(a \mid x) f(x)} \sum_{i=1,2} \int_0^{t^*} \frac{H_{ij}(s, t^* \mid a, x)}{S_C(s \mid a, x)}\,\mathrm{d}M_i(s \mid a, x).
\end{aligned}
$$

□

Proof of lemma 1. We have

$$
\partial_\epsilon \psi_j(P_\epsilon) = \partial_\epsilon \operatorname{E}_{P_\epsilon}\{\tau_{j,P_\epsilon}(X)\} = \tau_j(X) + \operatorname{E}\{\partial_\epsilon \tau_{j,P_\epsilon}(X)\} - \psi_j(P).
$$

Applying lemma B.3 gives the result. □

Proof of lemma 2. The EIF of $\chi^l(P)$ is given by theorem 3 in Ziersen and Martinussen (2024). By remark 2 in Ziersen and Martinussen (2024), the EIF of $\Gamma^l_j(P)$ is given by $\tilde{\psi}_{\Gamma^l_j}$ provided that the CATE function $\tau_j(x)$ has Gateaux derivative given by $\frac{\mathbb{1}(X=x)}{f(x)}\left[\varphi_j(O) - \tau_j(X)\right]$, which is seen to hold by lemma B.3. The EIF of Ω^l_j then follows from the chain rule. □

C Proof of asymptotic results

Lemma C.1. *The remainder term $P\{\varphi^a_j(\hat{\nu}) - \tau^a_j\}$, $a = 0, 1$, can be represented as*

$$
\begin{aligned}
\operatorname{E}\Bigg\{ &\sum_{i=1,2} \int_0^{t^*} S(s \mid a, X) \hat{H}_{ij}(s, t^* \mid a, X) \\
&\times \left(1 - \frac{\pi(a \mid X) S_C(s \mid a, X)}{\hat{\pi}(a \mid X) \hat{S}_C(s \mid a, X)} \right) \mathrm{d}\left[\hat{\Lambda}_i(s \mid a, X) - \Lambda_i(s \mid a, X) \right] \Bigg\}
\end{aligned}
$$

where the expectation is taken with respect to an observation O, considering the nuisance estimators, $\hat{\nu}$, fixed. Under assumption B3 it holds that $P\{\varphi_j(\hat{\nu}) - \tau_j\} = o_p(n^{-1/2})$.

Proof of lemma C.1. Throughout the proof, expectations and conditional expectations will be taken with respect to an observation O, considering estimated nuisance parameters fixed. For the first statement we have

$$\begin{aligned} P(\phi_j^a(\hat{\nu}) - \tau_j^a) = &\, \mathrm{E}\left\{\hat{L}_j(0, t^* \mid a, X) - L_j(0, t^* \mid a, X)\right\} \\ &+ \mathrm{E}\left\{\frac{\mathbb{1}(A = a)}{\hat{\pi}(a \mid X)} \sum_{i=1,2} \int_0^{t^*} \frac{\hat{H}_{ij}(s, t^* \mid a, X)}{\hat{S}_C(s \mid a, X)} \,\mathrm{d}\hat{M}_i(s \mid a, X)\right\} \end{aligned} \tag{18}$$

and we will expand the two terms separately. For the first term note that

$$\begin{aligned} \hat{F}_j(t \mid a, x) - F_j(t \mid a, x) = &\int_0^t \hat{S}(s \mid a, x) - S(s \mid a, x) \,\mathrm{d}\hat{\Lambda}_j(s \mid a, x) \\ &+ \int_0^t S(s \mid a, x) \left[\hat{\Lambda}_j(s \mid a, x) - \Lambda_j(s \mid a, x)\right]. \end{aligned} \tag{19}$$

Define

$$\Lambda(s \mid a, x) = \Lambda_1(s \mid a, x) + \Lambda_1(s \mid a, x),$$

then Duhamel's equations (Gill and Johansen, 1990) gives

$$\hat{S}(s \mid a, x) - S(s \mid a, x) = \int_0^s \frac{S(u \mid a, x)}{\hat{S}(u \mid a, x)} \,\mathrm{d}\left[\Lambda(u \mid a, x) - \hat{\Lambda}(u \mid a, x)\right] \hat{S}(s \mid a, x).$$

Plugging this into the first term in (19) gives

$$\begin{aligned} &\int_0^t \hat{S}(s \mid a, x) - S(s \mid a, x) \,\mathrm{d}\hat{\Lambda}_j(s \mid a, x) \\ &= \int_0^t \int_0^s \frac{S(u \mid a, x)}{\hat{S}(u \mid a, x)} \,\mathrm{d}\left[\Lambda(u \mid a, x) - \hat{\Lambda}(u \mid a, x)\right] \hat{S}(s \mid a, x) \,\mathrm{d}\hat{\Lambda}_j(s \mid a, x) \\ &= \int_0^t \int_u^t \hat{S}(s \mid a, x) \frac{S(u \mid a, x)}{\hat{S}(u \mid a, x)} \,\mathrm{d}\hat{\Lambda}_j(s \mid a, x) \,\mathrm{d}\left[\Lambda(u \mid a, x) - \hat{\Lambda}(u \mid a, x)\right] \\ &= \int_0^t \frac{S(u \mid a, x)}{\hat{S}(u \mid a, x)} \left(\hat{F}_j(t \mid a, x) - \hat{F}_j(u \mid a, x)\right) \mathrm{d}\left[\Lambda(u \mid a, x) - \hat{\Lambda}(u \mid a, x)\right] \\ &= \int_0^t \frac{S(u \mid a, x)}{\hat{S}(u \mid a, x)} \left(\hat{F}_j(t \mid a, x) - \hat{F}_j(u \mid a, x)\right) \mathrm{d}\left[\Lambda_1(u \mid a, x) - \hat{\Lambda}_1(u \mid a, x)\right] \\ &\quad + \int_0^t \frac{S(u \mid a, x)}{\hat{S}(u \mid a, x)} \left(\hat{F}_j(t \mid a, x) - \hat{F}_j(u \mid a, x)\right) \mathrm{d}\left[\Lambda_2(u \mid a, x) - \hat{\Lambda}_2(u \mid a, x)\right] \end{aligned}$$

Using this expansion, we can write (19) as

$$
\begin{aligned}
&\hat{F}_j(t \mid a,x) - F_j(t \mid a,x) \\
&= \int_0^t \frac{S(u \mid a,x)}{\hat{S}(u \mid a,x)} \left(\hat{F}_j(u \mid a,x) - \hat{F}_j(t \mid a,x)\right) \mathrm{d}\left[\hat{\Lambda}_1(u \mid a,x) - \Lambda_1(u \mid a,x)\right] \\
&\quad + \int_0^t \frac{S(u \mid a,x)}{\hat{S}(u \mid a,x)} \left(\hat{F}_j(u \mid a,x) - \hat{F}_j(t \mid a,x)\right) \mathrm{d}\left[\hat{\Lambda}_2(u \mid a,x) - \Lambda_2(u \mid a,x)\right] \\
&\quad + \int_0^t S(s \mid a,x) \left[\hat{\Lambda}_j(s \mid a,x) - \Lambda_j(s \mid a,x)\right] \\
&= \sum_{i=1,2} \int_0^t S(s \mid a,x) \left(\mathbb{1}(i=j) + \frac{\hat{F}_j(s \mid a,x) - \hat{F}_j(t \mid a,x)}{\hat{S}(s \mid a,x)}\right) \mathrm{d}\left[\hat{\Lambda}_i(u \mid a,x) - \Lambda_i(u \mid a,x)\right]
\end{aligned}
\tag{20}
$$

For the second term in (18) note that

$$
\begin{aligned}
&\mathrm{E}\left(\int_0^{t^*} \frac{\hat{H}_{ij}(s,t^* \mid A,X)}{\hat{S}_C(s \mid A,X)} \,\mathrm{d}\hat{M}_i(s \mid A,X) \,\middle|\, A,X \right) \\
&= \int_0^{t^*} \frac{\hat{H}_{ij}(s,t^* \mid A,X)}{\hat{S}_C(s \mid A,X)} \,\mathrm{E}(\mathrm{d}N_i(s) \mid A,X) \\
&\quad - \int_0^{t^*} \frac{\hat{H}_{ij}(s,t^* \mid A,X)}{\hat{S}_C(s \mid A,X)} \,\mathrm{E}(\mathbb{1}(\tilde{T} \geq s) \mid A,X)\,\mathrm{d}\hat{\Lambda}_i(s \mid A,X) \\
&= \int_0^{t^*} \frac{\hat{H}_{ij}(s,t^* \mid A,X)}{\hat{S}_C(s \mid A,X)} S(s \mid A,X) S_C(s \mid A,X)\,\mathrm{d}\left[\Lambda_i(s \mid A,X) - \hat{\Lambda}_i(s \mid A,X)\right].
\end{aligned}
\tag{21}
$$

Hence, by collecting (20) and (21), and using iterated expectation, we can write (18) as

$$
\begin{aligned}
&\mathrm{E}\Bigg\{ \sum_{i=1,2} \int_0^t S(s \mid a,X) \left(\mathbb{1}(i=j) + \frac{\hat{F}_j(s \mid a,X) - \hat{F}_j(t \mid a,X)}{\hat{S}(s \mid a,X)}\right) \mathrm{d}\left[\hat{\Lambda}_i(u \mid a,X) - \Lambda_i(u \mid a,X)\right] \\
&- \frac{\pi(a \mid X)}{\hat{\pi}(a \mid X)} \sum_{i=1,2} \int_0^{t^*} \frac{\hat{H}_{ij}(s,t^* \mid a,X)}{\hat{S}_C(s \mid a,X)} S(s \mid a,X) S_C(s \mid a,X)\,\mathrm{d}\left[\hat{\Lambda}_i(s \mid a,X) - \Lambda_i(s \mid a,X)\right] \Bigg\} \\
=&\,\mathrm{E}\Bigg\{ \sum_{i=1,2} \int_0^{t^*} S(s \mid a,X) \hat{H}_{ij}(s,t^* \mid a,X) \\
&\quad \times \left(1 - \frac{\pi(a \mid X) S_C(s \mid a,X)}{\hat{\pi}(a \mid X) \hat{S}_C(s \mid a,X)}\right) \mathrm{d}\left[\hat{\Lambda}_i(s \mid a,X) - \Lambda_i(s \mid a,X)\right] \Bigg\}
\end{aligned}
\tag{22}
$$

which gives the first statement. For the second statement note that

$$
P\{\phi^j(\hat{\nu}) - \tau^j\} = P\{\phi^j_1(\hat{\nu}) - \tau^j_1\} - P\{\phi^j_0(\hat{\nu}) - \tau^j_0\}.
$$

Applying the representation above along with assumption B3 gives the result. □

Lemma C.2. *Let $g(s \mid a,x) = \pi(a \mid x)S_C(s \mid a,x)$ and $\hat{g}(s \mid a,x) = \hat{\pi}(a \mid x)\hat{S}_C(s \mid a,x)$. Under assumption B1 and B2, the uncentered EIF, φ_j, is bounded in $L_2(P)$-norm as:*

$$\begin{aligned}\|\varphi_j(\hat{\nu}) - \varphi_j(\nu)\| \leq & C_1^* \left\| \sup_{s \leq t^*} \left|\hat{\Lambda}_1(s \mid a,X) - \Lambda_1(s \mid a,X)\right|\right\| \\ & + C_2^* \left\| \sup_{s \leq t^*} \left|\hat{\Lambda}_2(s \mid a,X) - \Lambda_2(s \mid a,X)\right|\right\| \\ & + C_c^* \left\| \sup_{s \leq t^*} \left|\hat{g}(s \mid a,X) - g(s \mid a,X)\right|\right\|\end{aligned}$$

Proof of lemma C.2. Note that

$$\|\varphi_j(\hat{\nu}) - \varphi_j(\nu)\| \leq \left\|\varphi_j^1(\hat{\nu}) - \varphi_j^1(\nu)\right\| + \left\|\varphi_j^0(\hat{\nu}) - \varphi_j^0(\nu)\right\|.$$

Hence we need to show $\left\|\varphi_j^a(\hat{\nu}) - \varphi_j^a(\nu)\right\| = o_p(1)$ for $a = 0,1$. Observe that

$$\begin{aligned}& P\{\varphi_j^a(\hat{\nu}) - \varphi_j^a(\nu)\}^2 \\ \leq & 2\,\mathrm{E}\left\{\hat{L}_j(0,t^* \mid a,X) - L_j(0,t^* \mid a,X)\right\}^2 & (23) \\ & + 2\,\mathrm{E}\left\{\sum_{i=1,2}\int_0^{t^*} \frac{\hat{H}_{ij}(s,t^* \mid a,X)}{\hat{g}(s \mid a,X)} - \frac{H_{ij}(s,t^* \mid a,X)}{g(s \mid a,X)}\,\mathrm{d}N_i(s)\right\}^2 & (24) \\ & + 2\,\mathrm{E}\left\{\sum_{i=1,2}\left[\int_0^{t^* \wedge \tilde{T}} \frac{\hat{H}_{ij}(s,t^* \mid a,X)}{\hat{g}(s \mid a,X)}\,\mathrm{d}\hat{\Lambda}_i(s \mid a,X) - \int_0^{t^* \wedge \tilde{T}} \frac{H_{ij}(s,t^* \mid a,X)}{g(s \mid a,X)}\,\mathrm{d}\Lambda_i(s \mid a,X)\right]\right\}^2. \\ & & (25)\end{aligned}$$

We will deal with each of the terms separately, but first we will derive some results for the consistency of $\hat{F}_j$, $\hat{L}_j$ and $\hat{H}_{ij}$. For $\hat{F}_j$ we have that

$$\begin{aligned}& \hat{F}_j(t^* \mid a,x) - F_j(t^* \mid a,x) \\ = & \int_0^{t^*} \hat{S}(s \mid a,x)\,\mathrm{d}\left[\hat{\Lambda}_j(s \mid a,x) - \Lambda_j(s \mid a,x)\right] + \int_0^{t^*} \hat{S}(s \mid a,x) - S(s \mid a,x)\,\mathrm{d}\Lambda_j(s \mid a,x) \\ = & \hat{S}(t^* \mid a,x)\left[\hat{\Lambda}_j(t^* \mid a,x) - \Lambda_j(t^* \mid a,x)\right] - \hat{S}(0 \mid a,x)\left[\hat{\Lambda}_j(0 \mid a,x) - \Lambda_j(0 \mid a,x)\right] \\ & - \int_0^{t^*} \hat{\Lambda}_j(s \mid a,x) - \Lambda_j(s \mid a,x)\,\mathrm{d}\hat{S}(s \mid a,x) + \int_0^{t^*} \hat{S}(s \mid a,x) - S(s \mid a,x)\,\mathrm{d}\Lambda_j(s \mid a,x) \\ \leq & 2\sup_{s \leq t^*}\left|\hat{\Lambda}_j(s \mid a,x) - \Lambda_j(s \mid a,x)\right| + \hat{\Lambda}_j(t \mid a,x)\sup_{s \leq t^*}\left|\hat{S}(s \mid a,x) - S(s \mid a,x)\right| \\ \leq & 2\sup_{s \leq t^*}\left|\hat{\Lambda}_j(s \mid a,x) - \Lambda_j(s \mid a,x)\right| \\ & + \log(\eta^{-1})e^C \sup_{s \leq t^*}\left[\left|\hat{\Lambda}_1(s \mid a,x) - \Lambda_1(s \mid a,x)\right| + \left|\hat{\Lambda}_2(s \mid a,x) - \Lambda_2(s \mid a,x)\right|\right] \\ = & C_1 \sup_{s \leq t^*}\left|\hat{\Lambda}_1(s \mid a,x) - \Lambda_1(s \mid a,x)\right| + C_2 \sup_{s \leq t^*}\left|\hat{\Lambda}_2(s \mid a,x) - \Lambda_2(s \mid a,x)\right| & (26)\end{aligned}$$

for constants $C_1 > 0$ and $C_2 > 0$. The second equality follows from partial integration and the second inequality follows from the mean value theorem with $C > 0$ is some value between $\hat{\Lambda}(s \mid a, x)$ and $\Lambda(s \mid a, x)$ together with assumption B1.

From (26) it follows immediately that

$$\begin{aligned}
&\hat{L}_j(0, t^* \mid a, x) - L_j(0, t^* \mid a, x) \\
\leq & t^* C_1 \sup_{s \leq t^*} \left| \hat{\Lambda}_1(s \mid a, x) - \Lambda_1(s \mid a, x) \right| + t^* C_2 \sup_{s \leq t^*} \left| \hat{\Lambda}_2(s \mid a, x) - \Lambda_2(s \mid a, x) \right| .
\end{aligned} \tag{27}$$

For $\hat{H}_{ij}$, observe

$$\begin{aligned}
&\hat{H}_{ij}(s, t^* \mid a, x) - H_{ij}(s, t^* \mid a, x) - H_{ij}(s, t^* \mid a, x) - H_{ij}(s, t^* \mid a, x) \\
= & \int_s^{t^*} \frac{\hat{F}_j(s \mid a, x) - F_j(s \mid, a, x) - (\hat{F}_j(u \mid a, x) - F_j(u \mid, a, x))}{\hat{S}(s \mid a, x)} \, \mathrm{d}u \\
& + \int_s^{t^*} \left(\frac{1}{\hat{S}(s \mid a, x)} - \frac{1}{S(s \mid a, x)} \right) (F_j(s \mid a, x) - F(u)) \, \mathrm{d}u \\
\leq & t^* \left| \frac{\hat{F}_j(s \mid a, x) - F_j(s \mid a, x)}{\hat{S}(s \mid a, x)} \right| + \eta^{-1} \int_s^{t^*} \left| \hat{F}_j(u \mid a, x) - F_j(u \mid a, x) \right| \mathrm{d}u \\
& + \eta^{-2} \left| \hat{S}(s \mid a, x) - (s \mid a, x) \right| \left(F_j(s \mid a, x) - \int_s^{t^*} F_j(u \mid a, x) \, \mathrm{d}u \right) \\
\leq & C_3 \sup_{s \leq t^*} \left| \hat{\Lambda}_1(s \mid a, x) - \Lambda_1(s \mid a, x) \right| + C_4 \sup_{s \leq t^*} \left| \hat{\Lambda}_2(s \mid a, x) - \Lambda_2(s \mid a, x) \right|
\end{aligned} \tag{28}$$

for some constants $C_3 > 0$ and $C_4 > 0$, which follows from (26). Now we proceed to bound each of the three terms initially stated. The bound of (23) follows directly from (27):

$$\begin{aligned}
& 2\,\mathrm{E} \left\{ \hat{L}_j(0, t^* \mid a, X) - L_j(0, t^* \mid a, X) \right\}^2 \\
\leq & 2\,\mathrm{E} \left\{ C_5 \sup_{s \leq t^*} \left| \hat{\Lambda}_1(s \mid a, x) - \Lambda_1(s \mid a, x) \right| + C_6 \sup_{s \leq t^*} \left| \hat{\Lambda}_2(s \mid a, x) - \Lambda_2(s \mid a, x) \right| \right\}^2 .
\end{aligned} \tag{29}$$

For (24) we have

$$\begin{aligned}
&\mathrm{E}\left\{\sum_{i=1,2}\int_0^{t^*}\frac{\hat{H}_{ij}(s,t^*\mid a,X)}{\hat{g}(s\mid a,X)}-\frac{H_{ij}(s,t^*\mid a,X)}{g(s\mid a,X)}\,\mathrm{d}N_i(s)\right\}^2\\
=&\,\mathrm{E}\left\{\sum_{i=1,2}\int_0^{t^*}\frac{\hat{H}_{ij}(s,t^*\mid a,X)-H_{ij}(s,t^*\mid a,X)}{\hat{g}(s\mid a,X)}\,\mathrm{d}N_i(s)\right.\\
&\left.+\int_0^{t^*}H_{ij}(s,t^*\mid a,x)\left(\frac{1}{\hat{g}(s\mid a,X)}-\frac{1}{g(s\mid a,X)}\right)\mathrm{d}N_i(s)\right\}^2\\
\leq&\,\mathrm{E}\left\{\eta^{-2}\sup_{s\leq t^*}\left|\hat{H}_{ij}(s\mid a,X)-H_{ij}(s\mid a,X)\right|\right. &&(30)\\
&\left.+\eta^{-4}\sup_{s\leq t^*}\left|\hat{g}(s\mid a,X)-g(s\mid a,X)\right|\left|H_{ij}(s,t^*\mid a,X)\right|\right\}^2\\
\leq&\,\mathrm{E}\left\{C_7\sup_{s\leq t^*}\left|\hat{\Lambda}_1(s\mid a,X)-\Lambda_1(s\mid a,X)\right|+C_8\sup_{s\leq t^*}\left|\hat{\Lambda}_2(s\mid a,X)-\Lambda_2(s\mid a,X)\right|\right.\\
&\left.\qquad+C_9\sup_{s\leq t^*}\left|\hat{g}(s\mid a,X)-g(s\mid a,X)\right|\right\}^2. &&(31)
\end{aligned}$$

The first inequality follows from assumption B1 and the second inequality follows from (28) together with

$$\begin{aligned}
H_{ij}(s,t^*\mid a,x)\leq&\,\eta^{-1}\int_s^{t^*}\left|F_j(s)-F_j(u)\right|\mathrm{d}u+t^*-s\\
\leq&\,\eta^{-1}2t^* &&(32)
\end{aligned}$$

by assumption B1.

Lastly, for (25) we have

$$\begin{aligned}
&\mathrm{E}\left\{\sum_{i=1,2}\left[\int_0^{t^*\wedge\tilde{T}}\frac{\hat{H}_{ij}(s,t^*\mid a,X)}{\hat{g}(s\mid a,X)}\,\mathrm{d}\hat{\Lambda}_i(s\mid a,X)-\int_0^{t^*\wedge\tilde{T}}\frac{H_{ij}(s,t^*\mid a,X)}{g(s\mid a,X)}\,\mathrm{d}\Lambda_i(s\mid a,X)\right]\right\}^2\\
=&E\left\{\underbrace{\sum_{i=1,2}\left[\int_0^{t^*\wedge\tilde{T}}\hat{H}_{ij}(s,t^*\mid a,X)\left(\frac{1}{\hat{g}(s\mid a,x)}-\frac{1}{g(s\mid a,x)}\right)\mathrm{d}\hat{\Lambda}_i(s\mid a,x)\right]}_{(a)}\right.\\
&\left.+\underbrace{\int_0^{t^*\wedge\tilde{T}}\sum_{i=1,2}\frac{\hat{H}_{ij}(s,t^*\mid a,X)}{g(s\mid a,X)}\,\mathrm{d}\hat{\Lambda}_i(s\mid a,X)-\int_0^{t^*\wedge\tilde{T}}\sum_{i=1,2}\frac{H_{ij}(s\mid a,X)}{g(s\mid a,X)}\,\mathrm{d}\Lambda_i(s\mid a,X)}_{(b)}\right\}^2.\\
& &&(33)
\end{aligned}$$

and we will bound (a) and (b) separately. For (a) we have that for almost all x

$$\begin{aligned}
&\int_0^{t^*\wedge\tilde{T}} \hat{H}_{ij}(s,t^* \mid a,x)\left(\frac{1}{\hat{g}(s \mid a,x)} - \frac{1}{g(s \mid a,x)}\right)\mathrm{d}\hat{\Lambda}_i(s \mid a,x)\\
\leq &\sum_{i=1,2}\sup_{s\leq t^*}\left|\hat{H}_{ij}(s,t^* \mid a,x)\right|\sup_{s\leq t^*}\left|\frac{1}{\hat{g}(s \mid a,x)} - \frac{1}{g(s \mid a,x)}\right|\hat{\Lambda}_i(t^* \mid a,x)\\
\leq &4\eta^{-5}t^*\log(\eta)\sup_{s\leq t^*}\left|\hat{g}(s \mid a,x) - g(s \mid a,x)\right|. \qquad (34)
\end{aligned}$$

by (32) together with assumption B1. Next, to control (b), define

$$H_j(s,t^* \mid a,x) = \int_s^{t^*}\frac{F_j(s \mid a,x) - F_j(u \mid a,x)}{S(s)}\,\mathrm{d}u$$

and let

$$\Lambda(s \mid a,x) = \Lambda_1(s \mid a,x) + \Lambda_2(s \mid a,x)$$

such that $S(s \mid a,x) = e^{-\Lambda(s|a,x)}$. Observe that

$$\begin{aligned}
&H_j(s,t^* \mid a,x)\\
=&-\int_s^{t^*}\frac{F_j(u \mid a,x) - F_j(s \mid a,x)}{S(s)}\,\mathrm{d}u\\
=&-\int_s^{t^*}\int_s^u\frac{S(v \mid a,x)}{S(s \mid a,x)}\Lambda_j(dv \mid a,x)\,\mathrm{d}u\\
\overset{(*)}{=}&-\int_s^{t^*}\int_s^u\left(1-\int_s^v\frac{S(v \mid a,x)}{S(w \mid a,x)}\Lambda(dw \mid a,x)\right)\Lambda_j(dv \mid a,x)\,\mathrm{d}u\\
=&-\left[\int_s^{t^*}\int_s^u\Lambda_j(\mathrm{d}v \mid a,x)\,\mathrm{d}u - \int_s^{t^*}\int_s^u\int_s^v\frac{S(v \mid a,x)}{S(w \mid a,x)}\Lambda(\mathrm{d}w \mid a,x)\Lambda_j(\mathrm{d}v \mid a,x)\,\mathrm{d}u\right]\\
=&-\left[\int_s^{t^*}\int_v^{t^*}\mathrm{d}u\Lambda_j(\mathrm{d}v \mid a,x) - \int_s^{t^*}\int_w^{t^*}\int_w^u\frac{S(v \mid a,x)}{S(w \mid a,x)}\Lambda_j(dv \mid a,x)\,\mathrm{d}u\Lambda(dw \mid a,x)\right]\\
=&-\left[\int_s^{t^*}\int_v^{t^*}\mathrm{d}u\Lambda_j(\mathrm{d}v \mid a,x) - \int_s^{t^*}\int_w^{t^*}\frac{F_j(u \mid a,x) - F_j(w \mid a,x)}{S(w \mid a,x)}\,\mathrm{d}u\Lambda(dw \mid a,x)\right]\\
=&-\int_s^{t^*}\sum_{i=1,2}H_{ij}(w,t^* \mid a,x)\Lambda_i(\mathrm{d}w \mid a,x)
\end{aligned}$$

where $(*)$ follows from the backward equation (theorem 5, Gill and Johansen, 1990). Hence

$$H_j(\mathrm{d}s,t^* \mid a,x) = \sum_{i=1,2}H_{ij}(s,t^* \mid a,x)\Lambda_i(\mathrm{d}s \mid a,x).$$

From this expression we can derive a bound for (b) using integration by parts:

$$
\begin{aligned}
&\int_0^{t^*\wedge\tilde{T}} \sum_{i=1,2} \frac{\hat{H}_{ij}(s,t^* \mid a,x)}{g(s \mid a,x)} \,\mathrm{d}\hat{\Lambda}_i(s \mid a,x) - \int_0^{t^*\wedge\tilde{T}} \sum_{i=1,2} \frac{H_{ij}(s \mid a,x)}{g(s \mid a,x)} \,\mathrm{d}\Lambda_i(s \mid a,x) \\
&= \int_0^{t^*\wedge\tilde{T}} \frac{1}{g(s \mid a,x)} \,\mathrm{d}\left[\hat{H}_j(s,t^* \mid a,x) - H_j(s,t^* \mid a,x)\right] \\
&= \frac{1}{g(t^* \mid a,x)} \left[\hat{H}_j(t^*\wedge\tilde{T},t^* \mid a,x) - H_j(t^*\wedge\tilde{T},t^* \mid a,x)\right] \\
&\quad - \frac{1}{g(0 \mid a,x)} \left[\hat{H}_j(0,t^* \mid a,x) - H_j(0,t^* \mid a,x)\right] \\
&\quad - \int_0^{t^*\wedge\tilde{T}} \left[\hat{H}_j(s,t^* \mid a,x) - H_j(s,t^* \mid a,x)\right] \left(\frac{1}{g}\right)(ds \mid a,x) \\
&\leq 3\eta^{-1} \sup_{s\leq t^*} \left|\hat{H}_j(s,t^* \mid a,x) - H_j(s,t^* \mid a,x)\right| \\
&\leq C_{10} \sup_{s\leq t^*} \left|\hat{\Lambda}_1(s \mid a,x) - \Lambda_1(s \mid a,x)\right| + C_{11} \sup_{s\leq t^*} \left|\hat{\Lambda}_2(s \mid a,x) - \Lambda_2(s \mid a,x)\right| \qquad (35)
\end{aligned}
$$

for some constants $C_{10} > 0$ and $C_{11} > 0$. Applying the bounds (34) and (35) to (33) gives that (25) is bounded by

$$
\begin{aligned}
&E\left\{C_{10} \sup_{s\leq t^*} \left|\hat{\Lambda}_1(s \mid a,X) - \Lambda_1(s \mid a,X)\right| + C_{11} \sup_{s\leq t^*} \left|\hat{\Lambda}_2(s \mid a,X) - \Lambda_2(s \mid a,X)\right|\right. \\
&\left.+ C_{12} \sup_{s\leq t^*} \left|\hat{g}_c(s \mid a,X) - g(s \mid a,X)\right|\right\}^2 . \qquad (36)
\end{aligned}
$$

Thus, applying (29), (30) and (36) to (23), (24) and (25), respectively, gives the result. □

Proof of Theorem 1. Consider the decomposition

$$
\mathbb{P}_n^k \varphi(\hat{\nu}_{-k}) = \mathbb{P}_n^k \tilde{\psi}_{\psi_j} + (\mathbb{P}_n^k - P)(\varphi(\hat{\nu}_{-k}) - \varphi(\nu)) + P\varphi(\hat{\nu}_{-k})
$$

such that

$$
\begin{aligned}
\hat{\psi}_j^{CF} - \psi_j &= \sum_{k=1}^{K} \frac{n_k}{n} \mathbb{P}_n^k \varphi(\hat{\nu}_{-k}) - \psi_j \\
&= \mathbb{P}_n \tilde{\psi}_{\psi_j} + \underbrace{\sum_{k=1}^{K} \frac{n_k}{n} (\mathbb{P}_n^k - P)(\varphi(\hat{\nu}_{-k}) - \varphi(\nu))}_{\text{empirical process term}} + \underbrace{\sum_{k=1}^{K} \frac{n_k}{n} P(\varphi(\hat{\nu}_{-k}) - \psi_j)}_{\text{remainder term}} .
\end{aligned}
$$

Now, if both the empirical process term and the remainder term in the above display are $o_p(n^{-1/2})$, it follows that $\hat{\psi}_j^{CF}$ is asymptotically linear with influence function given by $\tilde{\psi}_j$. By Proposition 2 in Kennedy (2022) this is achieved if $\|\varphi(\hat{\nu}_{-k}) - \varphi(\nu)\| = o_p(1)$ for each k and if the remainder is $o_p(n^{-1/2})$. Under assumption B for each k, the former is achieved by lemma C.2 and the latter is achieved by lemma C.1. An application of the central limit theorem together with Slutsky's lemma gives the convergence in distribution. □

Proof of Theorem 2. We will show that $\hat{\Gamma}_j^{l,CF}$ and $\hat{\chi}_j^{l,CF}$ are asymptotically linear with influence function given by $\tilde{\psi}_{\Gamma_j^l}$ and $\tilde{\psi}_{\chi_j^l}$, respectively. Since $\hat{\Omega}_j^{l,CF}$ is a ratio of the two, it follows from the functional delta method that it is asymptotically linear with influence function given by $\tilde{\psi}_{\Omega_j}$ (van der Vaart, 2000, ch. 25.7).

That $\hat{\chi}^{l,CF}$ is asymptotically linear follows from Theorem 5 in Ziersen and Martinussen (2024). To show that $\hat{\Gamma}_j^{l,CF}$ is asymptotically linear, we consider the decomposition

$$\mathbb{P}_n^k \phi_{\Gamma_j^l}(\hat{\nu}_{-k}) = \mathbb{P}_n^k \tilde{\psi}_{\Gamma_j^l} + (\mathbb{P}_n^k - P)(\phi_{\Gamma_j^l}(\hat{\nu}_{-k}) - \phi_{\Gamma_j^l}(\nu)) + P\phi_{\Gamma_j^l}(\hat{\nu}_{-k})$$

such that

$$\begin{aligned}\hat{\Gamma}_j^{l,CF} - \psi_j &= \sum_{k=1}^K \frac{n_k}{n} \mathbb{P}_n^k \phi_{\Gamma_j^l}(\hat{\nu}_{-k}) - \Gamma_j^l \\ &= \mathbb{P}_n \tilde{\psi}_{\Gamma_j^l} + \underbrace{\sum_{k=1}^K \frac{n_k}{n}(\mathbb{P}_n^k - P)(\phi_{\Gamma_j^l}(\hat{\nu}_{-k}) - \phi_{\Gamma_j^l}(\nu))}_{\text{empirical process term}} + \underbrace{\sum_{k=1}^K \frac{n_k}{n} P(\phi_{\Gamma_j^l}(\hat{\nu}_{-k}) - \Gamma_j^l)}_{\text{remainder term}}.\end{aligned}$$

Again, by Proposition 2 in Kennedy (2022), the desired asymptotic linearity follows if we can show that $\left\|\phi_{\Gamma_j^l}(\hat{\nu}_{-k}) - \phi_{\Gamma_j^l}(\nu)\right\| = o_p(1)$ for each k and that the remainder term is $o_p(n^{-1/2})$. For both results, we will consider arguments that are similar to the ones given in the proof of Theorem 4 in Ziersen and Martinussen (2024).

Empirical process term

Consider the following expansion for a given k:

$$\begin{aligned}\phi_{\Gamma_j^l}(\hat{\nu}_{-k})(O) - \phi_{\Gamma_j^l}(\nu)(O) =& [\varphi_j(\hat{\nu}_{-k})(O) - \varphi_j(\nu)(O)][X_l - \hat{E}_{-k}^l(X_{-l})] \\ &- [\hat{\tau}_{j,-k}^l(X_{-l}) - \tau_j^l(X_{-l})][X_l - \hat{E}_{-k}^l(X_{-l})] \\ &- [\hat{E}_{-k}^l(X_{-l}) - E(X_l \mid X_{-l})][\varphi_j(\nu)(O) - \tau_j^l(X_{-l})].\end{aligned}$$

For the first term we have

$$\mathrm{E}\left\{[\varphi_j(\hat{\nu}_{-k})(O) - \varphi_j(\nu)(O)][X_l - \hat{E}_{-k}^l(X_{-l})]\right\}^2 \leq M\,\mathrm{E}\left\{\varphi_j(\hat{\nu}_{-k})(O) - \varphi_j(\nu)(O)\right\}^2 = o_p(1),$$

where the inequality follows from assumption (i) and the equality follows from lemma C.2, since we have assumed that assumption B2 holds for each k. Consistency in $L_2(P)$ of the second term follows by analogous arguments, replacing assumption B2 with assumption (ii). Consistency of the third term in $L_2(P)$ follows from assumption (iii) if $[\varphi_j(\nu)(O) - \tau_j^l(X_{-l})]^2$

is bounded almost surely. To that end, observe that for almost all $o \in \mathcal{O}$ we have

$$
\begin{aligned}
&[\varphi_j(\nu)(o) - \tau_j^l(x_{-l})]^2 \\
\leq & 2\left[\tau(x) - \tau_j^l(x_{-l})\right]^2 + 4\left[\sum_{i=1,2} \frac{\mathbb{1}(A=1)}{\pi(1 \mid x)} \int_0^{t^*} \frac{H_{ij}(s, t^* \mid 1, x)}{S_c(s \mid 1, x)} \,\mathrm{d}M_i(s \mid, a, x)\right]^2 \\
&+ 4\left[\sum_{i=1,2} \frac{\mathbb{1}(A=0)}{\pi(0 \mid x)} \int_0^{t^*} \frac{H_{ij}(s, t^* \mid 0, x)}{S_c(s \mid 0, x)} \,\mathrm{d}M_i(s \mid, 0, x)\right]^2 .
\end{aligned}
$$

Hence consistency of the empirical process term follows, if we can bound each term in the display above. The first term is clearly bounded, since τ is bounded by t^*. For the second and third term observe that

$$
\begin{aligned}
&\left[\sum_{i=1,2} \frac{\mathbb{1}(A=a)}{\pi(a \mid x)} \int_0^{t^*} \frac{H_{ij}(s, t^* \mid a, x)}{S_c(s \mid a, x)} \,\mathrm{d}M_i(s \mid, a, x)\right]^2 \\
\leq & 2\eta^{-2}\left[\sum_{i=1,2} \int_0^{t^*} H_{ij}(s, t^* \mid a, x) \,\mathrm{d}N_i(s \mid, a, x)\right]^2 \\
&+ 2\eta^{-2}\left[\sum_{i=1,2} \int_0^{t^*} H_{ij}(s, t^* \mid a, x) \mathbb{1}(\tilde{T} \geq s) \,\mathrm{d}\Lambda_i(s \mid, a, x)\right]^2 \\
\leq & 2\eta^{-2}\left[\sum_{i=1,2} 2\eta^{-1}t^*\right]^2 + 2\eta^{-2}\left[2\eta^{-1}t^* \sum_{i=1,2} \int_0^{t^*} \mathbb{1}(\tilde{T} \geq s) \,\mathrm{d}\Lambda_i(s \mid, a, x)\right]^2 \\
= & 32\eta^{-4}(t^*)^2 + 16\eta^{-4}(t^*)^2\left[\int_0^{t^*} \mathbb{1}(\tilde{T} \geq s) \,\mathrm{d}\Lambda(s \mid, a, x)\right]^2 \\
\leq & \eta^{-4}(t^*)^2(32 + 16\log(\eta^{-1})^2)
\end{aligned}
$$

where the first and third inequality follows from assumption B1 and the second inequality follows from (32). Thus it follows that $\left\|\phi_{\Gamma_j^l}(\hat{\nu}_{-k}) - \phi_{\Gamma_j^l}(\nu)\right\| = o_p(1)$ for each k.

Remainder term

As in Ziersen and Martinussen (2024), we consider the decomposition

$$
\begin{aligned}
P\{\phi_{\Gamma_j^l}(\hat{\nu}_{-k}) - \phi_{\Gamma_j}(\nu)\} = & E\left\{[\varphi(\hat{\nu}_{-k})(O) - \varphi_j(\nu)(O)][X_l - \hat{E}_{-k}^l(X_{-l})]\right\} \\
& - \mathrm{E}\left\{[\hat{\tau}_{j,-k}^l(X_{-l}) - \tau_j^l(X_{-l})][X_j - \hat{E}_{-k}^j(X_{-l})]\right\} \\
& - \mathrm{E}\left\{[\hat{E}_{-k}^l(X_{-j}) - E(X_l \mid X_{-l})][\varphi_j(\nu)(O) - \tau_j^l(X_{-l})]\right\} \qquad (37)
\end{aligned}
$$

and we want to show that each term is $o_p(n^{-1/2})$. For the third term we note that

$$\begin{aligned}
&E(\varphi_j(\nu)(O) - \tau_j^l(X_{-l}) \mid A, X) \\
=&\left(\frac{\mathbb{1}(A=1)}{\pi(1 \mid X)} - \frac{\mathbb{1}(A=0)}{\pi(0 \mid X)}\right) \sum_{i=1,2} E\left(\int_0^{t^*} \frac{H_{ij}(s, t^* \mid A, X)}{S_C(s \mid A, X)} \,\mathrm{d}M_i(s \mid A, X)\middle| A, X\right) \\
=&0
\end{aligned}$$

since the integral is a martingale conditional on A and X, and hence, that the third term is equal to 0 by iterated expectation. The second term is $o_p(n^{-1/2})$ by Cauchy-Schwarz together with assumption (ii) and (iii). Following the derivations in the proof of C.1, we can use iterated expectation to write the first term as

$$\begin{aligned}
\mathrm{E}\Bigg\{&[X_l - \hat{E}^l_{-k}(X_{-l})] \sum_{a=0,1}\sum_{i=1,2}\int_0^{t^*} S(s \mid a, X)\hat{H}_{ij}(s, t^* \mid a, X) \\
&\times\left(1 - \frac{\pi(a \mid X)S_C(s \mid a, X)}{\hat{\pi}(a \mid X)\hat{S}_C(s \mid a, X)}\right) \mathrm{d}\left[\hat{\Lambda}_i(s \mid a, X) - \Lambda_i(s \mid a, X)\right]\Bigg\}.
\end{aligned}$$

By assumption (i) this is bounded by

$$\begin{aligned}
\sqrt{M}\sum_{a=0,1}\mathrm{E}\Bigg\{\Bigg|&\sum_{i=1,2}\int_0^{t^*} S(s \mid a, X)\hat{H}_{ij}(s, t^* \mid a, X) \\
&\times\left(1 - \frac{\pi(a \mid X)S_C(s \mid a, X)}{\hat{\pi}(a \mid X)\hat{S}_C(s \mid a, X)}\right) \mathrm{d}\left[\hat{\Lambda}_i(s \mid a, X) - \Lambda_i(s \mid a, X)\right]\Bigg|\Bigg\}
\end{aligned}$$

which is $o_p(n^{-1/2})$ by assumption B3.

We have now shown that summands of the empirical process term and remainder term are $o_p(n^{-1/2})$ for each k, and hence it follows from Proposition 2 in Kennedy (2022) that $\hat{\Gamma}_j^{l,CF}$ is asymptotically normal with influence function given by $\tilde{\psi}_{\Gamma_j^l}$. The convergence in distribution of $\sqrt{n}(\hat{\Omega}_j^{l,CF} - \Omega_j^l)$ follows from the central limit theorem together with Slutsky's lemma. □