FACULTY OF HEALTH SCIENCES
UNIVERSITY OF COPENHAGEN

PhD thesis

Helene Charlotte Rytgaard
Section of Biostatistics, University of Copenhagen

Targeted causal learning for longitudinal data

Academic advisors:
Thomas Alexander Gerds
Claus Thorn Ekstrøm
Mark van der Laan

This thesis has been submitted to the
Graduate School of Health and Medical
Sciences, University of Copenhagen
on January 1, 2020

Targeted causal learning for longitudinal data

PhD thesis

Helene Charlotte Rytgaard

Section of Biostatistics
Department of Public Health
Faculty of Health and Medical Sciences
University of Copenhagen
Øster Farimagsgade 5, Entr. B.
DK-1014 Copenhagen K

Academic advisors:

Thomas Alexander Gerds, University of Copenhagen

Claus Thorn Ekstrøm, University of Copenhagen

Mark van der Laan, University of California, Berkeley

Preface

This thesis has been submitted to the Graduate School of Health and Medical Sciences, University of Copenhagen. The work included was carried out at the Section of Biostatistics between January 2017 and December 2019. Four months were spent visiting Professor Mark van der Laan at the University of California, Berkeley, during Spring 2018. The project was funded by Danmarks Frie Forskningsfond (DFF).

The thesis includes four manuscripts that develop methods for causal effect estimation in observational longitudinal studies where data are collected over time. The work is the product of collaborations with my supervisors Thomas Gerds, Claus Ekstrøm and Mark van der Laan, and with Lars Kessing who provided the application that served as motivation for much of the work. Two main directions have been taken in the manuscripts; targeted minimum loss-based estimation (TMLE) and random forest methodology. I refer to the proposed methods collectively as 'targeted' estimation methods. The introductory part of the thesis seeks to provide the reader a general background for the topics of the manuscripts, and for the methodological direction taken.

The problems considered in this thesis are not easy ones, and many questions remain open. I am very thankful to have had three years to immerse myself in this area, and I plan to continue my future research on the same problems.

Summary

This thesis develops statistical methodology for causal inference based on observational longitudinal data. Such data, that consist of repeated observations over a period of time, are common in medical research. Often we are interested in the estimation of a treatment effect on an outcome of interest, however, the non-experimental nature of the data gives rise to complex confounding patterns which must be accounted for in the statistical analysis.

During the three PhD years, the work developed into two related, but distinct, directions:

- Targeted minimum loss-based estimation, abbreviated TMLE; Manuscript I.
- Random forest methodology; Manuscripts II–IV.

The common basis is that learning from the data is 1) guided by semiparametric and nonparametric models to account for realistic complexity of the underlying data-generating mechanism, and 2) targeted towards low-dimensional statistical parameters with a causal interpretation under additional assumptions.

Chapter 1 motivates the research by considering a particular application from the Danish registries. Chapters 2–7 are intended to introduce the reader to relevant statistical topics as a background for the manuscripts. Chapter 8 summarizes Manuscripts I–IV. Chapter 9 discusses the proposed methodology and considers extensions and future directions.

Manuscript I is the main theoretical contribution of the thesis. It proposes an extension of the existing TMLE methodology to the continuous-time setting where changes in both covariates, treatment and outcome can happen on an arbitrarily fine timescale. The manuscript presents an efficiency theorem for the estimation problem and shows theoretically how an efficient estimator is constructed.

Manuscript II is a review of random forest algorithms for survival analysis, and Manuscript III and Manuscript IV are concerned with the adaptation of the generalized random forest methodology to the survival and competing risks setting. Manuscript IV is mostly focused on setting up a causal search machinery for variable importance analysis and takes an inverse weighting approach to target different parameters, while Manuscript III revises the tree building process itself for the analysis of time-to-event outcomes.

Resumé

Denne afhandling udvikler statistisk metode til kausal inferens for observationel longitudinal data. Sådanne data, der består af gentagne observationer over tid, forekommer ofte i medicinsk forskning. Vi er ofte interesseret i at estimere behandlingseffekter på et interesseoutcome, imidlertid giver den ikke-eksperimentielle natur anledning til komplekse konfoundingmønstre, som skal håndteres i den statistiske analyse.

I løbet af de tre år har PhD-arbejdet forgrenet sig i to relaterede, men forskellige, retninger:

- Targeted minimum loss based estimation, forkortet TMLE; Manuskript I.
- Random forest; Manuskript II–IV.

Den fælles basis består i, at læringen fra data er 1) guidet af semiparametriske og ikke-parametriske metoder for at tage højde for realistisk kompleksitet af den underliggende data-genererende mekanisme og 2) målrettet mod lavdimensionale statistiske parameter, der kan tilskrives en kausal fortolkning under supplerende antagelser.

Kapitel 1 motiverer forskningen ved at betragte en bestemt anvendelse fra de danske registre. Kapitel 2 til 7 har til hensigt at introducere læseren til relevant statistiske emner som baggrund for manuskripterne. Kapitel 8 opsummerer manuskripterne. Kapitel 9 diskuterer metoderne og gennemgår udvidelser og fremtidige retninger.

Manuskript I udgør det primære teoretiske bidrag i afhandlingen. Manuskriptet foreslår en udvidelse af existerende TMLE-metodik til kontinuert tid, hvor ændringer i kovariater, behandling og outcome kan ske på en arbitrært fin tidsskala. Manuskriptet præsenterer et teorem for efficient estimation og viser teoretisk, hvordan en efficient estimator kan konstrueres.

Manuskript II er en gennemgang af random forest-algoritmer til overlevelsesanalyse, og Manuskript III og Manuskript IV handler om en tilpasning af generalized random forest til overlevelsesanalyse og competing risks analyse. Manuskript IV er primært fokuseret på at konstruere en kausal søgealgoritme og foreslår sandsynlighedsvægtning for at målrette forskellige parametre, mens Manuskript III reviderer selve trækonstruktionen for at imødekomme analysen af overlevelsesdata.

Contents

Abbrevations

ATE	Average Treatment Effect
CAR	Coarsening At Random
iid	Independent and Identically Distributed
HAL	Highly Adaptive Lasso
IPCW	Inverse Probability of Censoring Weighting
IPTW	Inverse Probability of Treatment Weighting
IPW	Inverse Probability Weighting
TMLE	Targeted Minimum Loss-based (Maximum Likelihood) Estimation

Notation

O	Observed data
P_0	The distribution of O
$\mathcal{M}$	The statistical model assumed to contain P_0
$G = G(P)$	Interventional part of distribution $P \in \mathcal{M}$
$Q = Q(P)$	Non-interventional part of distribution $P \in \mathcal{M}$
$\Psi : \mathcal{M} \to \mathbb{R}$	Target parameter
$\psi_0 = \Psi(P_0)$	Value of the target parameter
$\mathbb{P}_n$	Empirical distribution of the data $O_1, \ldots, O_n$
Pf	$\int f \, dP$
$\mathbb{E}_P[X]$	The expectation of the random variable X under P: $\int X dP$
$\|f\|_P$	$\sqrt{Pf^2}$ (the $L_2(P)$-norm)
$\xrightarrow{P}$	Convergence in probability
$\xrightarrow{\mathcal{D}}$	Convergence in distribution
$X_n = o_P(R_n)$	$R_n^{-1} X_n \xrightarrow{P} 0$
$\sigma(X)$	σ-algebra generated by the stochastic variable X
$\mathbb{1}\{B\}$	Indicator for an event B

Chapter 1.
Introduction and overview

This thesis develops statistical methodology for causal inference based on observational longitudinal data. The present chapter provides an overall introduction. Section 1.1 describes the application that served as motivation for much of the work. Section 1.2 gives an overview of the thesis, including the introductory chapters and the contributions of the manuscripts. Section 1.3 gives a short account of the statistical framework that we work within.

1.1 Motivation

Data from Danish nation-wide population-based registries provide time and subject-specific information about purchases of all prescribed drugs, hospital admission and deaths. We are interested in drug repurposing, that is, the study of drugs already in clinical use and their potential use for treating other diseases than they were developed for. Our question is if we can identify drugs with unintended protecting effects against onset (recorded as a hospital admission) of depression and bipolar disorder. We find that two aspects are particularly important to consider when approaching this problem. We return to both repeatedly throughout the thesis.

First, the data are observational in nature and we should distinguish between associational and causal effects. We do not want to detect drug effects, or lack of same, that comes from confounding by indication. Think of the following example: A patient gets treated with aspirin due to severe headache. The same headache may cause depression for the patient. If we simply describe the association between depression and aspirin intake, it may appear that aspirin causes depression. Clearly, this is not the conclusion we want to make. Our questions are causal in nature and we ask if aspirin has a *causal effect* on depression, not if aspirin is a good *predictor* of depression.

Second, the effect of a drug treatment must be summarized by some parameter. Statistical inference deals with learning about such parameters based on a model for the distribution of the observed data. One may, for example, specify a parametric regression model for the underlying distribution. When the regression model is correctly specified, estimators based hereon are often consistent and asymptotically efficient. However, model misspecification may introduce severe asymptotic bias, and, moreover, when the model is revised in order to meet the assumptions the subsequent inference ignores that the model was actually data-adaptively selected and is thus in-

valid. In contrast, semiparametric models provide a much more flexible approach to learning from the data. Still, it would be a big misunderstanding to think that we can simply apply machine learning methods blindly to data. Parameter interpretability and statistical inference remain key to provide useful statistical methodology.

In the next section we briefly describe the analysis of the considered problem as we implemented and carried it out in Kessing et al. (2019a,b) (see Table 1.1). This is used to further motivate the direction taken in the thesis work. In both papers (Kessing et al., 2019a,b) the focus was on six drugs of interest that were a priori hypothesized to be associated with depression and/or bipolar disorder.

1.1.1 A Poisson regression approach

Data are obtained by linking Danish population-based registers using the unique personal identification number, which is assigned to all persons living in Denmark. In the considered study, data are available in a fixed calendar period $[0, \tau]$ and include daily information on prescribed medical purchases and hospital admissions. The exposure to a specific drug of interest for a given subject at time t, $A(t)$, is summarized as a grouping of the number of purchases made so far (in the considered time period) of that drug. We may choose to collect hospital admissions with other diagnoses than depression as well as purchases of other drugs in a time-varying covariate $L(t)$. Moreover, we denote by L_0 a vector of baseline covariates, including information such as sex and socioeconomic status. The subjects are included from the beginning of the registry (time $t = 0$) and followed until event of interest (first admission with depression), date of death or end of study (time $t = \tau$). Poisson regression can be used to fit a Cox regression model (Cox, 1972) with a piecewise constant baseline hazard rate. To apply Poisson regression to our data for analyzing associations between prescription history of purchases of the drug and hospital admission with the disease, we assume a piecewise constant hazard for our event of interest and postulate the following regression model:

$$\lambda(t \,|\, A(t), L(t)) = \sum_{k=1}^{K} \theta_k \, \mathbb{1}\{t \in (t_{k-1}, t_k]\} \, e^{\gamma A(t) + \beta^\top L(t) + \alpha^\top L_0}. \tag{1.1}$$

In this model, $\gamma \in \mathbb{R}$ represents the effect of the treatment and $\theta_k \geq 0$ the baseline hazard rate in the kth time interval where $\cup_{k=1}^{K}(t_{k-1}, t_k] = (0, \tau]$. When all the observed covariates are categorical, the parameters (γ, β, α) can be estimated by maximum likelihood using standard software, and, notably, the data is only needed in aggregated form with event counts for all possible combinations of covariate levels. The reason is that the likelihood for the model (1.1) is proportional to a Poisson

Kessing et al. (2019a)	*New drug candidates for bipolar disorder – a nation-wide population-based study*	Published in *Bipolar disorders*, 2019
Kessing et al. (2019b)	*New drug candidates for depression – a nation-wide population-based study*	Published in *Acta Psychiatrica Scandinavica*, 2019

Table 1.1: Main co-authored publications.

likelihood. Indeed, we can write the likelihood for (1.1) as follows,

$$\mathcal{L}_n(\gamma,\beta,\alpha) = \prod_{k=1}^{K}\prod_{j=1}^{J}\big(\theta_k \exp(\gamma a + \beta^\top \ell(j) + \alpha^\top \ell_0(j))\big)^{D_k(j)} \\ \exp\big(-\theta_k \exp(\gamma a + \beta^\top \ell(j) + \alpha^\top \ell_0(j)) R_k(j)\big),$$

where $D_k(j)$ is the total number of events observed in the time interval $(t_{k-1}, t_k]$ and $R_k(j)$ is the total observed risk time in $(t_{k-1}, t_k]$, both among subjects with $(A, L_i(t), L_{0,i}(t)) = (a, \ell(j), \ell_0(j))$ with $(\ell(j), \ell_0(j))_{j=1}^J$ denoting distinct combinations of values of observed covariates. In practice we fit a Poisson regression model with event counts $D_k(j)$, mean $\gamma A(t) + \beta^\top L(t) + \alpha^\top L_0$ and offset $\log R_k(j)$.

From association to causation

Poisson regression provides an efficient algorithm for estimation of and inference for associational effects of the treatment on the hazard scale. Now, what if we want to move from the Poisson regression to a causal interpretation? We point out the following issues that arise in this regard.

Model (mis)specification. Covariates and their different functional forms included in (1.1) can be chosen in many different ways corresponding to very different interpretations of the resulting estimand γ. As for example shown by Grøn et al. (2016), estimation based on a misspecified Poisson regression may reverse the sign of the effect estimates.

Time-dependent confounding. What time-varying covariates $L(t)$ should be included in (1.1)? We could choose $L(t)$ as the collection of purchases of other drugs and hospitalizations with other diseases. The problem is that we condition on $L(t)$ at any given time t, which means that we disregard any earlier treatment effect that was mediated through $L(t)$. We formalize the concept of time-dependent confounding in Chapter 2.

Parameter interpretation. In model (1.1), γ represents the treatment effect on the hazard scale. The interpretation is that it is the average over time of the log hazard ratio of increasing the number of prescribed purchases of the drug of interest, controlling for the included covariates at any time. Either we include time-varying covariates, but then we may block some of the effect, or we do not include them and fail to adjust properly for confounding by indication. Furthermore, the hazard is defined conditional on survival that may be affected by treatment. This means that the hazard ratio generally cannot be ascribed a causal interpretation if there is a treatment effect on survival (Hernán, 2010). In any case, it is hard to argue that the parameter γ represents anything but an associational effect.

1.2 Overview of thesis

In this thesis we take a different approach to the motivating problem from Section 1.1. Our overall aims are 1) to avoid the use of parametric techniques and concern ourselves with semiparametric and nonparametric statistical models and the incorporation of machine learning methods, and 2) to target low-dimensional statistical parameters with a clear interpretation representing our research question of interest. The general line of thought we take is due to important work by Robins (1986); Gill and Robins (2001); van der Laan and Robins (2003); van der Laan and Rubin (2006) among others.

The thesis contains four manuscripts with the common aim of developing causal estimation methods for observational longitudinal data like encountered in the motivating problem from Section 1.1. The first manuscript is concerned with a novel continuous-time generalization of targeted minimum loss based estimation (TMLE) (van der Laan and Rose, 2011, 2018) and constitutes the main theoretical contribution of the thesis:

Manuscript I. *Continuous-time targeted minimum-loss based estimation of intervention-specific mean outcomes.*

The three other manuscripts contain work concerned with random forest methodology (Breiman, 2001; Athey et al., 2019) for event history analysis:

Manuscript II. *Random forests for survival analysis.*

Manuscript III. *Application of generalized random forests for survival analysis.*

Manuscript IV. *Average treatment effects with generalized random forests for survival and competing risks analysis.*

The manuscripts focus on different aspects of complications due to right-censoring, competing risks and time-dependent confounding. The common basis is that observations are given at random times in some bounded time interval $[0, \tau]$. Table 1.2 gives an overview of the publication status of each manuscript.

Manuscript I	*In revision for the Annals of Statistics (revision due May 2020)*
Manuscript II	*Published in Wiley StatsRef: Statistics Reference Online (2018)*
Manuscript III	*Published in Proceedings of the 21st European Young Statisticians Meeting (2019)*
Manuscript IV	*In preparation*

Table 1.2: Publication status for the manuscripts included in this thesis.

The thesis consists of nine chapters followed by the manuscripts. Chapters 2–7 present an introduction to the topics of the manuscripts, reviewing established ideas and results from the literature to cover and relate the material. Some key references are Robins (1986); Andersen et al. (1993); Bickel et al. (1993); van der Vaart (2000); van der Laan and Robins (2003); van der Laan and Rose (2011, 2018); Wager and Athey (2018); Athey et al. (2019). Specifically, Chapters 2, 3, 4 and 7 seek to provide a general background for the work and the methodological direction taken, using different examples and data structures. Chapters 5 and 6, on the other hand, give an overview of concepts relevant for moving on to present the work of the manuscripts. Table 1.3 outlines the overall structure of the thesis.

As a reading guide we also note the following:

- A background for Manuscript I is provided by Chapters 2–6.
- A background for Manuscript II is provided by Chapters 5 and 7.
- A background for Manuscript III is provided by Chapters 2, 3 and 5–7.
- A background for Manuscript IV is provided by Chapters 2 and 5–7.

Chapter 8 summarizes the manuscripts. Chapter 9 discusses the thesis work and suggests directions for future work.

1.3 Targeted causal estimation

In this thesis we consider learning methods that are all *targeted* towards a particular parameter, the *target parameter*, that represents our research question of interest. Throughout, we assume that observations $O_1, \ldots, O_n \overset{iid}{\sim} P_0$ are given, where P_0 belongs to some statistical model $\mathcal{M}$ imposing as few restrictions as possible. We define the target parameter Ψ as a mapping from the model $\mathcal{M}$ to the real line.

Our approach for defining the target parameter follows Robins (1986, 1987). As we detail in Chapter 2, we introduce *counterfactual variables* to represent the outcome under hypothetical experiments and relate the counterfactuals to the observed data by structural assumptions. Often the distribution of the counterfactuals can be identified in the following way. We specify a statistical model for the data-generating

Chapter 2	Chapter 2 introduces counterfactuals, and characterizes time-dependent confounding and time-varying treatment regimes using a discrete longitudinal data setting.
Chapters 3, 4, 7	Chapters 3, 4 and 7 review central concepts from semiparametric efficiency theory, targeted minimum loss based estimation (TMLE) and random forests (particularly, generalized random forests) using a simple example without time (referred to as our 'running example' or Example 1).
Chapters 5, 6	Chapters 5 and 6 give an overview of concepts relevant for moving on to the manuscripts. Chapter 5 reviews counting processes for modeling events happening randomly in time. Sections 5.1.1–5.1.2 introduce censoring and competing risks. Chapter 6 collects some aspects of causal survival analysis.
Manuscript I	Manuscript I applies TMLE methodology to estimate treatment effects in a continuous-time longitudinal data setting with time-dependent covariates and time-dependent treatment strategies.
Manuscripts II–IV	Manuscript II reviews existing random forest algorithms for survival analysis. Manuscripts III and IV apply generalized random forest methodology for treatment effect estimation in right-censored and competing risks settings (in continuous time) with baseline covariates and baseline treatment. Manuscript IV implements a causal search algorithm for the motivating problem from Section 1.1.

Table 1.3: Overview of chapters, manuscripts and considered data structures.

distribution P_0 as,

$$\mathcal{M} = \{P = P_{Q,G} : Q \in \mathcal{Q}, G \in \mathcal{G}\},$$

so that any $P \in \mathcal{M}$ is factorized into a part $Q = Q(P)$, on which our target parameter depends, and a part $G = G(P)$ that represents the remaining factors of the distribution. We also write, with some abuse of notation, $\Psi(P) = \Psi(Q)$. The target parameter Ψ is now defined as a mapping of $P = P_{Q,G}$ through P_{Q,G^*}, given from

$P_{Q,G}$ by substituting an intervention G^* for G. Under causal assumptions, P_{Q,G^*} can be interpreted as a distribution of counterfactual variables. The target parameter is a statistical quantity, but can now be ascribed a particular (causal) interpretation under a set of assumptions.

Estimating $\psi_0 = \Psi(P_0)$ based on the observed data $O_1, \ldots, O_n$ involves estimation of $Q_0 = Q(P_0)$, which is some possibly infinite-dimensional quantity such as a collection of conditional expectations and (or) densities. To make sure to capture realistic complexity of Q_0, we should employ as flexible methods as possible (Bang and Robins, 2005; van der Laan and Rose, 2011). One attractive idea is to estimate Q_0 in a 'targeted way', by focusing the estimation of Q_0 towards optimal estimation of ψ_0.

Chapter 2.
Causal effects

The questions we are interested in are often not associational but rather causal in nature. Causal questions are about what would have happened in a system if it had been subjected to a particular change. For example, for our motivating problem introduced in Chapter 1: What would have happened if, contrary to fact, all patients had been treated with drug X compared to if no one had been treated with drug X?

The change corresponds to an intervention on parts of the data-generating system. We are interested in the behavior of the rest of the system under interventions. A causal model tells us how probabilities change as a result of such external interventions, and causal questions are conveniently framed in terms of counterfactuals (Rubin, 1974; Neyman, 1923; Robins, 1986, 1987) or structural causal models (SCMs) and causal graphs (Pearl, 1995, 2009; Spirtes et al., 2000). In this chapter we formulate causal questions and define causal parameters in terms of counterfactuals. The counterfactuals may be identifiable from the data simply as a consequence of study design, but otherwise we need to set up distributional assumptions that allow us to express the counterfactuals from what we actually observe, the factuals.

Counterfactuals allow us to define what it means for a variable to have a causal effect on another variable. Counterfactuals represent the ideal interventions: If we wanted to say something about whether a drug worked on a particular patient, we would want to see what happened in the two different scenarios where the patient 1) took the drug and 2) did not take the drug. In reality, however, the patient either took the drug or did not. Counterfactual outcomes are defined as the outcome under the two scenarios, one where the patient took the drug and one where the patient did not. We say there is a causal effect if there is a difference between, for example, the expected values of the two counterfactual outcomes.

To more formally introduce relevant concepts we consider in this chapter a general longitudinal data setting in which subjects of a population are all measured at equally spaced (clinical) visits such as regular doctor follow-up visits (Robins, 1986, 1987; Daniel et al., 2013; Robins, 1989; Robins et al., 1992, 2000; Bang and Robins, 2005; Stitelman et al., 2011; Petersen et al., 2014; Robins and Hernán, 2008; van der Laan and Rose, 2011, 2018). The aim of our presentation is both to introduce notation for a particular setting, and also to exemplify a more general practice:

1. Clearly define interventions and causal parameters of interest.
2. Establish conditions under which we can identify the causal parameter from the observed data.

Indeed, once we have completed 1. and 2., we have in our hands a plain statistical estimation problem. Later chapters will deal with estimation in such problems. It is a key message that this allows us to keep statistical and causal concepts separate: The statistical estimation problem starts with the estimation of the parameter; causality follows under 1. and 2. above.

2.1 Experimental and observational data

We usually make the distinction between *experimental data* and *observational data*. In experimental data, we make sure that data are generated such as to enable direct estimation of causal parameters. An example is a randomized controlled trial (RCT) that we could imagine: A randomized trial is initiated at a point in calendar time where subjects of a population are randomized to be either exposed or not exposed to a certain drug. The randomization implies that the causal effect of the drug can be estimated directly: The expected risk of onset of disease for the exposed group can be compared to the expected risk of onset of disease for the unexposed group. In observational data, on the other hand, subjects are treated for a reason. If we compute the expected risk of onset of disease for the exposed group compared to the unexposed group, the difference may not be due to the treatment of interest but rather due to various confounders, that is, other factors that predict both treatment and outcome of interest. We return to this in Section 2.2.1.

RCTs are often considered the golden standard, but may not always be feasible. We cannot randomize people to be obese, to have a higher education, or to smoke. In the motivating example from Chapter 1 we would not be able to conduct a randomized trial for all drugs on the market. Moreover, causal effects found in randomized trials may not generalize to the general population: Trial participants are, on average, healthier than typical patients for whom prescriptions are routinely written. Hence, it is generally of great interest to turn to observational data for causal effect estimation. The counterfactual framework allows us to define the RCT we would have liked to have conducted but did not, and tells us what we need to deal with to analyze the effect of interest.

2.2 Interventions, counterfactuals and the g-computation formula

We assume the subject-level observed data to be given as a temporally ordered vector O consisting of a time series of measurements of a treatment variable A over $K+1$ time points, $A = (A_0, \ldots, A_K)$, and a time series of covariate measurements over

$K+2$ time points, $L = (L_0, \ldots, L_{K+1})$. We let the outcome of interest be L_{K+1}, which we denote by Y and assume it to be real-valued or binary. (By doing so we note that $Y = L_{K+1}$ does not have the same dimension as L_k). We use upper case letters to denote random variables,

$$O = (L_0, A_0, L_1, A_1, \ldots, L_K, A_K, L_{K+1} = Y),$$

with values in a space $\mathcal{O}$, and lower case letters to denote realizations of the same variables $o = (\ell_0, a_1, \ell_1, a_1, \ \ldots, \ell_K, a_K, \ell_{K+1} = y)$. We use overbars to denote the history of a variable up to a particular time-point k, for instance $\bar{A}_k = (A_0, \ldots, A_k)$. For $k = 0, 1, \ldots, K$, we assume that A_k takes value in a finite set $\mathcal{A}$ and that L_k takes values in a subset of $\mathbb{R}^p$. Let $\mathcal{A}_k = \mathcal{A}^{k+1}$ denote the product space where the vector $\bar{A}_k$ takes its value. We follow Robins (1986, 1987) and rely on the counterfactual outcome framework (Rubin, 1974; Neyman, 1923) to formalize the notion of interventions and causal effects.

Counterfactuals. In the observed data, any individual follows a particular treatment regime $\bar{A}_K$ along which covariate variables $\bar{L}_K$ and lastly outcome $Y = L_{K+1}$ are observed. Additional to the *factual variables*, the observed variables, we hypothesize existence of *counterfactual variables* $\bar{L}_{K+1}^{\bar{a}_K}$: The variables that would have been observed, had this individual, possibly contrary to fact, been treated with $\bar{a}_K \in \mathcal{A}_K$ rather than the observed $\bar{A}_K$. In particular, $Y^{\bar{a}_K} = L_{K+1}^{\bar{a}_K}$ is the *counterfactual outcome* that would have been observed in this hypothetical treatment scenario.

Causal effects. The causal effect of A on Y can now be framed in terms of contrasts between $Y^{\bar{a}_K}$ for different treatment regimes $\bar{a}_K \in \mathcal{A}_K$. For example, if the distribution of $Y^{\bar{a}_K^1}$ equals the distribution of $Y^{\bar{a}_K^2}$ for all $\bar{a}_K^1, \bar{a}_K^2 \in \mathcal{A}_K$ we say that there is no causal effect of A on Y. On the other hand, to conclude a causal effect we must be clear about what treatment regimes $\bar{a}_K \in \mathcal{A}_K$ we are interested in and how we summarize the effect of a regime. For example, we may focus on two particular regimes $\bar{a}_K^1, \bar{a}_K^2 \in \mathcal{A}_K$ and measure the causal effect as the contrast $\mathbb{E}[Y^{\bar{a}_K^1}] - \mathbb{E}[Y^{\bar{a}_K^2}]$. One particular example is the expected difference between the outcome $Y^{\bar{a}_K}$ when treated according to the regime $\bar{a}_K$ and the outcome $Y^{\bar{0}_K}$ when never treated, defining,

$$\psi = \mathbb{E}[Y^{\bar{a}_K}] - \mathbb{E}[Y^{\bar{0}_K}]. \tag{2.1}$$

If $\psi \neq 0$ we say that there is a causal effect of regime $\bar{a}_K$, compared to never treated, on the outcome Y.

Example 1. *(Running example.) Consider the special case $K = 0$ with $O = (L, A, Y)$. Let $L \in \mathbb{R}^p$, $A \in \{0, 1\}$ and $Y \in \{0, 1\}$. Then there are two counterfactual outcomes, Y^1 and Y^0, the outcomes that would have been observed had the*

subject been treated and not treated with A, respectively. A causal parameter can now be defined as,

$$\psi = \mathbb{E}[Y^1] - \mathbb{E}[Y^0], \tag{2.2}$$

generally referred to as the average treatment effect (ATE). Other contrasts could also be of interest, such as risk ratios or odds ratios. ●

Observed distribution. We denote the distribution of the observed data O by P_0 and assume that $P_0 \in \mathcal{M}$. Let p denote the density (with respect to an appropriate dominating measure) of $P \in \mathcal{M}$. The density p factorizes according to the time order of the observations,

$$p(O) = \prod_{k=0}^{K} g_k(A_k \mid \bar{L}_k, \bar{A}_{k-1}) \prod_{k=0}^{K+1} q_k(L_k \mid \bar{L}_{k-1}, \bar{A}_{k-1}), \tag{2.3}$$

using the notation g_k for the conditional density of A_k given the variables preceding it and q_k for the conditional density of L_k given the variables preceding it. We parametrize p accordingly,

$$p = p_{q,g} = gq. \tag{2.4}$$

Note that, in the above, A_{-1}, L_{-1} by convention are the empty set.

Counterfactual distribution. Associated with each regime $\bar{a}_K \in \mathcal{A}_K$ there is a distribution $P^{\bar{a}_K}$ with density $p^{\bar{a}_K}$ representing the distribution of data had the subject, possibly contrary to fact, been treated according to that treatment regime. We refer to $P^{\bar{a}_K}$ as the counterfactual distribution or the postinterventional distribution. Quite generally, we define causal parameters in terms of characteristics of $P^{\bar{a}_K}$ for various $\bar{a}_K \in \mathcal{A}_K$, for example as in (2.1). The following provides conditions under which we can identify the counterfactual distribution $P^{\bar{a}_K}$, and by that the causal parameter, from the observed data for a given treatment regime $\bar{a}_K \in \mathcal{A}_K$. We also refer to Gill and Robins (2001) or Hernán and Robins (2020, Chapter 19).

(A1) *(Consistency).* $Y^{\bar{a}_K} = \bar{L}_{K+1}^{\bar{a}_K}$ is observed for those subjects who followed $\bar{a}_K$ in the observed data. Specifically, $Y^{\bar{a}_K} = Y$ if $\bar{A}_K = \bar{a}_K$. This assumption entails that one subject's counterfactual outcome does not depend on the treatment received by others, and that there is only one way to administrate the treatment.

(A2) *(Sequential randomization).* Conditional on the covariate and treatment history, the treatment is independent of the counterfactual outcome,

$$A_k \perp\!\!\!\perp Y^{\bar{a}_K} \mid (\bar{L}_k, \bar{A}_{k-1}),$$

for all $\bar{a}_K \in \mathcal{A}_K$ and all $k = 0, \ldots, K$. This is also called *no unmeasured confounding*, *exchangeability* or *ignorability*.

(A3) *(Sequential positivity).* The intervention that we impose on the treatment regime must have support in the data,

$$\sup_{O \in \mathcal{O}} \prod_{k=0}^{K} \frac{\mathbb{1}\{A_k = a_k\}}{g_k(A_k \,|\, \bar{L}_k, \bar{A}_{k-1})} < \infty \quad a.s.,$$

i.e., the treatment regime that we wish to measure the effect of must have been followed with a nonzero probability in the factual data.

Under (A1)–(A3), the counterfactual density is obtained from (2.4) simply by deleting the g-factor and replacing the conditioning sets in the q-factor by the interventional value:

$$p^{\bar{a}_K}(O) = \prod_{k=0}^{K+1} q_k(L_k \,|\, \bar{L}_{k-1}, \bar{A}_{k-1} = \bar{a}_k), \tag{2.5}$$

under which we can compute the expectation of Y under the intervention,

$$\mathbb{E}[Y^{\bar{a}_K}] = \int y \, dP^{\bar{a}_K}(o).$$

Formula (2.5) is due to seminal work by James Robins (Robins, 1986) and is commonly referred to as the *g-computation formula*.

Example 1. *(Continued.) We assume* (A1) $Y = Y^1 A + Y^0 (1 - A)$, (A2) $A \perp\!\!\!\perp (Y^1, Y^0) \,|\, L$, *and* (A3) $0 < P(A = 1 \,|\, L) < 1$ *a.s. We let* $q_Y(y \,|\, a, \ell) = P(Y = y \,|\, A = a, L = \ell)$, *for* $y = 0, 1$, *and* $q_L(\ell)$ *denote the density for the distribution of* L *with respect to a dominating measure* μ_L. *Then we can rewrite* (2.2) *as:*

$$\begin{aligned}
\psi &= \mathbb{E}[Y^{A=1}] - \mathbb{E}[Y^{A=0}] \\
&= \mathbb{E}\big[\mathbb{E}[Y^{A=1} \,|\, A = 1, L]\big] - \mathbb{E}\big[\mathbb{E}\big[Y^{A=0} \,|\, A = 0, L]\big] \\
&= \mathbb{E}\big[\mathbb{E}[Y \,|\, A = 1, L]\big] - \mathbb{E}\big[\mathbb{E}\big[Y \,|\, A = 0, L]\big],
\end{aligned}$$

or, equivalently, as

$$\psi = \int \big(\bar{Q}(1, L) - \bar{Q}(0, L)\big) \, q_L(\ell) \, d\mu_L(\ell),$$

with $\bar{Q}(A, L) := \mathbb{E}[Y \,|\, A, L]$. *Either way, the parameter* ψ *is now identifiable from the observed data.* ●

Static, dynamic and stochastic interventions. Defining $g_k^*(a_k \mid \bar{\ell}_k, \bar{a}_{k-1}) = \mathbb{1}\{A_k = a_k\}$ we can present (2.5) on a more general form, as,

$$p^{g^*}(O) = p_{q,g^*}(O) = \prod_{k=0}^{K} g_k^*(A_k \mid \bar{L}_k, \bar{A}_{k-1}) \prod_{k=0}^{K+1} q_k(L_k \mid \bar{L}_{k-1}, \bar{A}_{k-1}). \tag{2.6}$$

Setting $\bar{A}_k$ to a fixed value $\bar{a}_k$ is an example of a static intervention. More generally we can define interventions in terms of other choices of a density $g_k^*(A_k \mid \bar{L}_k, \bar{A}_{k-1})$ for the intervention variable A_k, often referred to as stochastic interventions (Robins et al., 2004; Dawid and Didelez, 2010; Muñoz and van der Laan, 2012). This includes static interventions (on the form $g_k^*(A_k \mid \bar{L}_k, \bar{A}_{k-1}) = \mathbb{1}\{A_k = a_k\}$) and dynamic interventions (Chakraborty and Moodie, 2013) (where $g_k^*(A_k \mid \bar{L}_k, \bar{A}_{k-1}) = d_k(\bar{L}_k)$ is a fixed rule applied to the observed history) as special cases. We still refer to (2.6) as the g-computation formula. Let $O^{g^*} = (L_0^{g^*}, A_0^{g^*}, L_1^{g^*}, A_1^{g^*}, \ldots, L_{K+1}^{g^*} = Y^{g^*})$ denote the counterfactual variables generated in a hypothetical world where the stochastic regime g^* is assigned rather than g. Under the assumptions that,[1]

$$Y^{g^*} = Y \text{ if } \bar{A}_K = \bar{A}_K^{g^*}, \tag{A1*}$$

$$A_k \perp\!\!\!\perp Y^{g^*} \mid (\bar{L}_k, \bar{A}_{k-1}) \quad \forall\, k = 0, \ldots, K, \tag{A2*}$$

$$\sup_{o \in \mathcal{O}} \prod_{k=0}^{K} \frac{g_k^*(A_k \mid \bar{L}_k, \bar{A}_{k-1})}{g_k(A_k \mid \bar{L}_k, \bar{A}_{k-1})} < \infty \quad a.s., \tag{A3*}$$

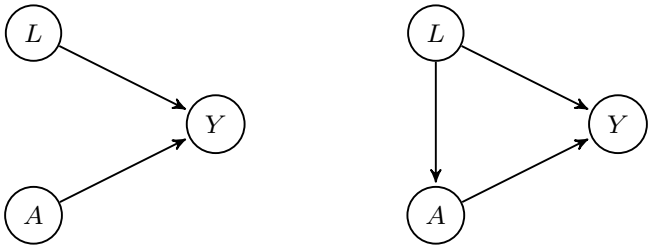
the distribution $P^{g^*} = P_{q,g^*}$ of O^{g^*} is characterized by the density p^{g^*} given by the g-computation formula (2.6) (Gill and Robins, 2001). The causal parameter can now be defined as a contrast under counterfactual distributions P_{q,g^*} with g^* varying over a set of interventions (static, dynamic, stochastic).

2.2.1 Confounding and time-dependent confounding

We consider again our motivating problem from Chapter 1 and let L represent hospital admissions (other than depression), A purchases of a particular drug of interest and Y a final outcome of interest. In the following we make use of directed acyclic graphs (DAGs)[2] (Pearl, 1995, 2009) to display various causal assumptions.

[1] We note that the consistency assumption becomes a bit more involved for stochastic interventions. We have tried to give a short account here but otherwise refer to Gill and Robins (2001, Assumption A1**).

[2] DAGs are useful for displaying conditional independencies, with nodes denoting variables and arrows between nodes representing the assumed direction of a causal influence. For a DAG to be causal we need to include both observed and unobserved variables such that if any observed variables share a cause, that cause is represented as an unobserved variable on the graph. Conditional independencies are identified if they are d-separated by a set of nodes in the graph, see Pearl (2009, Definition 1.2.3).



(a) Randomized treatment. **(b)** Time-fixed confounding.

Figure 2.1: Single time-point setting.

Time-fixed confounding. Assume that $K = 0$. Then we are in the simple setting with $O = (L, A, Y)$ (Example 1) that includes only baseline confounding followed by a single treatment decision. In Figure 2.1 we display these three variables as nodes in a DAG. In Figure 2.1a there is no arrow from L to A which means that L is not a cause of A and thus not a confounder: We have that $Y^a \perp\!\!\!\perp A$. In Figure 2.1b, L is a confounder for the effect of A on Y. In this case we have that $Y \perp\!\!\!\perp A^a \,|\, L$.

Time-dependent confounding that is not affected by treatment. For simplicity we consider $K = 1$, so that $O = (L_0, A_0, L_1, A_1, L_2 = Y)$. In Figure 2.2b we display the variables as nodes and include arrows where we allow a direct causal effect. In this case we allow for time-varying confounding on the treatment decisions at the two time-points, but we assume that baseline treatment A_0 does not affect the subsequent covariate L_1. We then have $Y \perp\!\!\!\perp A_0 \,|\, L_0$ and $Y \perp\!\!\!\perp A_1 \,|\, (L_0, L_1)$, and controlling for (L_0, L_1) allows identifiability of the effect of (A_0, A_1) on Y. Our data from Chapter 1 would fit into this setting if it was certain that hospital admissions do not to affect later treatment decisions.

Time-dependent confounding that is also affected by treatment. We consider again $O = (L_0, A_0, L_1, A_1, L_2 = Y)$. In Figure 2.2c we display the same graph as in Figure 2.2b, however, we also include the arrow $A_0 \to L_1$. In this case, we allow for time-varying confounding on the treatment decisions at the two time-points and we also allow for treatment to have an effect on later covariates. Then we have $Y \perp\!\!\!\perp A_0 \,|\, L_0$ and $Y \perp\!\!\!\perp A_1 \,|\, (L_0, A_0, L_1)$, but we do *not* have $Y \perp\!\!\!\perp A_1 \,|\, (L_0, L_1)$. Controlling for (L_0, L_1) includes the collider L_1 and does not allow identifiability of the effect of (A_0, A_1) on Y. In the data setting from Chapter 1, hospital admissions are an example of time-dependent confounding if we for instance believe that they are affected by earlier treatment and that treatment affects later hospital admissions.

Figure 2.2a and Figure 2.2d show two other settings, one with no measured and no unmeasured confounding (Figure 2.2a) and one with both measured and unmeasured time-dependent confounding (Figure 2.2d). For the latter, we cannot identify the causal parameter.

2.3 Final comments: Target parameter

In Section 1.3 we presented our general estimation problem: Given observed data $O \sim P_0$, where P_0 belongs to a statistical model $\mathcal{M}$,

$$\mathcal{M} = \{P = P_{Q,G} : Q \in \mathcal{Q}, G \in \mathcal{G}\},$$

we are interested in the target parameter $\Psi : \mathcal{M} \to \mathbb{R}$ defined as a mapping of P through P_{Q,G^*} with G^* an intervention substituted for G. This is the g-computation formula. Usually we work directly with the g-computation formula and provide the additional causal assumptions under which it has a desired interpretation. In this chapter we had $Q = q$ and $G = g$ and the g-computation formula P_{q,g^*} characterized by the density defined in (2.6). Now, for example, we may define our target parameter as:

$$\Psi(P) = \mathbb{E}_{P_{q,g^*}}[Y] = \int_{\mathcal{O}} Y \, dP_{q,g^*}. \tag{2.7}$$

This is a statistical parameter, but, as described in Section 2.2, it can be given a causal interpretation as the counterfactual mean outcome $\Psi(P) = \mathbb{E}[Y^{g^*}]$ under additional causal identifiability assumptions.

In the next couple of chapters, we will work mostly with the setting of Example 1 where $K = 0$ and $O = (L, A, Y)$. Again we assume that $O \sim P_0$ where P_0 belongs to a statistical model $\mathcal{M}$. In this example we target a parameter $\Psi : \mathcal{M} \to \mathbb{R}$ that is defined as the following difference:

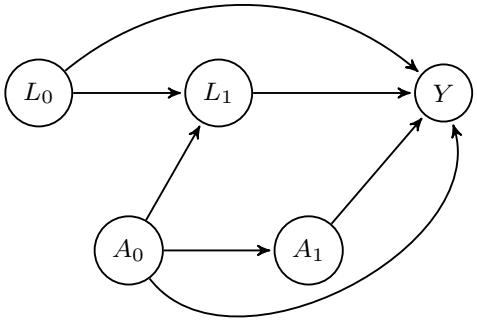
$$\Psi(P) = \int \left(\bar{Q}(1, L) - \bar{Q}(0, L)\right) q_L(\ell) \, d\mu_L(\ell), \qquad P \in \mathcal{M}. \tag{2.8}$$

Note that $\Psi(P)$ only depends on P through $Q = (\bar{Q}, q_L)$. Under the identifiability assumptions (see Example 1, page 13), this parameter has the desired interpretation,

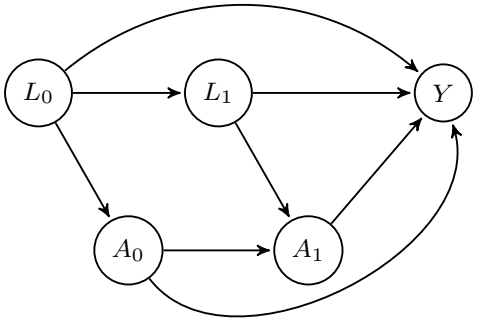
$$\Psi(P) = \mathbb{E}[Y^1] - \mathbb{E}[Y^0],$$

but the statistical estimation problem deals directly with $\Psi(P)$ as defined in (2.8).

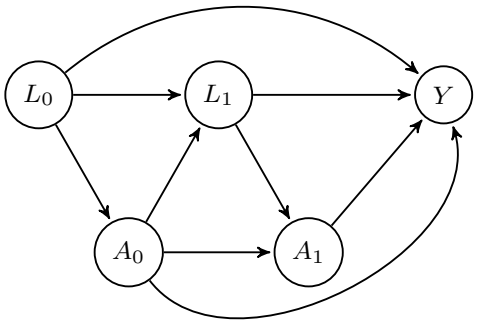
We return to the longitudinal setting again in Section 4.3. In Chapter 5 and Chapter 6 we move on to right-censored data, that gives rise to yet another layer of counterfactual reasoning.



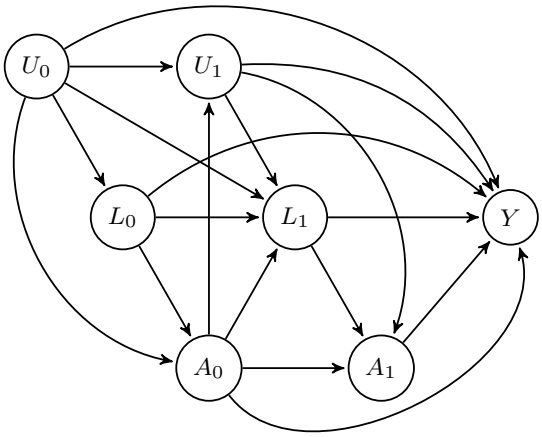
(a) No unmeasured or measured confounding for the time-varying treatment.



(b) Time-varying confounding that is not affected by earlier treatment.



(c) Time-dependent confounding that is affected by earlier treatment.



(d) Measured and unmeasured time-dependent confounding.

Figure 2.2: A number of situations that can occur in the time-varying setting. A sequential randomized study is presented in **(a)**. Here the counterfactual mean outcome $\mathbb{E}[Y^{g*}]$ is simply the mean outcome among those who followed treatment regime g^*. In observational studies represented by **(b)** and **(c)**, $\mathbb{E}[Y^{g*}]$ is identifiable under all strategies g^*. In observational studies represented by **(d)** no counterfactual mean outcome $\mathbb{E}[Y^{g*}]$ can be identified.

Chapter 3.
Semiparametric efficiency

In this chapter we review some fundamental results of semiparametric efficiency theory (Bickel et al., 1993). We do not consider this in depth, but rather make an informal overview based on Bickel et al. (1993); van der Vaart (2000); van der Laan and Robins (2003); Tsiatis (2007); van der Laan and Rose (2011, Appendix A).

We consider the general estimation problem as introduced in Section 1.3 and Chapter 2. Our aim is to estimate the target parameter $\psi_0 = \Psi(P_0) \in \mathbb{R}$ that behaves well under as few restrictions as possible on the statistical model $\mathcal{M}$ for the data-generating distribution P_0. We distinguish between *parametric models* that can be indexed by a finite-dimensional parameter $\theta \in \mathbb{R}^d$, for some $d \in \mathbb{N}$, and *semiparametric models* that cannot be indexed by a finite-dimensional $\theta \in \mathbb{R}^d$ alone. In this terminology, semiparametric models cover also *nonparametric models* that impose no parametric assumptions on any part of the data-generating distribution. The level of restrictions that we impose on $\mathcal{M}$ is generally a tradeoff between flexibility and robustness on the one hand, and statistical inference and efficiency on the other. Nonparametric models provide a more general and adaptive approach to learning from the data by allowing for realistic complexity of the true underlying probability distribution P_0, but the same adaptivity makes estimation and inference a more difficult exercise.

Results from semiparametric efficiency theory gives us efficiency bounds in models with minimal restrictions on the data-generating process. To understand the implications for our estimation problem, we need some technical tools. What does it mean for a semiparametric estimator to be *asymptotically efficient*? Summarizing existing work, we aim for an intuitive and practical understanding of the concepts in this regard.

To start out, we present some notation. We use $\mathbb{P}_n$ to denote the empirical distribution of the data: For $f : O \mapsto \mathbb{R}$, we have $\mathbb{P}_n f = \frac{1}{n}\sum_{i=1}^n f(O_i)$. For general P, we use $Pf = \int f \, dP$ to denote the expectation with respect to P. We further define $L_2(P)$-norm $\|f\|_P = \sqrt{Pf^2}$. Moreover, we adopt the use of stochastic o-notation:

$$X_n = o_P(R_n) \quad \text{meaning} \quad R_n^{-1} X_n \xrightarrow{P} 0 \text{ as } n \to \infty.$$

Finally, $\sigma(X)$ denotes the σ-algebra generated by the stochastic variable X.

3.1 Asymptotically linear estimators

Generally, we are interested in estimators that are *asymptotically linear*. We say that an estimator $\hat{\Psi}_n$ of ψ_0 is asymptotically linear if there exists a function $\phi(P) : \mathcal{O} \to \mathbb{R}$ (indexed by $P \in \mathcal{M}$) such that,

$$\sqrt{n}\big(\hat{\Psi}_n - \psi_0\big) = \sqrt{n}\,\mathbb{P}_n\,\phi(P_0) + o_P(1), \tag{3.1}$$

where $P_0\phi(P_0) = 0$ and $P_0\phi(P_0)^2 < \infty$. The function $\phi(P_0)$ is referred to as the *influence function* of the estimator $\hat{\Psi}_n$.

It is a simple consequence of Slutsky's theorem and the central limit theorem that (3.1) implies asymptotic normality,

$$\sqrt{n}\big(\hat{\Psi}_n - \psi_0\big) \xrightarrow{\mathcal{D}} N(0, \sigma_0^2),$$

with $\sigma_0^2 = P_0\phi(P_0)^2$. Thus, the asymptotic distribution of an asymptotically linear estimator, based on which we can construct confidence intervals and provide p-values for hypothesis tests, is identified entirely through its influence function. In addition, we can compare competing estimators for a specific parameter of interest by looking at the variances of their influence functions.

Example 1. *(Continued.) Recall that $g(A \mid L)$ denotes the conditional distribution of A given L. We may estimate our target parameter by means of inverse probability of treatment weighting (IPTW),*

$$\hat{\psi}_n^{\text{IPTW}} = \frac{1}{n}\sum_{i=1}^{n} \frac{(2A_i - 1)Y_i}{\hat{g}_n(A_i \mid L_i)}, \tag{3.2}$$

for an estimator $\hat{g}_n(A \mid L)$ of $g(A \mid L)$. Suppose we know the true g. Then $\hat{\psi}_n^{\text{IPTW}}$ is a linear estimator and thereby also an asymptotically linear estimator, and we see that,

$$\phi_{\text{IPTW}}(P)(O_i) = \frac{\mathbb{1}\{A_i = 1\} - \mathbb{1}\{A_i = 0\}}{g(A_i \mid L_i)} Y_i - \Psi(P), \tag{3.3}$$

is the influence function of $\hat{\psi}_n^{\text{IPTW}}$. ●

In the next section we introduce the *efficient influence function*, also known as the *canonical gradient*, that characterizes the "best variance" and thereby efficient estimation in the semiparametric model. Next we get into how we can construct estimators that have influence function equal to the efficient influence function.

3.2 Semiparametric efficiency

Our aim is to estimate $\psi_0 = \Psi(P_0)$ for a given parameter $\Psi : \mathcal{M} \to \mathbb{R}$. For starters, note that it is harder to estimate ψ_0 for all $P \in \mathcal{M}$ than it is for some submodel $\mathcal{M}' \subset \mathcal{M}$. For smooth parametric models $\mathcal{M}' = \{P_\theta : \theta \in \mathbb{R}^d\}$, classical statistical theory gives the Fisher information for estimating $\Psi(P_\theta)$. The information for estimating $\Psi(P_\theta)$ for the semiparametric model $\mathcal{M}$ is defined as the infimum of the information of all parametric submodels (van der Vaart, 2000).

Let $L_2(P)$ denote the Hilbert space of all functions $f : \mathcal{O} \to \mathbb{R}$ with $Pf = 0$ and $Pf^2 < \infty$ endowed with inner product $\langle f_1, f_2 \rangle_P = Pf_1f_2$. We are interested in a particular subspace $\dot{\mathcal{P}}(P) \subset L_2(P)$ known as the *tangent space*. As a technical device, we consider smooth one-dimensional submodels $\mathcal{M}' = \{P_\varepsilon : \varepsilon \in \mathbb{R}\} \subset \mathcal{M}$ (see Figure 3.1) that are constructed such that they pass through P at $\varepsilon = 0$ and has score $S \in L_2(P)$ given as,

$$S = \frac{d}{d\varepsilon}\bigg|_{\varepsilon=0} \log dP_\varepsilon. \tag{3.4}$$

If we consider a rich collection of such submodels we obtain a corresponding collection of score functions. The tangent space of the model $\mathcal{M}$ at P is defined as the closure of the linear span of the collection of score functions (van der Vaart, 2000).

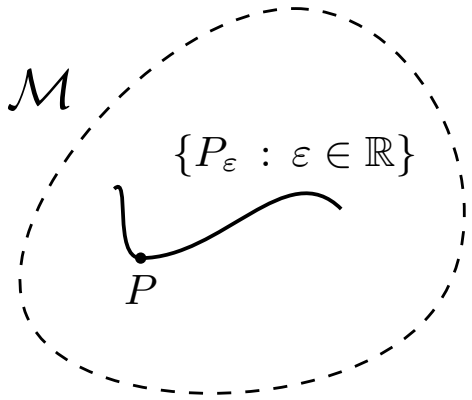


Figure 3.1: Illustration of a parametric submodel $\{P_\varepsilon : \varepsilon \in \mathbb{R}\} \subset \mathcal{M}$.

The parameter $\Psi : \mathcal{M} \to \mathbb{R}$ is *pathwise differentiable* at P if there exists $\phi(P) : \mathcal{O} \to \mathbb{R}$ such that for every score $S \in \dot{\mathcal{P}}(P)$ and submodel P_ε with score S,

$$\frac{d}{d\varepsilon}\bigg|_{\varepsilon=0} \Psi(P_\varepsilon) = P\phi(P)S. \tag{3.5}$$

Any function $\phi(P) \in L_2(P)$ that fulfills (3.5) is called a *gradient*. The gradient that is an element of the tangent space $\mathscr{T}(P)$ is referred to as the *canonical gradient*. We denote it $\phi^*(P)$. For any $\phi^\perp$ in the orthogonal complement of the tangent space, we have that,

$$P(\phi^*(P) + \phi^\perp)S = P\phi^*(P)S + P\phi^\perp S = P\phi^*(P)S,$$

for all $S \in \mathscr{T}(P)$. This implies that $\phi^*(P) + \phi^\perp$ for any $\phi^\perp$ is a gradient. In fact any gradient can be represented as $\phi^*(P) + \phi^\perp$ for some $\phi^\perp$, and the canonical gradient can be obtained by projecting any gradient onto the tangent space. Moreover, the canonical gradient $\phi^*(P)$ is the gradient that has the smallest variance of all gradients,

$$P(\phi^*(P) + \phi^\perp)^2 = P(\phi^*(P))^2 + P(\phi^\perp)^2 \geq P(\phi^*(P))^2.$$

We formulate this as the following lemma.

Lemma 3.1 (van der Vaart (2000), Lemma 25.19)**.** *The canonical gradient $\phi^*(P)$ defines a lower bound for estimating $\Psi(P)$ in the model $\mathcal{M}$ in the sense that,*

$$P\phi(P)^2 \geq P\phi^*(P)^2,$$

for all $\phi(P) \in L_2(P)$.

We restrict attention to asymptotically linear estimators that also satisfy a certain regularity condition. One can show that the influence function of any regular and asymptotically linear estimator equals a gradient of the pathwise derivative. According to Lemma 3.1 the canonical gradient is the gradient with the smallest variance. Thus the canonical gradient characterizes the efficient estimator (see also van der Vaart, 2000, Section 25.3).

Lemma 3.2. *An estimator $\hat{\Psi}_n^*$ is asymptotically efficient among all regular and asymptotically linear estimators if and only if,*

$$\sqrt{n}\big(\hat{\Psi}_n^* - \psi_0\big) = \sqrt{n}\,\mathbb{P}_n\phi^*(P_0) + o_P(1),$$

i.e., $\hat{\Psi}_n^*$ is asymptotically linear with influence function equal to the canonical gradient.

Since an asymptotically linear estimator is asymptotically efficient if and only if its influence function equals the canonical gradient, the canonical gradient is also naturally referred to as the *efficient influence function*. Thus, a key part of constructing an asymptotically efficient estimator of $\Psi(P)$ given the model $\mathcal{M}$ is to derive the canonical gradient.

3.2.1 A recipe for deriving the canonical gradient

Throughout, as outlined in Section 1.3, we consider models $\mathcal{M} = \{P = P_{Q,G} : Q \in \mathcal{Q}, G \in \mathcal{G}\}$ and a target parameter $\Psi : \mathcal{M} \to \mathbb{R}$ that depends only on P through the part $Q = Q(P)$. Let $\mathscr{T}_Q(P)$ denote the tangent space at P in the submodel $\mathcal{M}_Q \subset \mathcal{M}$ that assumes G to be known. We can derive the efficient influence function by characterizing the tangent space $\mathscr{T}_Q(P)$ and providing the projection formula; then the canonical gradient can be found as the projection of any other gradient of the pathwise derivative of $\Psi : \mathcal{M}_G \to \mathbb{R}$ at P onto $\mathscr{T}_Q(P)$ (van der Laan and Robins, 2003). Recall that the projection of a function $\phi(P)$ onto a subspace $\mathscr{T}(P)$ of a Hilbert space, denoted $\Pi(\phi(P) \,|\, \mathscr{T}(P))$, is defined by the two properties:

1. $\Pi(\phi(P) \,|\, \mathscr{T}(P)) \in \mathscr{T}(P)$,
2. $\big\langle \Pi(\phi(P) \,|\, \mathscr{T}(P)) - \phi(P), f \big\rangle_P = 0$ for any $f \in \mathscr{T}(P)$.

As an initial gradient, one can use the influence function of an estimator that takes the intervention part as known (van der Laan and Robins, 2003; van der Laan and Rose, 2011, Appendix A.5, A.7).

Example 1. *(Continued.) We have $Q = (q_Y, q_L)$ and $G = g$. We can use the influence function $\phi_{\mathrm{IPTW}}(P)$ from* (3.3) *as an initial gradient. Following the lines of van der Laan and Rose (2011, Appendix A.7) we note that the factorization of the density $p(o) = q_Y(y \,|\, a, \ell)\, g(a \,|\, \ell)\, q_L(\ell)$ implies an orthogonal decomposition of the tangent space,*

$$\mathscr{T}_Q(P) = \mathscr{T}_{q_Y}(P) \oplus \mathscr{T}_{q_L}(P),$$

and a corresponding decomposition of the projection operator,

$$\Pi(\phi(P) \,|\, \mathscr{T}(P)) = \Pi(\phi(P) \,|\, \mathscr{T}_{q_Y}(P)) + \Pi(\phi(P) \,|\, \mathscr{T}_{q_L}(P)).$$

The projections onto $\mathscr{T}_{q_Y}(P)$ and $\mathscr{T}_{q_L}(P)$ are characterized by,

$$\begin{aligned} \Pi(\phi(P) \,|\, \mathscr{T}_{q_L}(P)) &= \mathbb{E}[\phi(P) \,|\, L], \\ \Pi(\phi(P) \,|\, \mathscr{T}_{q_Y}(P)) &= \phi(P) - \mathbb{E}[\phi(P) \,|\, A, L]. \end{aligned}$$

Projecting $\phi_{\mathrm{IPTW}}(P)$ onto $\mathscr{T}(P)$ now gives an expression for the canonical gradient,

$$\begin{aligned} \phi^*(P)(O_i) = \left(\frac{\mathbb{1}\{A_i = 1\} - \mathbb{1}\{A_i = 0\}}{g(A_i \,|\, L_i)} \right) \big(Y - \bar{Q}(A_i, L_i)\big) \\ + \big(\bar{Q}(1, L_i) - \bar{Q}(0, L_i)\big) - \Psi(P), \end{aligned} \tag{3.6}$$

where $g(a \,|\, L) = P(A = a \,|\, L)$ and $\bar{Q}(A, L) = \mathbb{E}[Y \,|\, A, L]$. ●

3.3 Efficient substitution estimation

Let $\hat{\Psi}_n = \Psi(\hat{P}_n)$ be a substitution estimator based on an estimator $\hat{P}_n$ of (relevant parts of) P_0. We say that $\hat{\Psi}_n$ *solves the efficient influence curve equation* if,

$$\mathbb{P}_n \phi^*(\hat{P}_n) = o_P(n^{-1/2}). \tag{3.7}$$

By Lemma 3.2, the canonical gradient characterizes efficient estimation. In this section we sketch how the asymptotic behavior of an estimator that solves the efficient influence curve equation can be analyzed. We define the remainder term,

$$R_2(P, P_0) := \Psi(P) - \Psi(P_0) + P_0\phi^*(P).$$

Pathwise differentiability of $\Psi : \mathcal{M} \to \mathbb{R}$ implies a first-order expansion,

$$\begin{aligned}
&\Psi(P) - \Psi(P_0) \\
&\quad = -P_0\phi^*(P) + R_2(P, P_0) \\
&\quad = (\mathbb{P}_n - P_0)(\phi^*(P) - \phi^*(P_0)) + (\mathbb{P}_n - P_0)\phi^*(P_0) - \mathbb{P}_n\phi^*(P) + R_2(P, P_0).
\end{aligned}$$

Specifically, substituting the estimator $\hat{P}_n$ for P in the above expression gives a decomposition for the estimator $\hat{\Psi}_n = \Psi(\hat{P}_n)$,

$$\begin{aligned}
\Psi(\hat{P}_n) - \Psi(P_0) = {} & \mathbb{P}_n\phi^*(P_0) \\
& - P_0\phi^*(P_0) - \mathbb{P}_n\phi^*(\hat{P}_n) \\
& + (\mathbb{P}_n - P_0)(\phi^*(\hat{P}_n) - \phi^*(P_0)) && \text{(C1)} \\
& + R_2(\hat{P}_n, P_0). && \text{(C2)}
\end{aligned}$$

If $\hat{P}_n$ solves the efficient influence curve equation (3.7), then $\mathbb{P}_n\phi^*(\hat{P}_n) = o_P(n^{-1/2})$. Furthermore, we have that $P_0\phi^*(P_0) = 0$. We now see that $\hat{\Psi}_n = \Psi(\hat{P}_n)$ is asymptotically linear and efficient if we can construct $\hat{P}_n$ such that (C1) and (C2) are $o_P(n^{-1/2})$, since we then have,

$$\sqrt{n}\big(\hat{\Psi}^*_n - \psi_0\big) = \sqrt{n}\,\mathbb{P}_n\phi^*(P_0) + o_P(1).$$

Thus asymptotic linearity and efficiency $\hat{\Psi}^*_n$ follows from Lemma 3.2. What remains is to show that the terms in (C1) and (C2) are $o_P(n^{-1/2})$.

Remainder term. The last term (C2) is a remainder term for the expansion $R_2(P, P_0) = \Psi(P) - \Psi(P_0) + P_0\phi^*(P)$ and must be analyzed for the particular problem at hand. Recall that the target parameter only depends on P through $Q = Q(P)$. The canonical gradient and the remainder, on the other hand, typically depend on

both $Q = Q(P)$ and $G = G(P)$. In certain problems, the remainder $R_2(P, P_0)$ can be represented in terms of expressions like,

$$\int \big(H_1(Q) - H_1(Q_0)\big)\big(H_2(G) - H_2(G_0)\big) f(P, P_0) dP_0, \tag{3.8}$$

where $H_1(\cdot), H_2(\cdot)$ are some functions of Q and G, respectively. We say that $R_2(P, P_0)$, or, equivalently, the canonical gradient, admits a *double robustness* structure (van der Laan and Robins, 2003) with respect to misspecification of Q_0 or G_0,

$$P_0\phi^*(P) = \Psi(P_0) - \Psi(P), \quad \text{if} \quad G(P) = G(P_0) \text{ or } Q(P) = Q(P_0),$$

see also van der Laan (2017). The double robustness property implies that if an estimator $\hat{P}_n = (\hat{Q}_n, \hat{G}_n)$ solves the efficient influence curve equation, then $\hat{\Psi}_n = \Psi(\hat{P}_n)$ will be a consistent estimator of ψ_0 if either $\hat{Q}_n$ or $\hat{G}_n$ is a consistent estimator.

Example 1. *(Continued.) The remainder is given as (see, e.g. van der Laan, 2017),*

$$\begin{aligned} R_2(P, P_0) &= \Psi(P) - \Psi(P_0) + P_0\phi^*(P) \\ &= \sum_{a=0,1} (2a-1) \int \frac{g(a \mid L) - g_0(a \mid L)}{g(a \mid L)} \big(\bar{Q}(a, L) - \bar{Q}_0(a, L)\big)\, dP_0, \end{aligned} \tag{3.9}$$

with $g(a \mid L) = P(A = a \mid L)$ and $\bar{Q}(A, L) = \mathbb{E}[Y \mid A, L]$. We note that $R_2(P, P_0)$ in (3.9) admits a double robustness structure which implies that $R_2(P, P_0) = 0$ if either $\bar{Q} = \bar{Q}_0$ or $g = g_0$. In addition, we can bound $R_2(P, P_0)$ by use of the Cauchy-Schwartz inequality,

$$\begin{aligned} R_2(P, P_0) &= \sum_{a=0,1} (2a-1) \int \frac{\big(g(a \mid L) - g_0(a \mid L)\big)}{g(a \mid L)} \big(\bar{Q}(a, L) - \bar{Q}_0(a, L)\big)\, dP_0 \\ &\leq \sum_{a=0,1} (2a-1) \frac{1}{g(a \mid L)} \|g - g_0\|_{P_0} \|\bar{Q} - \bar{Q}_0\|_{P_0} \\ &\leq \sum_{a=0,1} (2a-1) \frac{1}{\delta} \|g - g_0\|_{P_0} \|\bar{Q} - \bar{Q}_0\|_{P_0}, \end{aligned}$$

assuming the stronger version of Assumption (A3) (Section 2.2), $\delta < g(1 \mid L) < 1 - \delta$ for some $\delta > 0$. Now, substitute an estimator $\hat{P}_n = (\hat{\bar{Q}}_n, \hat{g}_n)$ for P in $R_2(P, P_0)$. We see that one way to achieve $R_2(\hat{P}_n, P_0) = o_P(n^{-1/2})$ is by ensuring that $\|\hat{g}_n - g_0\|_{P_0} = o_P(n^{-1/4})$ and $\|\hat{\bar{Q}}_n - \bar{Q}_0\|_{P_0} = o_P(n^{-1/4})$. This holds if $\hat{\bar{Q}}_n$ and $\hat{g}_n$ are based on correctly specified parametric models, but, as we will see in the next chapter, it can also hold for estimators based on much larger models. ●

Donsker class conditions. We can analyze the asymptotic behavior of the term (C1) by applying a useful result from empirical process theory (van der Vaart and Wellner, 1996; van der Vaart, 2000). We do not consider this in depth but rather sketch what is needed here. We let $\mathscr{F}$ denote a class of functions $f : \mathcal{O} \to \mathbb{R}$ and consider the *empirical process* $\{\mathbb{G}_n f : f \in \mathscr{F}\}_{n\geq 1}$ for increasing sample size n. The important result that we use again and again is the continuous mapping theorem as formulated in van der Vaart (2000, Lemma 19.24): Suppose $\hat{f}_n \in \mathscr{F}$ where $\mathscr{F}$ is a *Donsker class* $\mathscr{F}$, and suppose $P_0(\hat{f}_n - f_0)^2$ converges to zero in probability. Then the continuous mapping theorem states that:

$$\mathbb{G}_n \hat{f}_n = \mathbb{G}_n f_0 + o_P(1).$$

Applying this result with $\hat{f}_n = \phi^*(\hat{P}_n)$, $f_0 = \phi^*(P_0)$ and $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P_0)$ takes care of the convergence rate of the term (C1), if $\phi^*(\hat{P}_n) \in \mathcal{F}_\phi$ where $\mathcal{F}_\phi$ is a Donsker class and if $P_0(\phi^*(\hat{P}_n) - \phi^*(P_0))^2$ converges to zero in probability.[1] In general, one can show that a function class is Donsker by using bracketing and covering numbers, but often we use that "nice" transformations of Donsker function classes are again Donsker (van der Vaart and Wellner (1996), Section 2.10). In that case we put Donsker class conditions directly on the function class for $\hat{P}_n$ and prove that the mapping ϕ^* preserves the Donsker property. A very important Donsker class we make use of is the class of càdlàg functions with finite variation norm (Gill et al., 1995), but Donsker classes also include, for instance, parametric classes.

3.4 Final comments

In this chapter we have reviewed some important results from semiparametric efficiency theory. We can now talk about asymptotically efficient estimation of a target parameter $\Psi : \mathcal{M} \to \mathbb{R}$ in a given semiparametric statistical model $\mathcal{M}$. Notably, to construct an efficient estimator, one first needs to derive the efficient influence function for the given estimation problem.

The following Theorem 3.3 states conditions for efficient substitution estimation as discussed in Section 3.3. For any estimator $\hat{\Psi}^*_n = \Psi(\hat{P}^*_n)$ that meets the conditions of Theorem 3.3, we have that,

$$\sqrt{n}\big(\hat{\Psi}^*_n - \psi_0\big) \xrightarrow{\mathcal{D}} N(0, \sigma_0^2),$$

The next chapter introduces targeted minimum loss based estimation as a specific methodology for constructing asymptotically linear and efficient estimators.

[1] Note that $P_0(\phi^*(\hat{P}_n) - \phi^*(P_0))^2 = \int (\phi^*(\hat{P}_n)(o) - \phi^*(P_0)(o))^2 \, dP_0(o)$, i.e., the expectation is taken over randomness in $O \mapsto \phi^*(\hat{P}_n)(O)$, with $\hat{P}_n$ considered fixed.

Theorem 3.3 (Asymptotically efficient estimation)**.** *Suppose an estimator $\hat{P}_n^*$ of P_0 is constructed such that:*

(C1) *The estimator solves the efficient influence curve equation:* $\mathbb{P}_n\phi^*(\hat{P}_n^*) = o_P(n^{-1/2})$,
(C2) $\phi^*(\hat{P}_n^*)$ *takes value in a Donsker class and* $P_0(\phi^*(\hat{P}_n^*) - \phi^*(P_0))^2$ *converges to zero in probability,*
(C3) $R_2(\hat{P}_n^*, P_0) = o_P(n^{-1/2})$,

then the estimator $\hat{\Psi}_n = \Psi(\hat{P}_n^*)$ *is an asymptotically linear and efficient estimator for* $\psi_0 = \Psi(P_0)$.

We conclude this chapter by applying Theorem 3.3 specifically to Example 1. We have seen that our target parameter (2.8) can be parametrized by $Q = (\bar{Q}, q_L)$ with the conditional expectation $\bar{Q}(A, L) = \mathbb{E}[Y \mid A, L]$ and the covariate density q_L. In the same manner we note that the efficient influence function for the estimation problem (see Example 1, page 23) is a function of P only through Q and g,

$$\phi^*(Q, g)(O_i) = \left(\frac{\mathbb{1}\{A_i = 1\} - \mathbb{1}\{A_i = 0\}}{g(A_i \mid L_i)}\right)\big(Y - \bar{Q}(A_i, L_i)\big) \tag{3.10}$$

$$+ \big(\bar{Q}(1, L_i) - \bar{Q}(0, L_i)\big) - \Psi(Q). \tag{3.11}$$

Notably, we can provide an estimator for the target parameter if we can estimate $Q = (\bar{Q}, q_L)$, but, as expressed by Theorem 3.3, we need both Q and g to construct an efficient estimator. Now suppose $(\hat{Q}_n, \hat{g}_n)$ of (Q_0, g_0) are constructed such that:

(C1) The efficient influence curve equation is solved: $\mathbb{P}_n\phi^*(\hat{Q}_n, \hat{g}_n) = o_P(n^{-1/2})$,
(C2) $\phi^*(\hat{Q}_n, \hat{g}_n)$ takes value in a Donsker class and $P_0(\phi^*(\hat{Q}_n, \hat{g}_n) - \phi^*(Q_0, g_0))^2$ converges to zero in probability,
(C3) $\|\hat{g}_n - g_0\|_{P_0}\|\hat{\bar{Q}}_n - \bar{Q}_0\|_{P_0} = o_P(n^{-1/2})$,

then, by Theorem 3.3, the estimator $\hat{\Psi}_n = \Psi(\hat{Q}_n)$ is asymptotically linear and efficient.

Chapter 4.
TMLE

Targeted minimum loss based estimation (TMLE) (van der Laan and Rose, 2011, 2018) is a general framework for construction of asymptotically linear estimators for causal parameters. The methodology was first proposed in van der Laan and Rubin (2006) but has since been extended for a large variety of problems (see, e.g., van der Laan and Gruber, 2012; van der Laan and Luedtke, 2014; Chambaz and van der Laan, 2014; Sofrygin and van der Laan, 2017; van der Laan et al., 2018; Cai and van der Laan, 2019).

Given observations $O_1, \ldots, O_n \overset{iid}{\sim} P_0$ with P_0 belonging to a semiparametric model $\mathcal{M}$, we let $\Psi : \mathcal{M} \to \mathbb{R}$ be our target parameter with efficient influence function $\phi^*(P)$. TMLE is based on substitution estimation: Estimation of the target parameter requires estimators for the infinite-dimensional nuisance parameters, as in our running example where we need estimators for $Q = (\bar{Q}, q_L)$ and for g. The idea is that the estimators are constructed exactly such as to meet the conditions of Theorem 3.3 from the previous chapter.

The general TMLE template can be summarized as follows:

Step 1: Initial estimation. Construct an initial estimator $\hat{P}_n^0$ for P_0.

Step 2: Targeting step. Update $\hat{P}_n^0 \mapsto \hat{P}_n^*$ such that,

$$\mathbb{P}_n \, \phi^*(\hat{P}_n^*) = o_P(n^{-1/2}).$$

We mainly focus on the second step, the targeting step. We start by giving a general description in Section 4.1, then we apply it specifically to our running example for illustration. In Section 4.2 we give a brief outline of the first step; for sake of presentation we focus only on the running example. We finish the chapter with the (discrete) time-varying setting from Section 4.3, which is the setting that we generalize to continuous time in Manuscript I.

4.1 The targeting step

The targeting step for the parameter $\Psi : \mathcal{M} \to \mathbb{R}$ with influence function $\phi^*(P)$, given an initial estimator $\hat{P}_n^0$, is defined by the following:

(i) A choice of a loss function for P, $(O, P) \mapsto \mathcal{L}(P)(O)$.

(ii) A parametric fluctuation model parametrized by $\varepsilon \in \mathbb{R}$ through P at $\varepsilon = 0$, $\{P_\varepsilon : \varepsilon \in \mathbb{R}\} \subset \mathcal{M}$.

(iii) An updating algorithm.

Notably, (i) and (ii) must be defined such that the score of P_ε equals the efficient influence function for the target parameter evaluated in P itself,

$$\left.\frac{d}{d\varepsilon}\right|_{\varepsilon=0} \mathcal{L}(P_\varepsilon)(O_i) = \phi^*(P)(O). \tag{4.1}$$

Recall that the end goal of the targeting step is to fluctuate, or update, the initial estimator $\hat{P}_n^0$ into $\hat{P}_n^*$ such that $\hat{P}_n^*$ solves the efficient influence curve equation. To carry out the fluctuation, we consider the fluctuation model defined by (ii) through the initial estimator $\hat{P}_n^0$: Denote this by $\hat{P}_{n,\varepsilon}^0$. The parameter ε determines the amount of fluctuation. We now estimate ε in the model $\hat{P}_{n,\varepsilon}^0$ based on the observed data by,

$$\hat{\varepsilon}_n := \operatorname*{argmin}_\varepsilon \mathbb{P}_n \mathcal{L}(\hat{P}_{n,\varepsilon}^0).$$

Then we update $\hat{P}_n^0$ into $\hat{P}_n^1$ by defining $\hat{P}_n^1 := \hat{P}_{n,\hat{\varepsilon}_n}^0$. Notably, the minimizer $\hat{\varepsilon}_n$ of $\varepsilon \mapsto \mathbb{P}_n \mathcal{L}(\hat{P}_{n,\varepsilon}^0)$ solves,

$$\mathbb{P}_n \left.\frac{d}{d\varepsilon}\right|_{\varepsilon=\hat{\varepsilon}_n} \mathcal{L}(\hat{P}_{n,\varepsilon}^0) = 0.$$

Depending on the problem at hand we may already be done at this point; if we repeat the same procedure with $\hat{P}_n^1$ substituted for $\hat{P}_n^0$, and we have that,

$$\mathbb{P}_n \left.\frac{d}{d\varepsilon}\right|_{\varepsilon=0} \mathcal{L}(\hat{P}_{n,\varepsilon}^1) = 0, \tag{4.2}$$

then $\hat{P}_n^* := \hat{P}_n^1$ by the property ensured in (4.1) solves the efficient influence curve equation without further updating. We are done and we define the TMLE estimator for the target parameter as $\hat{\Psi}_n^* = \Psi(\hat{P}_n^*)$.

For many problems, however, we need to iterate the fluctuation step before we get to (4.2). This is for instance the case if the loss is indexed by other components of the model that are also updated. We will see an example of this later in Chapter 6. Starting from a current estimator $\hat{P}_n^k$ and substituting $\hat{P}_n^k$ for $\hat{P}_n^0$, the kth update is carried out by solving $\hat{\varepsilon}_k := \operatorname{argmin}_\varepsilon \hat{P}_{n,\varepsilon}^k$ and defining $\hat{P}_n^{k+1} := \hat{P}_{n,\hat{\varepsilon}_k}^k$. These steps are repeated until $\hat{\varepsilon}_{k^*}$ is approximately zero. Then $\hat{P}_n^* := \hat{P}_n^{k^*}$ solves the efficient influence curve equation, $\mathbb{P}_n \phi^*(\hat{P}_n^{k^*}) = o_P(n^{-1/2})$, and we define our TMLE estimator for the

target parameter as $\hat{\Psi}_n^* = \Psi(\hat{P}_n^*)$.

Consider specifically the target parameter that only depends on P through $Q = Q(P)$ but the efficient influence function depends on both $Q = Q(P)$ and $G = G(P)$. We then need initial estimators for both Q and G, but a targeting step only for Q; constructed such that the targeted estimator $\hat{Q}_n^*$ for a given estimator $\hat{G}_n$ solves the efficient influence curve equation,

$$\mathbb{P}_n \phi^*(\hat{Q}_n^*, \hat{G}_n) = o_P(n^{-1/2}).$$

This involves a choice of a loss function for Q, potentially indexed by G, $(O, Q) \mapsto \mathcal{L}(Q)(O)$, and a parametric fluctuation model through Q. In the following section we consider our running example for illustration.

4.1.1 The targeting step for Example 1

In our running example, suppose we have constructed initial estimators $\hat{g}_n, \hat{\bar{Q}}_n^0$ for $g, \bar{Q}$, respectively. We estimate q_L by the empirical distribution $\hat{q}_{L,n}$ of $L_1, \ldots, L_n$. Let $\hat{Q}_n^0 = (\hat{\bar{Q}}_n^0, \hat{q}_{L,n})$. In the targeting step we update $\hat{\bar{Q}}_n^0 \mapsto \hat{\bar{Q}}_n^*$ such that $\hat{Q}_n^* = (\hat{\bar{Q}}_n^*, \hat{q}_{L,n})$ solves,

$$\mathbb{P}_n \, \phi^*(\hat{Q}_n^*, \hat{g}_n) = o_P(n^{-1/2}).$$

This involves a choice of a loss function for $\bar{Q}$ and a corresponding path indexed by $\varepsilon \in \mathbb{R}$ through $\bar{Q}$, such that the generated score equals the first term (3.10) of the canonical gradient. In other words, we define a parametric fluctuation model $\{\bar{Q}_\varepsilon : \varepsilon \in \mathbb{R}\} \subset \mathcal{M}$ through $\bar{Q}$ at $\varepsilon = 0$ and a loss function $(O, \bar{Q}) \to \mathcal{L}(\bar{Q})(O)$, such that,

$$\phi^*(Q, G)(O_i) = \underbrace{\left(\frac{\mathbb{1}\{A_i = 1\} - \mathbb{1}\{A_i = 0\}}{g(A_i \mid L_i)} \right) \big(Y_i - \bar{Q}(A_i, L_i)\big)}_{= \frac{d}{d\varepsilon}|_{\varepsilon=0} \mathcal{L}(\bar{Q}_\varepsilon)(O_i)} + \bar{Q}(1, L_i) - \bar{Q}(0, L_i) \\ - \Psi(\bar{Q}, q_L). \tag{4.3}$$

To achieve this, we use the log-likelihood loss,

$$\mathcal{L}(\bar{Q})(O) = -\big(Y \log \bar{Q}(A, L) + (1 - Y) \log(1 - \bar{Q}(A, L))\big), \tag{4.4}$$

along with a logistic regression model,

$$\bar{Q}_\varepsilon(A, L) = \operatorname{expit}\big(\operatorname{logit}(\bar{Q}(A, L)) + \varepsilon H(A, L)\big), \tag{4.5}$$

with the so-called "clever covariate",

$$H(A, L) := \frac{\mathbb{1}\{A = 1\} - \mathbb{1}\{A = 0\}}{g(A \mid L)}, \tag{4.6}$$

that depends on g but not on $\bar{Q}$. Substituting $\hat{g}_n$ for g in (4.6), we now define,

$$\hat{\varepsilon}_n := \underset{\varepsilon}{\operatorname{argmin}}\ \mathbb{P}_n\, \mathcal{L}(\hat{\bar{Q}}^0_{n,\varepsilon}),$$

and the updated estimator $\hat{\bar{Q}}^*_n(a, L) := \hat{\bar{Q}}^0_{n,\hat{\varepsilon}_n}(a, L)$. Per construction of our loss function (4.4) together with the fluctuation model (4.5) and the clever covariate (4.6), we now have that,[1]

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\mathbb{1}\{A_i = 1\} - \mathbb{1}\{A_i = 0\}}{\hat{g}_n(A_i \mid L_i)}\right)\left(Y_i - \hat{\bar{Q}}^*_n(A_i, L_i)\right) = 0, \tag{4.7}$$

Finally we estimate q_L by the empirical distribution $\hat{q}_{L,n}$ of $L_1, \ldots, L_n$. This defines the TMLE estimator as,

$$\hat{\Psi}^*_n = \Psi(\hat{\bar{Q}}^*_n, \hat{q}_{L,n}) = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\bar{Q}}^*_n(1, L_i) - \hat{\bar{Q}}^*_n(0, L_i)\right),$$

and we solve,

$$\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\bar{Q}}^*_n(1, L_i) - \hat{\bar{Q}}^*_n(0, L_i) - \Psi(\hat{\bar{Q}}^*_n)\right) = 0, \tag{4.8}$$

which is the leftover term (3.11) of the efficient influence function. Combining (4.7) and (4.8) now specifically implies that,

$$\mathbb{P}_n\, \phi^*(\hat{Q}^*_n, \hat{g}_n) = 0,$$

exactly as we wanted, and we have completed the targeting step.

4.2 Initial estimation for Example 1

TMLE relies on initial estimators for the relevant parts of the data-generating distribution. To illustrate the construction of such estimators we continue with our running example. In this example, we need initial estimators $\hat{\bar{Q}}^0_n, \hat{g}_n$ for $\bar{Q}(A, L) = \mathbb{E}[Y \mid A, L]$ and $g(a \mid L) = P(A = a \mid L)$. (Only $\hat{\bar{Q}}^0_n$ is indexed by the superscript '0', since only this, and not $\hat{g}_n$, is updated in the targeting step). Estimation of $\bar{Q}(A, L)$ and $g(a \mid L)$ can be formulated as a prediction problem. One could for example specify a logistic regression of Y on A, L such as,

$$\bar{Q}(A, L) = \mathbb{E}[Y \mid A, L] = \operatorname{expit}\left(\beta_0 + \beta_1 A + \beta_2^\top L\right), \tag{4.9}$$

[1] The score of a coefficient of a covariate in a parametric logistic regression is equal to the covariate times the residual.

and a logistic regression of A on L such as,

$$g(1 \mid L) = \mathbb{E}[A \mid L] = \operatorname{expit}\big(\gamma_0 + \gamma_1^\top L\big). \tag{4.10}$$

Assuming that the models in (4.9) and (4.10) are correctly specified, the assumptions of Theorem 3.3 are met, and we are done: The resulting estimator after performing the targeting step is asymptotically linear and efficient. However, often we would like to avoid such assumptions and, as we discussed in Chapter 3, work with larger models than parametric ones.

4.2.1 Loss-function based cross-validation

We write general estimators as $\hat{\bar{Q}}_k$ and $\hat{g}_k$, where $\mathbb{P}_n \mapsto \hat{\bar{Q}}_k(\mathbb{P}_n)$ and $\mathbb{P}_n \mapsto \hat{g}_k(\mathbb{P}_n)$ map the empirical distribution $\mathbb{P}_n$ of the observed data to an estimator for $\bar{Q}$ and g, respectively. Our goal is to construct an estimator for $\bar{Q}$ (and g equivalently) that is 'close' to the truth. Recall that we defined a loss function $(O, \bar{Q}) \to \mathcal{L}(\bar{Q})(O)$ in (4.4). We refer to $P\mathcal{L}(\bar{Q})$ as the corresponding risk (expected loss) under P. The true $\bar{Q}_0$ is identified as the minimizer of the true risk, $P_0\mathcal{L}(\bar{Q}_0) = \mathbb{E}_{P_0}[\mathcal{L}(\bar{Q}_0)(O)]$. To assess the performance of a given estimator we can use the risk difference,

$$d_0(\bar{Q}, \bar{Q}_0) = P_0\mathcal{L}(\bar{Q}) - P_0\mathcal{L}(\bar{Q}_0),$$

as a measure of 'closeness'. We consider now a collection, or a library, of estimators $\hat{\bar{Q}}_1, \ldots, \hat{\bar{Q}}_K$. (Notice that we are here using indexation $k = 1, \ldots, K$, which is not to be confused with k used in the previous sections to index the targeting iterations). The estimator that is closest to the true $\bar{Q}_0$ in the library is the one that minimizes the risk difference $d_0(\bar{Q}, \bar{Q}_0)$, corresponding to minimizing the true risk $P_0\mathcal{L}(\bar{Q})$.

To estimate the true risk, we work with the empirical risk obtained by V-fold cross validation. A V-fold cross-validation scheme defines $v = 1, \ldots, V$ sample splits into a training sample $\{1, \ldots, n\} \setminus \mathrm{Val}(v)$ used to construct the estimator and a validation sample $\mathrm{Val}(v) \subset \{1, \ldots, n\}$ used to evaluate it. We note that $(\mathrm{Val}(v))_{1 \le v \le V}$, forms a partitioning of the total sample $\{1, \ldots, n\}$. Let $\mathbb{P}^0_{n,v}, \mathbb{P}^1_{n,v}$ denote the empirical distributions corresponding to the training and validation sample, respectively, for the vth sample split, $v = 1, \ldots, V$. We can now define the cross-validation selector of k as,

$$\hat{k}_n^Q := \operatorname*{argmin}_k \frac{1}{V} \sum_{v=1}^{V} \mathbb{P}^1_{n,v}\, \mathcal{L}(\hat{\bar{Q}}_k(\mathbb{P}^0_{n,v})),$$

that adaptively selects an estimator among the given library of K estimators $\hat{\bar{Q}}_1, \ldots, \hat{\bar{Q}}_K$. Nice statistical properties have been established for this cross-validation selector:

Under the following conditions on the loss function and the model $\mathcal{M}$,

$$\sup_{\bar{Q}} \frac{\|\mathcal{L}(\bar{Q}) - \mathcal{L}(\bar{Q}_0)\|_{P_0}^2}{P_0(\mathcal{L}(\bar{Q}) - \mathcal{L}(\bar{Q}_0))} \leq M_1 < \infty, \tag{L1}$$

$$\sup_{O,\bar{Q}} |\mathcal{L}(\bar{Q}) - \mathcal{L}(\bar{Q}_0)| < M_2 < \infty, \tag{L2}$$

asymptotic optimality of the cross-validation selector is implied by the oracle inequality (van der Laan and Dudoit, 2003; van der Vaart et al., 2006; van der Laan et al., 2006; van der Laan et al., 2007; Polley et al., 2011) that compares the performance of $\hat{\bar{Q}}_{\hat{k}_n^Q}(\mathbb{P}_n)$ with the theoretical "oracle" estimator $\hat{\bar{Q}}_{\tilde{k}_n^Q}(\mathbb{P}_n)$ which is based on the "oracle" selector (see, e.g., van der Laan et al., 2007, Theorem 1),

$$\tilde{k}_n^Q := \operatorname*{argmin}_k \frac{1}{V} \sum_{v=1}^{V} P_0 \, \mathcal{L}(\hat{\bar{Q}}_k(\mathbb{P}_{n,v}^0)),$$

that minimizes the true risk. Specifically, the estimator $\hat{\bar{Q}}_{\hat{k}_n^Q}(\mathbb{P}_n)$ performs asymptotically as well as the best performing candidate estimator included in the library. These results have further been extended to the case where libraries also include, for example, convex combinations of their individual estimators (van der Laan et al., 2007); let $\hat{\bar{Q}}_k^V(O_i)$ denote the cross-validated prediction for the ith observation, i.e., $\hat{\bar{Q}}_k^V(O_i) = \sum_{v=1}^{V} \mathbb{1}\{i \in \text{Val}(v)\} \, \hat{\bar{Q}}_k(\mathbb{P}_{n,v}^0)(O_i)$ where $\mathbb{1}\{i \in \text{Val}(v)\}$ indicates whether the ith observation is included in the vth validation set. Define next $\hat{\bar{Q}}_\alpha^V = \alpha_1 \hat{\bar{Q}}_1^V + \cdots + \alpha_K \hat{\bar{Q}}_K^V$ with $\alpha = (\alpha_1, \ldots, \alpha_K)$ and $\sum_{k=1}^{K} \alpha_k = 1$, $\alpha_k \geq 0$, $k = 1, \ldots, K$. Then estimate α by,

$$\hat{\alpha}_n^Q = (\hat{\alpha}_1^Q, \ldots, \hat{\alpha}_K^Q) := \operatorname*{argmin}_\alpha \mathbb{P}_n \, \mathcal{L}(\hat{\bar{Q}}_\alpha^V),$$

which provides a corresponding weighted estimator $\hat{\bar{Q}}_{\hat{\alpha}_n^Q} = \hat{\alpha}_1^Q \hat{\bar{Q}}_1(\mathbb{P}_n) + \cdots + \hat{\alpha}_K^Q \hat{\bar{Q}}_K(\mathbb{P}_n)$ for $\bar{Q}$. This process is also referred to as *super learning* (van der Laan et al., 2007), and the estimator $\hat{\bar{Q}}_{\hat{\alpha}_n^Q}$ is called the *super learner* for $\bar{Q}$. To realize the flexibility of super learning, note that a library for example could consist of estimators based on many different parametric models: One could replace the linear forms of (4.9)–(4.10) with ones including any function of L or different combinations of interaction terms. The library could also include any machine learning algorithm for, in this case, binary outcome regression.

Super learners can be set up for any parameter that we can define as a risk minimizer. Just as for $\bar{Q}$, we can define $(O, g) \to \mathcal{L}_g(g)(O)$,

$$\mathcal{L}_g(g)(O) = -\big(A \log g(A \mid L) + (1 - A) \log(1 - g(A \mid L))\big),$$

and, as we outlined above for $\bar{Q}$, construct a super learner $\hat{g}_{\hat{\alpha}_n^g}$ for g. In our running example, the library of types of estimators can be the same as the one used for $\bar{Q}$. The oracle results tell us that the super learners for $\bar{Q}, g$ will perform asymptotically as well as the best performing combination of estimators included in their libraries. This will be important: Indeed, it can be shown that we can construct a particular estimator for $\bar{Q}$ and for g such that $\|\hat{g}_n - g_0\|_{P_0} = o_P(n^{-1/4})$ and $\|\hat{\bar{Q}}_n - \bar{Q}_0\|_{P_0} = o_P(n^{-1/4})$ under minimal conditions on $\mathcal{M}$, so that when this estimator is included in the libraries, condition (C3) of Theorem 3.3 is met. This estimator is called the highly adaptive lasso (HAL) estimator (van der Laan, 2015; Benkeser and van der Laan, 2016; van der Laan, 2017; van der Laan and Rose, 2018). In the next section we give a short presentation of the HAL estimator and some intuition on the theoretical properties.

4.2.2 HAL estimation and proof of convergence

The key to the following is a particular nonparametric smoothness assumption that we impose on the parameter spaces of the nuisance parameters. Importantly, the elements of $\mathcal{M}$ can be characterized by functions that can be discontinuous or non-differentiable, we only need them to be càdlàg and have finite sectional variation norm such that they each generate a signed measure (Gill et al., 1995). This implies a specific representation of the nuisance parameters, that we use to define the HAL estimator. Moreover, the class of functions that are càdlàg and have finite sectional variation norm is a Donsker class, and this will allow us to establish conditions (C2) and (C3) of Theorem 3.3 which, as we have seen, for our running example becomes:

(C2) $\phi^*(\hat{Q}_n, \hat{g}_n)$ takes value in a Donsker class and $P_0(\phi^*(\hat{Q}_n, \hat{g}_n) - \phi^*(Q_0, g_0))^2$ converges to zero in probability,

(C3) $\|\hat{g}_n - g_0\|_{P_0}\|\hat{\bar{Q}}_n - \bar{Q}_0\|_{P_0} = o_P(n^{-1/2})$.

We assume that for each $g \in \mathcal{G}$, $O \mapsto g(O)$ is càdlàg with finite sectional variation norm and $\delta < g(O) < 1 - \delta$ a.s. Likewise, for each $\bar{Q} \in \mathcal{Q}$, $O \mapsto \bar{Q}(O)$ is càdlàg and has finite sectional variation norm. The sectional variation norm for any d-variate real-valued càdlàg function f admits a representation in terms of its measures over sections (Gill et al., 1995). In particular, when the function is defined on a discrete support this representation reduces to a finite sum of interval indicator functions and the measure assigned to those intervals. Moreover, the variation norm becomes the sum of absolute values of coefficients. This implies that the estimation of the function corresponds to an L_1-penalized (Lasso) regression (Tibshirani, 1996) where the unknown β coefficients are the measures assigned to the intervals. Indeed, we can write a d-variate real-valued function f with $m \in \mathbb{N}$ support points z_j as,

$$f(x) = f(0) + \sum_{s \in \mathcal{P}(\{1,\ldots,d\})} \sum_j \mathbb{1}\{z_{j,s} \leq x_s\} f(dz_{j,s}), \tag{4.11}$$

where $\mathcal{P}(\{1,\ldots,d\})$ is the set of all subsets of $\{1,\ldots,d\}$, $z_{j,s}$ and x_s are defined by setting the coordinates not in the index set s to zero, and $f(dz_{j,s})$ is the point mass assigned to $z_{j,s}$. If we define $\beta_{j,s} := f(dz_{j,s})$, we have that $\|f\|_v = \|\beta\|_1$ where $\|\cdot\|_1$ denotes the L_1-norm and $\|\cdot\|_v$ denotes the variation norm (Gill et al., 1995),

$$\|f\|_v = |f(0)| + \sum_{s\in\mathcal{P}(\{1,\ldots,d\})} \sum_j |f(dz_{j,s})|.$$

For example, assume that $L \in \mathbb{R}_+$, and that we want to estimate $g(L) = P(A = 1 \,|\, L)$. We assume that g is càdlàg in L and has finite sectional variation norm. Let $(\ell_j)_{j=1}^J$ be the observed support points of L (if L is truly continuous, then $J = n$). Then the representation (4.11) for g_h (that jumps only at support points of L) as an approximation to g becomes,

$$g_h(L) = g_h(0) + \sum_{j=1}^{J} \mathbb{1}\{\ell_j \le L\}\, dg_h(L_j). \tag{4.12}$$

Define $\psi_j(L) = \mathbb{1}\{\ell_j \le L\}$ for $j = 1,\ldots,J$ with corresponding coefficients $\beta_j = dg_h(L_j)$, and the HAL estimator for a given choice of $M < \infty$,

$$\hat{g}_{n,M}^{\text{HAL}} = \underset{\beta=(\beta_0,\ldots,\beta_J)}{\operatorname{argmin}} \; \mathbb{P}_n\, \mathcal{L}_g(g) \qquad \text{s.t.} \quad \|\beta\|_1 \le M. \tag{4.13}$$

This corresponds to an L_1-penalized regression with a coefficient vector of size equal to the number of observed support points for the variable L. We may select M, corresponding to the bound that we put on the variation norm, by cross-validation. Now suppose that L is two-dimensional, i.e., $L = (L_1, L_2) \in \mathbb{R}_+ \times \mathbb{R}_+$. Let $(\ell_{j,1})_{j=1}^{J_1}$ be the observed support points of L_1 and $(\ell_{j,2})_{j=1}^{J_2}$ the observed support points of L_2. Then the representation of g_h in (4.12) as an approximation to g becomes,

$$\begin{aligned} g_h(L) = g_h(L_1, L_2) = g(0,0) &+ \sum_{j=1}^{J_1} \mathbb{1}\{\ell_{j,1} \le L_1\}\, dg_h(\ell_{j,1}, 0) \\ &+ \sum_{j=1}^{J_2} \mathbb{1}\{\ell_{j,2} \le L_2\}\, dg_h(0, \ell_{j,2}) \\ &+ \sum_{j_1=1}^{J_1} \sum_{j_2=1}^{J_2} \mathbb{1}\{\ell_{j_1,1} \le L_1\}\mathbb{1}\{\ell_{j_1,2} \le L_2\}\, dg_h(\ell_{j_1,1}, \ell_{j_2,2}). \end{aligned}$$

Define $\psi_{j,1}(L_1) = \mathbb{1}\{\ell_{j,1} \le L_1\}$ for $j = 1,\ldots,J_1$, $\psi_{j,2}(L_2) = \mathbb{1}\{\ell_{j,2} \le L_2\}$ for $j = 1,\ldots,J_2$, and $\psi_{j,12}(L_1, L_2) = \mathbb{1}\{\ell_{j(1),1} \le L_1, \ell_{j(2),2} \le L_2\}$ with $j = 1,\ldots,J_1 J_2$ indexing all combinations of (j_1, j_2). Moreover, define $\beta_{j,1} = dg_h(\ell_{j,1}, 0)$, $\beta_{j,2} = dg_h(0, \ell_{j,2})$

and $\beta_{j,12} = dg_h(\ell_{j,1}, \ell_{j,2})$. Again we can define the HAL estimator as in (4.13), but now the parameter β surely has a much larger dimension.

It is clear that the HAL representation (4.11) and the HAL estimation problem increase rapidly in complexity when the dimension of the support of L increases. Despite these potential practical problems (that may be solved by various discretizations or dimension reduction steps), the theory behind the HAL estimator is quite powerful: The main point is that for HAL estimation,

$$\|\hat{g}_n^{\text{HAL}} - g_0\|_{P_0} = o_P(n^{-1/4}) \quad \text{and} \quad \|\hat{\bar{Q}}_n^{\text{HAL}} - \bar{Q}_0\|_{P_0} = o_P(n^{-1/4}),$$

can be established (see van der Laan, 2017; Benkeser and van der Laan, 2016; van der Laan and Rose, 2018, Chapter 7), relying only on the assumptions on $\mathcal{M}$ stated in the beginning of this subsection. By the oracle properties of the super learner, a TMLE that uses a super learner which includes HAL estimation in its library for initial estimation thus meets the conditions of Theorem 3.3 requiring only that $(\bar{Q}_0, g_0)$ are càdlàg and have finite variation norm.

4.3 Longitudinal TMLE (LTMLE)

Here, we describe the TMLE template for the longitudinal data setting from Chapter 2 with observed data,

$$O = (L_0, A_1, L_1, A_1, \ldots, L_K, A_K, L_{K+1} = Y).$$

We are interested in estimating $\psi_0 = \Psi(P_0)$, where the target parameter $\Psi : \mathcal{M} \to \mathbb{R}$ is defined by,

$$\Psi(P) = \mathbb{E}_{P_{q,g^*}}[Y] = \int_{\mathcal{O}} Y \, dP_{q,g^*}.$$

Again, g^* represents the intervention of interest and P_{q,g^*} has density as defined by the g-computation formula (2.6). We also refer to van der Laan and Gruber (2012); Schnitzer et al. (2014); Petersen et al. (2014); Sofrygin et al. (2019) and van der Laan and Rose (2018, Chapter 3 and 4). We focus on the implementation of TMLE for longitudinal data that constructs an estimator by exploiting a sequential regression

representation of the target parameter (Robins, 2000; Bang and Robins, 2005):

$$\begin{aligned}
\bar{Q}_{K+1}(\bar{L}_K, \bar{A}_K) &= \mathbb{E}_{P_{q,g^*}}[Y \mid \bar{L}_K, \bar{A}_K] = Y, \\
\bar{Q}_K(\bar{L}_{K-1}, \bar{A}_{K-1}) &= \mathbb{E}_{P_{q,g^*}}[\bar{Q}_{K+1}(\bar{L}_K, \bar{A}_K) \mid \bar{L}_{K-1}, \bar{A}_{K-1}], \\
&\vdots \\
\bar{Q}_k(\bar{L}_{k-1}, \bar{A}_{k-1}) &= \mathbb{E}_{P_{q,g^*}}[\bar{Q}_{k+1}(\bar{L}_k, \bar{A}_k) \mid \bar{L}_{k-1}, \bar{A}_{k-1}], \\
&\vdots \\
\bar{Q}_2(\bar{L}_1, A_1) &= \mathbb{E}_{P_{q,g^*}}[\bar{Q}_3(\bar{L}_2, \bar{A}_2) \mid \bar{L}_1, \bar{A}_1], \\
\bar{Q}_1(L_0) &= \mathbb{E}_{P_{q,g^*}}[\bar{Q}_2(\bar{L}_1) \mid L_0] = \mathbb{E}[\bar{Q}_2(\bar{L}_1) \mid L_0].
\end{aligned} \tag{4.14}$$

The representation allows for evaluation of $\mathbb{E}_{P_{q,g^*}}[Y]$ by an integration process, where, at each time-point k, we first integrate out A_k with respect to the intervention g_k^* and then L_k with respect to q_k.

The canonical gradient (the efficient influence curve) of the pathwise derivative of $\Psi : \mathcal{M} \to \mathbb{R}$ at $P \in \mathcal{M}$ is given by (see, e.g., van der Laan and Gruber, 2012, Theorem 1),

$$\phi^*(P)(O) = \bar{Q}_1(L_0) - \Psi(P) \tag{4.15}$$

$$+ \sum_{k=1}^{K+1} \left(\prod_{\ell=0}^{k-1} \frac{g_\ell^*(\bar{A}_\ell \mid \bar{L}_\ell)}{g_\ell(\bar{A}_\ell \mid \bar{L}_\ell)} \right) \left(\bar{Q}_{k+1}(\bar{L}_k, \bar{A}_k) - \bar{Q}_k(\bar{L}_{k-1}, \bar{A}_{k-1}) \right). \tag{4.16}$$

The longitudinal TMLE (LTMLE) procedure is defined iteratively, along the lines of the representation in (4.14). Largely, it can be summarized as follows. The algorithm starts with an initial estimator of the conditional expectation at the final time-point $K+1$, and updates this estimator to solve the efficient influence function at time-point $K+1$. Then the updated conditional expectation serves as the outcome in the next (Kth) conditional expectation and is updated to solve the efficient influence function at time-point K. Proceeding iteratively in this way along the time sequence, a targeting step is performed at each time k to solve the efficient influence curve equation at time k. In the end we have an estimator for $\bar{Q}_1$ which is a function only of L_0, but was obtained from a sequence of estimators $(\hat{\bar{Q}}_k^* : k = 1, \ldots, K)$ constructed such as to solve the entire efficient influence curve equation, and we estimate the target parameter by taking the empirical average over $\hat{\bar{Q}}_1^*$.

To explore the details, consider here $K = 2$ and let the intervention be the static intervention $g_k^* = \mathbb{1}\{A_k = 0\}$, $k = 0, 1$. Our observed data are then: $O = (L_0, A_0, L_1, A_1, Y)$. For the implementation of the LTMLE, we need initial estimators $\hat{g}_k$ for g_k, $k = 0, 1$, and initial and updated estimators for $\bar{Q}_k$, $k = 1, 2$. For g_1 we can apply loss-based cross-validation as described in Section 4.2.1, now with outcome A_1 and conditioning set (A_0, L_1, L_1), and, similarly for g_0, with outcome A_0 and conditioning set L_0.

Estimation and targeting of the conditional expectations $\bar{Q}_1, \bar{Q}_2$ proceed as follows. First, for $\bar{Q}_2$, we define the log-likelihood loss function,

$$\mathcal{L}_2(\bar{Q}_2)(O) = -\big(Y \log \bar{Q}_2 + (1-Y)\log(1-\bar{Q}_2)\big).$$

This corresponds to a regression of Y on $(\bar{A}_1, \bar{L}_1)$, and defines an estimator $\hat{\bar{Q}}_2^0$ as a function of $(\bar{A}_1, \bar{L}_1)$. To fluctuate and target $\hat{\bar{Q}}_2^0$, we define the submodel $\{\bar{Q}_{2,\varepsilon} : \varepsilon \in \mathbb{R}\}$ as,

$$\bar{Q}_{2,\varepsilon} = \operatorname{expit}\big(\operatorname{logit} \bar{Q}_2 + \varepsilon H_2\big),$$

with the "clever covariate",

$$H_2(\bar{A}_1, \bar{L}_1) = \prod_{\ell=0,1} \frac{\mathbb{1}\{A_\ell = 0\}}{g_\ell(\bar{A}_\ell \mid \bar{L}_\ell)}. \tag{4.17}$$

We provide an estimator $\hat{H}_2$ by substituting $\hat{g}_\ell$, $\ell = 0, 1$, in (4.17). We define,

$$\hat{\varepsilon}_2 := \operatorname*{argmin}_{\varepsilon} \mathbb{P}_n \mathcal{L}_2(\hat{\bar{Q}}_{2,\varepsilon}^0),$$

and the updated estimator $\hat{\bar{Q}}_2^* := \hat{\bar{Q}}_{2,\hat{\varepsilon}_2}^0$. For this updated estimator it now holds that,

$$\frac{1}{n}\sum_{i=1}^{n} \left(\prod_{\ell=0,1} \frac{\mathbb{1}\{A_\ell = 0\}}{g_\ell(\bar{A}_\ell \mid \bar{L}_\ell)} \right) \Big(Y_i - \hat{\bar{Q}}_2^*(\bar{L}_{1,i}, \bar{A}_{1,i}) \Big) = 0, \tag{4.18}$$

and we say this it is targeted. This was the first step. In the second step we use the targeted estimator $\hat{\bar{Q}}_2^*$ to provide an estimator for $\bar{Q}_1$. We define the log-likelihood loss function, *indexed by* the estimator $\hat{\bar{Q}}_2^*$,

$$\mathcal{L}_1(\bar{Q}_1)(O) = -\big(\hat{\bar{Q}}_2^* \log \bar{Q}_1 + (1 - \hat{\bar{Q}}_2^*)\log(1-\bar{Q}_1)\big).$$

This corresponds to a regression of $\hat{\bar{Q}}_2^*$ on (A_0, L_0), and defines an estimator $\hat{\bar{Q}}_1^0$ as a function of (A_0, L_0). To target $\hat{\bar{Q}}_1^0$, we define the submodel $\{\hat{\bar{Q}}_{1,\varepsilon}^0 : \varepsilon \in \mathbb{R}\}$ as,

$$\hat{\bar{Q}}_{1,\varepsilon}^0 = \operatorname{expit}\big(\operatorname{logit} \bar{Q}_1 + \varepsilon H_1\big),$$

with,

$$H_1(A_0, L_0) = \frac{\mathbb{1}\{A_0 = 0\}}{g_0(A_0 \mid L_0)}. \tag{4.19}$$

We provide an estimator $\hat{H}_1$ by substituting $\hat{g}_0$ in (4.19). We define,

$$\hat{\varepsilon}_1 := \operatorname*{argmin}_{\varepsilon} \mathbb{P}_n \mathcal{L}_1(\hat{\bar{Q}}_{1,\varepsilon}^0),$$

and the updated estimator $\hat{\bar{Q}}_1^* := \hat{\bar{Q}}^0_{1,\hat{\varepsilon}_1}$. For this targeted estimator it now holds that,

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{1}\{A_0=0\}}{g_0(A_0\,|\,L_0)}\left(\hat{\bar{Q}}_2^*(\bar{L}_{1,i},\bar{A}_{1,i})-\hat{\bar{Q}}_1^*(L_{0,i},A_{0,i})\right)=0. \tag{4.20}$$

Now, note that (4.18)–(4.20) together sum up to the second part (4.16) of the efficient influence function $\phi^*(P)$. As in Section 4.1.1, we estimate the distribution of baseline covariates by the empirical distribution which takes care of the first term (4.15) of the efficient influence function $\phi^*(P)$. In the end, we define the targeted estimator $\hat{\Psi}_n^*$ as,

$$\hat{\Psi}_n^*=\frac{1}{n}\sum_{i=1}^{n}\hat{\bar{Q}}_1^*(L_{0,i}),$$

which solves the efficient influence curve equation,

$$\mathbb{P}_n\,\phi^*(\hat{Q}_n^*,\hat{Q}_n)=0,$$

with $\hat{Q}_n^*=(\hat{\bar{Q}}_1^*,\hat{\bar{Q}}_2^*)$ and $\hat{G}_n=(\hat{g}_0,\hat{g}_1)$.

4.4 Final comments

In Chapter 3 we presented Theorem 3.3 which characterizes conditions for asymptotically efficient estimation of a target parameter $\psi_0=\Psi(P_0)$. In this chapter we have reviewed TMLE as a two-step procedure for construction of an estimator that meets the conditions of this theorem:

1. By the targeting step, the TMLE estimator solves the efficient influence curve equation. This already implies nice properties that relates back to double robustness as we discussed in Chapter 3: An estimator that solves the efficient influence curve equation (for which the specific efficient influence function admits a double robustness structure) is consistent if either the estimator for Q or the estimator for G is consistent. That the TMLE estimator solves the efficient influence curve equation further provides the basis for establishing asymptotic normality and efficiency.

2. The targeting step can be supplemented by a data-adaptive super learning approach for initial estimation. Particularly, the super learner performs asymptotically as well as any algorithm used in its library. Combined with double robustness, including any estimator that attains a rate as fast as $n^{-1/4}$ in the library of the super learner is enough to take care of the remainder term in Theorem 3.3.

We note that TMLE is not the only method that exploits the efficient influence function for efficient estimation of the target parameter. Estimating equation methodology (van der Laan and Robins, 2003) provides a complementary approach that similarly exploits double robustness properties and asymptotic efficiency by identifying parameters as solutions to the right estimating equations. One simple example of an estimating equation is the inverse probability weighted estimating equation,

$$0 = \frac{1}{n}\sum_{i=1}^{n}\frac{(2A_i - 1)Y_i}{g(A_i \mid L_i)} - \psi. \tag{4.21}$$

which is solved by the IPTW estimator $\hat{\psi}_n^{\text{IPTW}}$ we had in Example 1 on page 20, when substituting $\hat{g}_n$ for g. The performance of the IPTW estimators relies entirely on estimation of g. However, we can improve upon the method by augmenting the estimating equation in (4.21) so that the solution solves the efficient influence curve equation (Robins and Rotnitzky, 1992; van der Laan and Robins, 2003). Notably, by substituting $\hat{g}_n$ for g and $\hat{\bar{Q}}_n$ in the estimating equation,

$$0 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\mathbb{1}\{A_i = 1\} - \mathbb{1}\{A_i = 0\}}{g(A_i \mid L_i)}\right)\left(Y - \bar{Q}(A_i, L_i)\right) + \left(\bar{Q}(1, L_i) - \bar{Q}(0, L_i)\right) - \psi,$$

we obtain an estimator $\hat{\psi}_n^*$ that, per construction, solves the efficient influence curve equation for the estimation problem of our running example.

In Section 4.3 we briefly outlined the longitudinal (LTMLE) procedure as a technique for causal effect estimation based on longitudinal data. Our Manuscript I is motivated by the encounter of a similar type of longitudinal data. However, instead of imposing a discrete structure on the underlying time-scale, we consider data given continuously in time as discussed in the next chapters.

Chapter 5.
Counting processes, right-censoring and competing risks

So far we have considered data observed in discrete time with no censoring. In contrast, all our manuscripts are concerned with longitudinal data observed in continuous time subject to right-censoring and/or competing risks. In this chapter we give an informal introduction to the counting process framework utilized in survival analysis to provide a background for the manuscripts. Counting processes allow for a mathematical formulation of events happening at random points in time.

We start by introducing standard concepts for counting processes, taking a practical approach. Section 5.1.1 and Section 5.1.2 introduces right-censoring and competing risks, respectively, and Section 5.1.3 extends the setting to longitudinal data with time-varying covariates. We note that:

- Manuscript I is concerned with a continuous-time longitudinal data setting with time-varying covariates and treatment.
- Manuscript II and Manuscript III are concerned with right-censored data.
- Manuscript IV is concerned with competing risks data.

Key concepts in this chapter are the intensity (and the hazard) function, coarsening at random (and independent censoring), and the correspondence between the hazard rate and the failure probability (the risk). Both this chapter and the next, where we consider treatment effect estimation in the right-censored survival analysis setting of Section 5.1.1, will be useful for presenting our manuscripts later in Chapter 8.

5.1 Counting processes

Formally, a counting process $N(t)$ is a càdlàg stochastic process indexed by time with $N(0) = 0$ and paths that are piecewise constant. A counting process only has positive integer-valued and finite jumps. We introduce the notion of a filtration $(\mathcal{F}_t)_{t\geq 0}$ on the measure space $(\Omega, \mathcal{F})$, which is a family of increasing σ-algebras indexed by $t \geq 0$. We say that a stochastic process $X = (X(t))_{t\geq 0}$ is adapted to $(\mathcal{F}_t)_{t\geq 0}$ if $X(t)$ is $\mathcal{F}_t$-measurable for all $t \geq 0$, in which case we write $X(t) \in \mathcal{F}_t$. We say that X is predictable if X is $\mathcal{F}_{t-}$-adapted. The filtration generated by the stochastic process $\mathcal{F}_t = \sigma(X_s : 0 \leq s \leq t)$ is the smallest σ-algebra that makes X_s measurable for all

$s \in [0, t]$. Heuristically, the filtration carries information on the history of the process X up to time t.

We give a characterization of the distribution of a counting process in terms of its intensity. In the following we consider a filtration $(\mathcal{F}_t)_{t\geq 0}$ such that $N(t)$ is adapted to $\mathcal{F}_t$. The counting process has a compensator $\Lambda(t)$, a left-continuous and predictable process such that $M(t) = N(t) - \Lambda(t)$ is a martingale with respect to $\mathcal{F}_t$. An adapted stochastic process $(M(t))_{t\geq 0}$ is a martingale if $\mathbb{E}[|M_t|] < \infty$ for all $t \geq 0$ and $\mathbb{E}[M_t \,|\, \mathcal{F}_s] = M_s$ for all $s \leq t$. The intensity process $\lambda(t)$ of $N(t)$ is defined by,

$$\Lambda(t) = \int_0^t \lambda(s)ds, \tag{5.1}$$

and we also refer to the compensator $\Lambda(t)$ as the cumulative intensity. Since $M(t)$ is a martingale, we have that $\mathbb{E}[dM(t) \,|\, \mathcal{F}_{t-}] = 0$, which moreover implies that,

$$\mathbb{E}[dN(t) \,|\, \mathcal{F}_{t-}] = \mathbb{E}[d\Lambda(t) \,|\, \mathcal{F}_{t-}] = d\Lambda(t) = \lambda(t)dt.$$

We use the notation $d\Lambda(t) = \Lambda(dt)$ and $\lambda(t)dt$ interchangeably. Conveniently one can think of $d\Lambda(t)$ as $P(dN(t) = 1 \,|\, \mathcal{F}_{t-})$, i.e., the risk of an event in the interval $[t, t+dt)$. We define the times where $N(t)$ jumps, also referred to as the event times,

$$T_k = \inf\{t > 0 \,:\, N(t) > k - 1\}. \tag{5.2}$$

The relation between the conditional probability of a jump of $N(t)$ and the intensity of $N(t)$ further provides a connection to hazard functions. In the following, we let $T = T_1 \in \mathbb{R}_+$ be the time where the first jump of $N(t)$ occurs. We define the hazard function of the distribution of T as,

$$\alpha(t) = \lim_{dt\to 0} \frac{P(T \in [t, t + dt) \,|\, T \geq t)}{dt}, \tag{5.3}$$

which is interpreted as the instantaneous rate of an event immediately after t given that it did not happen before t. The cumulative hazard function is defined by $\mathrm{A}(t) = \int_0^t \alpha(s)ds$. If we denote the at-risk indicator by $R(t) = \mathbb{1}\{T \geq t\}$, the hazard $\alpha(t)$ for the distribution of T and the intensity of the counting process are related by,

$$\lambda(t) = R(t)\alpha(t), \tag{5.4}$$

that is, for subjects at risk, the hazard and the intensity coincide. The hazard rate is connected to the distribution of an event time, whereas the intensity function describes the occurrence of a sequence of events over time. Let F be the distribution of T. We identify the *survival function* by,

$$S(t) = 1 - F(t) = \prod_{s\leq t} \big(1 - \alpha(t)dt\big),$$

where $\prod$ denotes the product integral (Andersen et al., 1993, Section II.6). When F is absolutely continuous this reduces to,

$$S(t) = \exp\left(-\int_0^t \alpha(s)ds\right).$$

A multivariate counting process,

$$N = (N_j : j = 1, \ldots, J),$$

is a vector of $J \in \mathbb{N}$ $\mathcal{F}_t$-adapted counting processes counting J types of events. For example, $j = 1, \ldots, J$ may indicate different causes of death in a competing risks model (see Section 5.1.2). Each component N^j has a compensator Λ^j, and the likelihood based on observing $(N(s) : s \leq t)$ can be characterized by,

$$\mathcal{L}_t = \prod_{s \leq t} \prod_{j=1}^{J} \left(\lambda^j(s)ds\right)^{dN^j(s)} \left(1 - \lambda^j(s)ds\right)^{1-dN^j(s)}. \tag{5.5}$$

Informally, we can think of the data generated recursively from infinitesimal time interval to infinitesimal time interval, with the probability of a jump of $N^j(t)$ in $[t, t+dt)$ characterized by $d\Lambda^j(t) = \lambda^j(t)dt$.

5.1.1 Right-censoring in survival analysis

In survival analysis we are interested in characterizing the distribution of the time until a particular event of interest happens (the survival time). Often this event time is subject to right-censoring, with T being right-censored if it is only known that T is larger than some value. Letting C denote the right-censoring time, the observed time is $\tilde{T} = \min(T, C)$ and we let the variable $\Delta = \mathbb{1}\{T \leq C\}$ be the indicator of event. We consider a counting process formulation. Based on the observed data we define:

- The observed counting process $N^1(t) = \mathbb{1}\{\tilde{T} \leq t, \Delta = 1\}$ with intensity $\lambda^1(t)$.
- The observed counting process $N^c(t) = \mathbb{1}\{\tilde{T} \leq t, \Delta = 0\}$ with intensity $\lambda^c(t)$.
- The observed at-risk indicator $R(t) = \mathbb{1}\{\tilde{T} \geq t\}$.

We let $\mathcal{F}_t = \sigma(N^1(t), N^c(t) : 0 \leq s \leq t)$ denote the filtration generated by the observed data. The likelihood of the observed data factorizes as,

$$\mathcal{L}_t = \prod_{s \leq t} \left(\lambda^1(s)ds\right)^{dN(s)} \left(1 - \lambda^1(s)ds\right)^{1-dN(s)} \left(\lambda^c(s)ds\right)^{dN^c(s)} \left(1 - \lambda^c(s)ds\right)^{1-dN^c(s)}. \tag{5.6}$$

Under independent censoring (see, e.g., Andersen et al., 1993, Definition III.2.1), or, more generally, coarsening at random (see Remark 5.1 below), the intensity $\lambda^1(t)$ of the observed process $N^1(t)$ is again related to the hazard function $\alpha(t)$ of the distribution of T through,

$$\lambda^1(t) = R(t)\alpha(t),$$

and we can model the distribution of T accordingly. Particularly, we identify the survival probability as,

$$S(t) = P(T > t) = \prod_{s \leq t} \big(1 - \alpha(s)ds\big) = \exp\left(-\int_0^t \alpha(s)ds\right).$$

Remark 5.1 (Coarsening at random (CAR) (Heitjan and Rubin, 1991; Jacobsen and Keiding, 1995; Gill et al., 1997))**.** *Let X be the full data structure and C the coarsening mechanism such that, for a known many-to-one mapping Φ, $O = \Phi(X, C)$ is the observed data. Let $\mathcal{X}$ be the sample space of X and $\mathcal{C}$ the sample space of C. Further let,*

$$C(o) = \{x \in \mathcal{X} \,:\, \Phi(x, c) = o \;\; \textit{for some} \;\; c \in \mathcal{C}\},$$

be the subset of $\mathcal{X}$ with elements consistent with the observation o. A general definition of CAR is given in (Gill et al., 1997) and is formulated as (van der Laan and Robins, 2003, Section 1.2.3):

$$dP(o \,|\, X = x) = dP(o \,|\, X = x') \quad \textit{on} \quad \{o \,:\, x \in C(o)\} \cap \{o \,:\, x' \in C(o)\}.$$

Informally, given X, the coarsening mechanism only depends on the observed part $c(o)$ of x. For example, for the right-censored setting considered here, the full data structure is $X = T$ whereas the observed data are $O = (\tilde{T}, \Delta)$, and CAR implies that the hazard of C only depends on the history of the observed data (see also, van der Laan and Robins, 2003, Example 1.8). Note that CAR models for right-censored data is usually also formulated with a (baseline) covariate L included in both the full and the observed data structure. By that it becomes a conditional independence assumption, and the richer the information contained in L the more likely it is to hold.

5.1.2 Competing risks

Competing risks models are concerned with time-to-event data, where each subject may experience one of $J \geq 2$ mutually exclusive types of failures. A competing risks model is useful, for example, if we are interested in analyzing the time to death of a

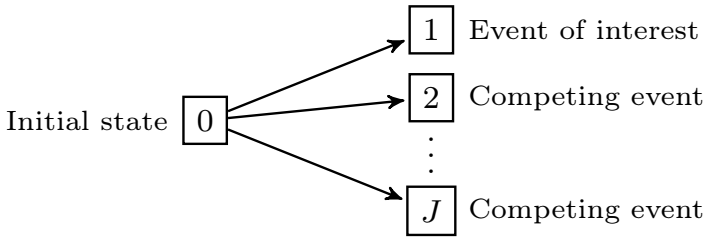


Figure 5.1: A multi-state representation of competing risks.

particular cause when other types of deaths are encountered as well.

As in Section 5.1.1, we let $\tilde{T} = \min(T, C)$, but now we observe only $\tilde{\Delta} = \mathbb{1}\{T \leq C\}\Delta$ where $\Delta \in \{1, \ldots, J\}$ is an indicator of the type of event, i.e., $\Delta = j$ if cause j was observed at time T, and $\tilde{\Delta} = 0$ indicates that the subject was right-censored. We define the observed multivariate counting process $N = (N^c, N^j : j = 1, \ldots, J)$ with $N^c(t) = \mathbb{1}\{\tilde{T} \leq t, \tilde{\Delta} = 0\}$ and $N^j(t) = \mathbb{1}\{\tilde{T} \leq t, \tilde{\Delta} = j\}$, $j = 1, \ldots, J$. The likelihood based on the observed data can be represented as,

$$\mathcal{L}_t = \prod_{s \leq t} \big(\lambda^c(s)ds\big)^{dN^c(s)} \big(1 - \lambda^c(s)ds\big)^{1-dN^c(s)} \prod_{j=1}^{J} \big(\lambda^j(s)ds\big)^{dN_j(s)} \big(1 - \lambda^j(s)ds\big)^{1-dN_j(s)}. \tag{5.7}$$

Under independent censoring, the intensity $\lambda^j(t)$ identifies the corresponding cause-j specific hazard function,

$$\alpha^j(t) = \lim_{dt \to 0} \frac{P(T \in [t, t + dt), \Delta = j \,|\, T \geq t)}{dt},$$

by the relation,

$$\lambda^j(t) = R(t)\alpha^j(t), \quad j = 1, \ldots, J.$$

The cause-j specific hazard $\alpha^j(t)$ at time t represents the risk of event j just after t for subjects who are event-free until just before time t, and thus characterizes the instantaneous probability of a jump from the initial state to state j (Figure 5.1).

In the situation without competing risks ($J = 1$), there is a one-to-one correspondence between the survival function and the hazard function, i.e., between the risk and the rate. One key aspect of competing risks analysis is that this is no longer the case. Accordingly, another way of analyzing the distribution of the time to a particular

event type of interest is by the cumulative incidence,

$$\begin{aligned} F^j(t) = P(T \leq t, \Delta = j) &= \int_0^t S(s-)\lambda^j(s)\, ds \\ &= \int_0^t \lambda^j(s) \exp\Big(-\sum_{l=1}^{J} \int_0^s \lambda^l(u)\, du \Big) ds. \end{aligned} \tag{5.8}$$

Notably, this also depends on other causes $l \neq j$ than the one of interest.

5.1.3 Longitudinal data with time-varying covariates

We can use a representation of longitudinal data with time-varying covariates in terms of counting processes, letting intensities characterize the rate of a finite number of continuous monitoring times for each individual conditional on their observed history. Suppose time-varying covariates only consist of a single process $N^\ell(t)$ that records, for example, hospital admissions. The subject-specific longitudinal data are collected in a multivariate counting process (N^1, N^c, N^ℓ) that tracks failure, censoring and covariate events. Jumps of N^1, N^c are terminating events, whereas N^ℓ can jump many times. We may represent our observed data as,

$$\bar{O}(t) = \big(s, N^1(s), N^c(s), N^\ell(s) \,:\, s \in \{T_k\}_{k=0}^{K_t}\big),$$

where T_k is the time of the kth monitoring of the subject, and $K_t < \infty$ is the number of monitoring times before time t. The likelihood can now be expressed as,

$$\begin{aligned} \mathcal{L}_t = \prod_{s \leq t} &\Lambda^1(ds)^{N(ds)} \big(1 - \Lambda^1(ds)\big)^{1-N^1(ds)} \Lambda^c(ds)^{N^c(ds)} \big(1 - \Lambda^c(ds)\big)^{1-N^c(ds)} \\ &\Lambda^\ell(ds)^{N^\ell(ds)} \big(1 - \Lambda^\ell(ds)\big)^{1-N^\ell(ds)}. \end{aligned}$$

A marked point process model provides a useful generalization to allow for more complex covariate events. We say that N^ℓ is a marked point process with respect to a given measurable mark space $(\mathcal{X}, \mathscr{X})$ if N^ℓ takes value in the product space $(\mathcal{T} \times \mathcal{X}, \mathscr{B}(\mathcal{T}) \otimes \mathscr{X})$. Here $\mathscr{B}$ denotes the Borel σ-algebra and $\mathcal{T}$ an interval of time. The compensator Λ^ℓ of N^ℓ defines a measure on $(\mathcal{T} \times \mathcal{X}, \mathscr{B}(\mathcal{T}) \otimes \mathscr{X})$. For sake of presentation, we assume that $\mathcal{X}$ is finite. Conditional on the past at time t, there is an event of N^ℓ in the time interval $[t, t + dt)$ with a mark at x with probability $\Lambda^\ell(dt, x)$ and there is no event with probability $1 - \Lambda^\ell(dt, \mathcal{X})$. Now, we can present the likelihood as,

$$\begin{aligned} \mathcal{L}_t = \prod_{s \leq t} &\Lambda^1(ds)^{N^1(ds)} \big(1 - \Lambda^1(ds)\big)^{1-N^1(ds)} \Lambda^c(ds)^{N^c(ds)} \big(1 - \Lambda^c(ds)\big)^{1-N^c(ds)} \\ &\prod_{x \in \mathcal{X}} \Lambda^\ell(ds, x)^{N^\ell(ds,x)} \big(1 - \Lambda^\ell(ds, \mathcal{X})\big)^{1-N^\ell(ds,\mathcal{X})}. \end{aligned}$$

This is a slightly different version of the likelihood that we work with in Manuscript I, where we have an additional process $N^a(t)$ tracking treatment changes.

5.2 Final comments

In event history analysis, the goal is to analyze the time until some event of interest happens. Often we model the intensities that characterize the process by which events happen as a function of covariate history for all time-points. A popular method in medical research for taking covariate information into account when characterizing the distribution of an event time is the Cox regression model (Cox, 1972). The Cox model assumes a multiplicative effect of covariates on the intensities, whereby covariate changes are associated with a higher or lower event rate compared to a baseline. A baseline hazard $\lambda_0(t)$ represents the change in the hazard rate over time. Consider particularly a setting with $A \in \mathcal{F}_0$ a baseline treatment variable, $L_0 \in \mathcal{F}_0$ a baseline (pre-treatment) covariate and $L(t) \in \mathcal{F}_t$ a time-varying covariate vector. We may then posit a Cox proportional hazards model as follows:

$$\lambda(t \mid L_0, A, L(t)) = \lambda_0(t) \exp(\alpha L_0 + \beta^\top L(t) + \psi A). \tag{5.9}$$

Say we are interested in ψ. Notably, this equals,

$$\log\left(\frac{\lambda(t \mid L_0, 1, L(t))}{\lambda(t \mid L_0, 0, L(t))}\right) = \psi,$$

but the interpretation is less clear, as we are comparing individuals with the same values of covariates $L(t)$ at time t. Since past covariate values may have been affected by treatment, controlling for covariates that are after treatment may block some of the treatment effect. If we exclude the covariates, on the other hand, we fail to model the event rate accurately if the data were truly generated according to (5.9). This is the same issue we discussed in Section 1.1.1 (and was pointed out by Robins, 1986, already in 1986).

In our manuscripts presented later, we have two aims that we keep distinct:

(1) To analyze treatment effects on time-to-event outcomes in presence of covariate-dependent censoring and competing risks without imposing, possibly restrictive, model assumptions such as proportionality of hazards.

(2) To incorporate time-dependent covariates such as to enable treatment effect estimation with a sound interpretation.

Manuscript I is particularly focused on (2) but also presents a framework for (1). Manuscripts III–IV are concerned with (1). In the next chapter we consider treatment effect estimation in the survival analysis setting from Section 5.1.1 to discuss a few more aspects relevant for the manuscripts.

Chapter 6.
Causal survival analysis

In this chapter we consider treatment effect estimation in the right-censored survival analysis setting from Section 5.1.1. This serves as an introduction to Manuscript III and Manuscript IV, but can also be viewed as a special-case of Manuscript I. Altogether, this chapter gives an overview of different aspects relevant for the manuscripts. We summarize a few points:

- Censoring adds another layer to the counterfactual analysis introduced in Chapter 2. Throughout this chapter we consider treatment and censoring acting together as a coarsening mechanism, as we detail below.
- Inverse probability weighting by the distribution of coarsening mechanisms identifies parameters representing treatment effects. Manuscripts III and IV follow this approach, along the lines of Section 6.1.1.
- The discrete longitudinal setting considered in Chapter 2 and Section 4.3 also covers right-censored data structures. We give a brief overview of existing TMLE methods for right-censored data analysis in Section 6.2.
- All existing TMLE procedures assume that time is discrete. For the survival analysis setting with only baseline covariates and treatment, the discrete-time TMLE (as, e.g., presented by Moore and van der Laan, 2009a) generalizes more or less directly to a continuous-time version. Section 6.3 outlines such a TMLE procedure for continuous-time survival analysis. We present the efficient influence function that we use for the procedure in Section 6.1.2.

We continue from Section 5.1.1. Assume that we have observations of an event time $\tilde{T} = \min(T, C)$, an event indicator $\Delta = \mathbb{1}\{T \leq C\} \in \{0, 1\}$, and now also a baseline treatment $A \in \{0, 1\}$ and a (pre-treatment) baseline covariate vector $L \in \mathbb{R}^p$. As in Chapter 2, we introduce counterfactuals. Let T^a, for $a = 0, 1$, be the counterfactual event time had treatment been $A = a$. We make the consistency assumption that the observed data are related to the counterfactuals in terms of $\tilde{T} = \min(T^A, C)$ and $\Delta = \mathbb{1}\{T^A \leq C\}$. Notice that T^0, T^1 are only partly observed for two reasons: Firstly since any subject was only either treated or not, and, secondly, due to censoring. In the causal framework we can think of censoring as adding another counterfactual layer, with T^a the counterfactual event time had there been no censoring and had treatment been $A = a$. This gives a general formulation of treatment and censoring acting together as a coarsening mechanism, where the coarsening at random assumption (see

Remark 5.1 from the previous chapter) allows for identification of the distribution of counterfactual event times (van der Laan and Robins, 2003; Tsiatis, 2007). We further note that (as shown by Robins et al., 2000) the sequential randomization assumption (A2) from Chapter 2 is equivalent to a coarsening at random assumption under reasonably general conditions. Here the full data are the set of counterfactual treatment-specific outcomes $(Y^{\bar{a}_K} : \bar{a}_K)$, and the observed data (under consistency) are the actual treatment received and the corresponding treatment-specific outcome $(\bar{A}_K, Y^{\bar{A}_K})$.

6.1 Target parameter

We let $P_0 \in \mathcal{M}$ denote the distribution of the observed data $O = (\tilde{T}, \Delta, A, L)$ and $(\mathcal{F}_t)_{t\geq 0}$ the filtration generated by the observed data (so that $A, L \in \mathcal{F}_t$ for all $t \geq 0$). The likelihood p_0 of the observed data factorizes as,

$$p_0(O) = q_L(L) \prod_t \big(\lambda^1(t \,|\, A, L)dt\big)^{N^1(dt)} \big(1 - \lambda^1(t \,|\, A, L)dt\big)^{1-N^1(dt)} \tag{6.1}$$

$$g(A \,|\, L)\big(\lambda^c(t \,|\, A, L)dt\big)^{N^c(dt)} \big(1 - \lambda^c(t \,|\, A, L)dt\big)^{1-N^c(dt)}, \tag{6.2}$$

where $\lambda^1(t \,|\, A, L)$ is the conditional intensity for the observed counting process $N^1(t) = \mathbb{1}\{\tilde{T} \leq t, \Delta = 1\}$, i.e.,

$$\mathbb{E}[dN^1(t) \,|\, \mathcal{F}_{t-}] = \mathbb{E}[d\Lambda^1(t \,|\, A, L) \,|\, \mathcal{F}_{t-}] = d\Lambda^1(t \,|\, A, L) = \lambda^1(t \,|\, A, L)dt,$$

likewise, $\lambda^c(s \,|\, A, L)$ is the conditional intensity for the process $N^c(t) = \mathbb{1}\{T \leq t, \Delta = 0\}$, q_L is the distribution of baseline covariates L and $g(a \,|\, L) = P(A = a \,|\, L)$ is the distribution of treatment A conditional on L. In the factorization, (6.2) represents the interventional part (G) and (6.1) the non-interventional part (Q). We are interested in characterizing the distribution of the counterfactual event time T^a. Essentially we achieve this by replacing G by G^* that puts all mass in no censoring and treatment $A = a$. To motivate this, however, we need to relate the resulting g-computation formula P_{Q,G^*} to the distribution of the counterfactual T^a.

We represent the observed data O as a missing data structure on the full data structure (L, T^1, T^0) coarsened by (A, C), and assume coarsening at random. Under coarsening at random, we can now identify the conditional hazard $\alpha(t \,|\, A, L)$ of the underlying event time T from the intensity of the observed counting process $N^1(t)$:

$$\lambda^1(t \,|\, A, L) = \mathbb{1}\{\tilde{T} \geq t\}\, \alpha(t \,|\, A, L).$$

This further means that we can identify the counterfactual survival probability,

$$P(T^a > t) = \mathbb{E}\big[P(T^a > t \,|\, L)\big] = \mathbb{E}\big[P(T > t \,|\, A = a, L)\big] = \prod_{s \leq t} \big(1 - \alpha(s \,|\, a, L)ds\big).$$

We now fix a time-point $t_0 > 0$ and define the target parameter $\Psi_{t_0} : \mathcal{M} \to \mathbb{R}$ by,

$$\Psi_{t_0}(P) = P(T^1 > t_0) - P(T^0 > t_0), \tag{6.3}$$

that is, the difference in survival beyond time t_0 in the counterfactual scenario where everyone is treated versus the counterfactual scenario where no one is treated.

The positivity assumption in this setting becomes,

$$P(C > t_0 \,|\, a, L)P(A = a \,|\, L) > \eta > 0 \quad a.s.,$$

for $a = 0, 1$, and we note that this may restrict possible choices for t_0 in a given application.

6.1.1 Inverse probability weighting

An alternative characterization of the target parameter in (6.3) can be given as follows. Under coarsening at random, we can factorize the observed data distribution as,

$$\begin{aligned}
&P(\tilde{T} \in dt, \Delta = 1, A = a, L \in d\ell) \\
&\quad = P(\Delta = 1 \,|\, T^a = t, A = a, L = \ell)P(A = a \,|\, T^a = t, L = \ell)P(T^a \in dt, L \in d\ell) \\
&\quad = P(C > t \,|\, A = a, L = \ell)P(A = a \,|\, L = \ell)P(T^a \in dt \,|\, L = \ell)P(L \in d\ell),
\end{aligned}$$

for $a = 0, 1$, which implies that,

$$P(T^a \in dt \,|\, L = \ell) = \frac{P(\tilde{T} \in dt, \Delta = 1, A = a \,|\, L = \ell)}{P(C > t \,|\, A = a, L = \ell)P(A = a \,|\, L = \ell)}.$$

This further means that,

$$\begin{aligned}
P(T^a > t_0 \,|\, L) &= \int_{t_0}^{\infty} P(T^a \in dt \,|\, L) \\
&= \int_{t_0}^{\infty} \frac{P(\tilde{T} \in dt, \Delta = 1, A = a \,|\, L)}{P(C > t \,|\, A = a, L = \ell)P(A = a \,|\, L)},
\end{aligned}$$

which immediately suggests an inverse probability weighted estimator for $P(T^a > t_0 \,|\, L)$. Let $\hat{\bar{G}}_n(t \,|\, A, L)$ denote an estimator for $\bar{G}(t \,|\, A, L) = P(C > t \,|\, A, L)$ and $\hat{g}_n(L)$ an estimator for $P(A = a \,|\, L)$.[1] Then a consistent estimator for (6.3) can be obtained as,

$$\hat{\psi}_n^{\text{IPW}} = \frac{1}{n} \sum_{i=1}^{n} \frac{(2A_i - 1)\mathbb{1}\{\tilde{T}_i > t_0\}}{\hat{\bar{G}}_n(t_0 \,|\, A_i, L_i)\hat{g}_n(A_i \,|\, L_i)}, \tag{6.4}$$

if $\hat{\bar{G}}_n(t \,|\, A, L)$ and $\hat{g}_n(L)$ are both consistent.

[1] Notice that we try to follow the notation from Manuscripts III and IV, but, to avoid confusion, we use $\bar{G}$ (rather than G) to denote the censoring survival distribution.

6.1.2 Efficient influence function

As we have discussed in earlier chapters, the efficient influence function for the estimation problem guides the construction of efficient estimators. Once we have the efficient influence function, we can consider TMLE methodology, or, as was mentioned in the end of Chapter 4, we can improve upon the inverse probability weighted estimator (6.4) directly by considering the efficient influence curve equation (van der Laan and Robins, 2003). In the next section we, again, consider the TMLE framework.

We here sketch how one may derive an expression for the efficient influence function for the parameter defined in (6.3) in the CAR model that assumes G to be known. This is then also the efficient influence function in the model with unknown G (van der Laan and Robins, 2003). Under CAR we have that:

$$\bar{G}(t \mid A, L) = P(C > t \mid A, L) = \prod_{s \leq t} \big(1 - \Lambda^c(dt \mid A, L)\big),$$
$$S(t \mid A, L) = P(T > t \mid A, L) = \prod_{s \leq t} \big(1 - \Lambda^1(dt \mid A, L)\big).$$

We use the influence function for the IPW estimator from (6.4) in Section 6.1.2 as an initial gradient,

$$\phi_{\text{IPW}}(P)(O) = \frac{\mathbb{1}\{A = 1\} - \mathbb{1}\{A = 0\}}{g(A \mid L) \prod_{t \leq t_0}(1 - \Lambda^c(dt \mid A, L))} 1\{\tilde{T} > t_0\} - \Psi(P), \tag{6.5}$$

Now, we need to project $\phi_{\text{IPW}}(P)$ onto the tangent space $\mathscr{T}_Q(P) = \mathscr{T}_{q_L}(P) \oplus \mathscr{T}_{\Lambda^1}(P)$ where $Q = (q_L, \Lambda^1)$. The projection onto $\mathscr{T}_{q_L}(P)$ equals $\mathbb{E}[\phi_{\text{IPW}}(P)(O) \mid L] = P(T > t_0 \mid A = 1, L) - P(T > t_0 \mid A = 0, L) - \Psi(P)$. The projection onto $\mathscr{T}_{\Lambda^1}(P)$ equals,

$$\int \Big(\mathbb{E}[\phi_{\text{IPW}}(P)(O) \mid N^1(dt) = 1, \mathcal{F}_{t-}] \\ - \mathbb{E}[\phi_{\text{IPW}}(P)(O) \mid N^1(dt) = 0, \mathcal{F}_{t-}] \Big) \big(N(dt) - \Lambda^1(dt)\big).$$

If we define,

$$H_t(O) := \mathbb{E}[\phi_{\text{IPW}}(P)(O) \mid N^1(dt) = 1, \mathcal{F}_{t-}] - \mathbb{E}[\phi_{\text{IPW}}(P)(O) \mid N^1(dt) = 0, \mathcal{F}_{t-}], \tag{6.6}$$

we can present the efficient influence function as,

$$\begin{aligned} \phi^*(P)(O) = &\int H_t(O) \big(N^1(dt) - \Lambda^1(dt)\big) \\ &+ P(T > t_0 \mid A = 1, L) - P(T > t_0 \mid A = 0, L) - \Psi(P). \end{aligned} \tag{6.7}$$

We note that,

$$\mathbb{E}[\phi_{\text{IPW}}(P)(O) \mid N^1(dt) = 1, \mathcal{F}_{t-}] = \left(\frac{\mathbb{1}\{A=1\} - \mathbb{1}\{A=0\}}{g(A \mid L)}\right) \mathbb{E}\left[\frac{1\{\tilde{T} > t_0\}}{\prod_{t \leq t_0}(1 - \Lambda^c(dt \mid A, L))} \,\middle|\, N^1(dt) = 1, \mathcal{F}_{t-}\right],$$

where,

$$\mathbb{E}\left[\frac{1\{\tilde{T} > t_0\}}{\prod_{t \leq t_0}(1 - \Lambda^c(dt \mid A, L))} \,\middle|\, N^1(dt) = 1, \mathcal{F}_{t-}\right] = \frac{\mathbb{1}\{t > t_0\}}{\prod_{t \leq t_0}(1 - \Lambda^c(dt \mid A, L))}.$$

Likewise,

$$\mathbb{E}[\phi_{\text{IPW}}(P)(O) \mid N^1(dt) = 0, \mathcal{F}_{t-}] = \left(\frac{\mathbb{1}\{A=1\}}{g(1 \mid L)} - \frac{\mathbb{1}\{A=0\}}{g(0 \mid L)}\right) \mathbb{E}\left[\frac{1\{\tilde{T} > t_0\}}{\prod_{t \leq t_0}(1 - \Lambda^c(dt \mid A, L))} \,\middle|\, N^1(dt) = 0, \mathcal{F}_{t-}\right],$$

where,

$$\begin{aligned}
&\mathbb{E}\left[\frac{1\{\tilde{T} > t_0\}}{\prod_{t \leq t_0}(1 - \Lambda^c(dt \mid A, L))} \,\middle|\, N^1(dt) = 0, \mathcal{F}_{t-}\right] \\
&= \frac{\mathbb{1}\{t > t_0\}}{\prod_{t \leq t_0}(1 - \Lambda^c(dt \mid A, L))} + \frac{\mathbb{1}\{t \leq t_0\}}{\prod_{s < t}(1 - \Lambda^c(ds \mid A, L))} P(T > t_0 \mid T > t, A, L) \\
&= \frac{\mathbb{1}\{t > t_0\}}{\prod_{t \leq t_0}(1 - \Lambda^c(dt \mid A, L))} + \frac{\mathbb{1}\{t \leq t_0\}}{\prod_{s < t}(1 - \Lambda^c(ds \mid A, L))} \frac{P(T > t_0 \mid A, L)}{P(T > t \mid A, L)}.
\end{aligned}$$

This now means that we can write the clever covariate in (6.6) as:

$$H_t(O) = -\left(\frac{\mathbb{1}\{A=1\} - \mathbb{1}\{A=0\}}{g(A \mid L)}\right) \frac{\mathbb{1}\{t \leq t_0\}}{\prod_{s < t}(1 - \Lambda^c(ds \mid A, L))} \frac{P(T > t_0 \mid A, L)}{P(T > t \mid A, L)}. \tag{6.8}$$

In Section 6.3 we outline a TMLE algorithm that uses the representation of the efficient influence function in (6.7) with the time-varying clever covariate H_t defined by (6.8). We also use the same influence function in Manuscript III, although it admits a different representation (that they are equivalent can be realized by combining van der Laan and Robins, 2003, Example 1.12 and Theorem 1.1).

6.2 TMLE for (discrete-time) survival analysis

We give a brief overview of some of the existing TMLE methods for survival analysis. We make the following distinction between these methods:

(1) Estimation of treatment-specific discrete-time survival at a fixed timepoint for right-censored survival outcomes using full likelihood based TMLE (van der Laan and Rubin, 2006; Moore and van der Laan, 2009a,b,c; Stitelman et al., 2011; Stitelman and van der Laan, 2011; van der Laan and Rose, 2011; Benkeser, 2015).

(2) Estimation of treatment-specific discrete-time survival at a fixed timepoint for right-censored survival outcomes using time-varying covariate-adjusted TMLE (LTMLE) (van der Laan and Gruber, 2012; Schnitzer et al., 2014; Petersen et al., 2014; Lendle et al., 2017).

Furthermore, Manuscript I of this thesis provides the groundwork for:

(3) Estimation of treatment-specific continuous-time survival at a fixed timepoint for right-censored survival outcomes using time-varying covariate-adjusted TMLE.

The TMLE methods (1) and (2) all consider the setting where each subject is monitored at discrete event times. This means that $\tilde{T}$ (and covariates and treatment) only takes values in the set $\{1, 2, \ldots, \tau\}$. In this case we can represent the observed data as in Section 2.2:

$$O = (L_0, A_1, L_1, A_2, \ldots, L_\tau, A_\tau, L_{\tau+1}),$$

with $N^1(t) = \mathbb{1}\{\tilde{T} \leq t, \Delta = 1\}$ being part of $L_t = (L_t^1, N^1(t))$ and $N^c(t) = \mathbb{1}\{\tilde{T} \leq t, \Delta = 0\}$ being part of the intervention variable, which now includes both censoring and treatment, $A_t = (A_t^1, N^c(t))$. Note that A_t^1 denotes treatment at time t and L_t^1 denotes time-varying covariates at time t (which may both be empty sets for $t \geq 1$ if we do not consider time-varying covariates and treatment). When $N^1(t) = 1$ or $N^c(t) = 1$, the variables are encoded with their last known value. The LTMLE procedure (2) (van der Laan and Gruber, 2012; Schnitzer et al., 2014; Petersen et al., 2014; Lendle et al., 2017) summarized in Section 4.3 applies directly to this setting, with the sequential regression steps at each step t carried out only among subjects uncensored and alive by this time.

The full likelihood based TMLEs (1) (van der Laan and Rubin, 2006; Moore and van der Laan, 2009a,b,c; Stitelman et al., 2011; Stitelman and van der Laan, 2011; van der Laan and Rose, 2011; Benkeser, 2015) are referred to as targeted *maximum likelihood* estimation methods, rather than targeted *minimum loss-based* estimation. Here all components of the likelihood (including potentially very high-dimensional covariate densities) are estimated and then updated in an iterative manner that is targeted towards the parameter of interest. In the next section we describe how one may define a TMLE procedure for the continuous-time setting of Section 6.1 that follows along the lines of the TMLE (targeted *maximum likelihood* estimation) procedure for discrete-time survival analysis.

6.3 Sketch of TMLE for continuous-time survival analysis

We sketch a TMLE procedure for the continuous-time survival analysis setting with baseline confounding. We emphasize that more details for a more complicated situation is found in Manuscript I. We consider estimation of the target parameter $\Psi : \mathcal{M} \to \mathbb{R}$ defined in (6.3) with influence function $\phi^*(P)$ displayed in (6.7).

Initial estimation. We need initial estimators for the censoring mechanism (Λ^c), for the treatment distribution (g), for the covariate distribution (q_L) and for the survival mechanism (Λ^1). For q_L we use the empirical distribution. We estimate g by $\hat{g}_n$ as a binary regression with outcome A and covariates L. For the sake of presentation, we form initial estimators for the intensities based on Cox regression working models.[2] Specifically,

$$\lambda^c(t \mid A, L) = R(t)\lambda_0^c(t)\exp(\alpha L + \gamma A),$$
$$\lambda^1(t \mid A, L) = R(t)\lambda_0^1(t)\exp(\beta L + \psi A),$$

where $R(t) = \mathbb{1}\{\tilde{T} \geq t\}$ denotes the at-risk indicator. We estimate the coefficients (α, γ) and (β, ψ) with their partial maximum likelihood estimators, and we use the Breslow estimator (Breslow, 1974) for the baseline cumulative hazard functions,

$$\hat{\Lambda}_{0,n}^c(t) = \int_0^t \frac{\sum_{i=1}^n dN_i^c(s)}{\sum_{i=1}^n R_i(s)\exp(\hat{\alpha}L_i + \hat{\gamma}A_i)}, \tag{6.9}$$

$$\hat{\Lambda}_{0,n}^1(t) = \int_0^t \frac{\sum_{i=1}^n dN_i^1(s)}{\sum_{i=1}^n R_i(s)\exp(\hat{\beta}L_i + \hat{\psi}A_i)}. \tag{6.10}$$

We note that $\hat{\Lambda}_{0,n}^c(t)$ jumps at the censoring times, and that $\hat{\Lambda}_{0,n}^1(t)$ jumps at the event times.

For each subject i and each time t, we compute,[3]

$$d\hat{\Lambda}_{k=0}^1(t \mid a, L) = \exp(\hat{\beta}L + \hat{\psi}a)\frac{\sum_{j=1}^n dN_j^1(t)}{\sum_{j=1}^n R_j(t)\exp(\hat{\beta}L_j + \hat{\psi}A_j)}, \quad a = 0, 1.$$

Note that this estimator is a step function of time, with changes at all event times in the data. This defines an initial estimator for,

$$\hat{\Lambda}_{k=0}^1(t \mid a, L) = \int_0^t d\hat{\Lambda}_{k=0}^1(s \mid a, L) = \exp(\hat{\beta}L + \hat{\psi}a)\,\hat{\Lambda}_{0,n}^1(t),$$

[2] In applications we would use super learning, see comments in Chapter 9.

[3] The notation here is a little tricky; note that '$k = 0$' is used to indicate 'initial estimator', not to be mistaken with $\Lambda_{0,n}^1$ which is the Breslow estimator for the baseline cumulative hazard.

based on which we form an initial estimator for the survival probability at time t,

$$\hat{S}_{k=0}(t \mid a, L) = \exp\big(-\hat{\Lambda}^1_{k=0}(t \mid a, L)\big). \tag{6.11}$$

This suggests an initial estimate of the target parameter defined by,

$$\hat{\Psi}^{k=0}_n = \frac{1}{n}\sum_{i=1}^{n}\big(\hat{S}_{k=0}(t \mid 1, L) - \hat{S}_{k=0}(t \mid 0, L)\big).$$

Targeting. Recall that the purpose of the targeting step is to update the initial estimator such as to solve the efficient influence curve equation. We presented an expression for the efficient influence function in (6.7) with a time-varying clever covariate H_t defined by (6.8). Note that the efficient influence $\phi^*(P)$ function depends on P only through $(\Lambda^1, g, \Lambda^c)$. We abuse notation and write $\phi^*(\Lambda^1, g, \Lambda^c)$. Likewise it is noted that the clever covariate H_t also depends on $(\Lambda^1, g, \Lambda^c)$, whereas our target parameter depends only on Λ^1.

To carry out the targeting step, we need an estimator for the clever covariate H_t. It was noted that H_t depends on Λ^c but only through the censoring survival function $G(t \mid A, L) = \exp(-\Lambda^c(t \mid A, L))$ and on Λ^1 but only through the survival function. We estimate the censoring survival function by,

$$\hat{\bar{G}}_n(t \mid A, L) = \exp\left(-\exp(\hat{\alpha}L + \hat{\gamma}A)\int_0^t \frac{\sum_{i=1}^n dN^c_i(s)}{\sum_{i=1}^n R_i(s)\exp(\hat{\alpha}L_i + \hat{\gamma}A_i)}\right).$$

Further, we substitute our estimator $\hat{g}_n$ for g and our initial estimator $\hat{S}_{k=0}(t \mid A, L)$ from (6.11) for $S(t \mid A, L)$. Accordingly, we obtain,

$$\hat{H}_{t,k=0}(A, L) = -\left(\frac{\mathbb{1}\{A=1\} - \mathbb{1}\{A=0\}}{\hat{g}_n(A \mid L)}\right)\frac{\mathbb{1}\{t \le t_0\}}{\hat{\bar{G}}_n(t- \mid A, L)}\left(-\frac{\hat{S}_{k=0}(t_0 \mid A, L)}{\hat{S}_{k=0}(t \mid A, L)}\right).$$

Note that the estimator for the clever covariate is indexed by '$k = 0$' since it depends on the estimator $\hat{S}_{k=0}(t \mid A, L)$, which, as we will see, is part of the updating procedure.

Our aim is to update the estimator $\hat{\Lambda}^1_{k=0}(t \mid A, L)$ in a way such as to solve the efficient influence curve equation. As dictated by Section 4.1, we need the following:

(i) A choice of a loss function for λ^1, $(O, \lambda^1) \mapsto \mathcal{L}(\lambda^1)(O)$. Here we consider the partial log-likelihood loss function for λ^1:

$$\mathcal{L}(\lambda^1) = \int_0^{t_0} \log\lambda^1(t)\, dN^1(t) - \lambda^1(t)dt.$$

(ii) A parametric fluctuation submodel parametrized by $\varepsilon \in \mathbb{R}$ through λ^1 at $\varepsilon = 0$, $\{\lambda^1_\varepsilon : \varepsilon \in \mathbb{R}\}$. Here we define the submodel:

$$\lambda^1_\varepsilon(t) = \lambda^1(t) \exp\big(\varepsilon H_t\big),$$

(iii) An updating algorithm.

For the choices of (i)–(ii) we note that:

$$\frac{d}{d\varepsilon}\bigg|_{\varepsilon=0} \mathcal{L}(\lambda^1_\varepsilon) = \int_0^{t_0} H_t\left(dN^1(t) - \lambda^1(t)dt\right) = \int_0^{t_0} H_t\left(dN^1(t) - d\Lambda^1(t)\right),$$

as desired, recognizing the expression on the right hand side as the first term of the efficient influence curve equation displayed in (6.7). Now, consider the submodel through the initial estimator $d\hat{\Lambda}^1_{k=0}(t) = \hat{\lambda}^1_{k=0}(t)dt$. Finding $\hat{\varepsilon}_{n,1}$ that minimizes $\varepsilon \mapsto \mathbb{P}_n \mathcal{L}(\hat{\lambda}^1_{k=0,\varepsilon})$, for the given estimator $\hat{H}_{t,k=0}$, corresponds to solving the equation,

$$\frac{1}{n}\sum_{i=1}^n \Big(\int_0^\tau \mathbb{1}\{t \le t_0\} \hat{H}_{t,k=0}(A_i, L_i)\big(dN_i(t) - \exp\big(\varepsilon \hat{H}_{t,k=0}(A_i, L_i)\big) d\hat{\Lambda}^1_{k=0}(t \mid A_i, L_i)\big)\Big) = 0,$$

with respect to ε. The updated estimator,

$$d\hat{\Lambda}^1_{k=1}(t \mid a, L) := d\hat{\Lambda}^1_{k=0}(t \mid a, L) \exp\big(\hat{\varepsilon}_{n,1} \hat{H}^1_{t,k=0}(a, L)\big), \quad \text{for } a = 0, 1,$$

now solves the relevant part of the efficient influence curve equation,

$$\mathbb{P}_n \int_0^{t_0} \hat{H}^1_{t,k=0}\left(dN^1(t) - d\hat{\Lambda}^1_{k=1}(t)\right) = 0,$$

however, with respect to the not yet updated $\hat{H}^1_{t,k=0}$. This means that the steps above must be iterated. Below we describe the step to update a current estimator $\hat{\Lambda}^1_k(t \mid a, L)$ into $\hat{\Lambda}^1_{k+1}(t \mid a, L)$:

1. Compute the current survival function for all $t \le t_0$ based on $\hat{\Lambda}^1_k(t \mid a, L)$,

$$\hat{S}^1_k(t \mid a, L) = \exp\big(-\hat{\Lambda}^1_k(t \mid a, L)\big), \quad a = 0, 1.$$

2. Compute the current value of the clever covariates for all $t \le t_0$:

$$\hat{H}^1_{t,k}(a, L) = -\left(\frac{\mathbb{1}\{A = 1\} - \mathbb{1}\{A = 0\}}{\hat{g}_n(A \mid L)}\right) \frac{\mathbb{1}\{t \le t_0\}}{\hat{\bar{G}}_n(t - \mid A, L)} \left(-\frac{\hat{S}_k(t_0 \mid A, L)}{\hat{S}_k(t \mid A, L)}\right)$$

3. Estimate $\hat{\varepsilon}_n$ as the solution to the following equation:

$$\frac{1}{n}\sum_{i=1}^{n}\Big(\int_0^{\tau}\mathbb{1}\{t\le t_0\}\hat{H}_{t,k}(A_i,L_i)\big(dN_i(t)-\exp\big(\varepsilon\hat{H}_{t,k}(A_i,L_i)\big)d\hat{\Lambda}^1_k(t\,|\,A_i,L_i)\big)\Big) = 0.$$

4. Use $\hat{\varepsilon}_n$ to update:

$$d\hat{\Lambda}^1_{k+1}(t\,|\,a,L) := d\hat{\Lambda}^1_k(t\,|\,a,L)\exp\big(\hat{\varepsilon}_n\hat{H}^1_{t,k}(a,L)\big),\quad a=0,1.$$

5. Compute:

$$\hat{\Lambda}^1_{k+1}(t\,|\,a,L)=\int_0^t d\hat{\Lambda}^1_{k+1}(t\,|\,a,L),\quad a=0,1.$$

The iterations are continued until $|\,\mathbb{P}_n\phi^*(\hat{\Lambda}^1_{k^*},\hat{g}_n,\hat{\Lambda}^c_n)\,|<\gamma_n$, where $\gamma_n=o_P(n^{-1/2})$ is some stopping criterion to ensure that we solve the efficient influence curve equation up to order $o_P(n^{-1/2})$. We define the final estimator for the counterfactual survival probability as,

$$\hat{S}_{k^*}(t_0\,|\,a,L)=\exp\left(-\int_0^{t_0}d\hat{\Lambda}^1_{k^*}(t\,|\,a,L)\right),\qquad a=0,1,$$

and estimate the target parameter by,

$$\hat{\Psi}^*_n=\frac{1}{n}\sum_{i=1}^{n}\big(\hat{S}_{k^*}(t_0\,|\,1,L_i)-\hat{S}_{k^*}(t_0\,|\,1,L_i)\big),\tag{6.12}$$

which is the final TMLE estimator Ψ^*_n for our target parameter. If the conditions of Theorem 3.3 are satisfied, Ψ^*_n follows an asymptotic normal distribution around the true value with a standard error that we can estimate by $\hat{\sigma}^2_n$,

$$\hat{\sigma}^2_n=\frac{1}{n}\sum_{i=1}^{n}\big(\phi^*(\hat{\Lambda}^1_{k^*},\hat{g}_n,\hat{\Lambda}^c_n)(O_i)\big)^2,$$

where, by (6.7),

$$\begin{aligned}&\phi^*(\hat{\Lambda}^1_{k^*},\hat{g}_n,\hat{\Lambda}^c_n)(O_i)\\&\quad=\int\mathbb{1}\{t\le t_0\}\hat{H}_{t,k^*}(A_i,L_i)\Big(dN_i(t)-\exp\big(\varepsilon\hat{H}_{t,k^*}(A_i,L_i)\big)d\hat{\Lambda}^1_{k^*}(t\,|\,A_i,L_i)\Big)\\&\qquad+\ \hat{S}_{k^*}(t_0\,|\,1,L_i)-\hat{S}_{k^*}(t_0\,|\,0,L_i)-\hat{\Psi}^*_n.\end{aligned}$$

6.4 Final comments

All our manuscripts consider parameters like $\Psi_{t_0} : \mathcal{M} \to \mathbb{R}$ defined by (6.3) in Section 6.1. Manuscript I uses the TMLE framework to estimate the effect of time-dependent treatment strategies on the survival probability in a continuous-time setting. In Section 6.3 we outlined the case with only baseline confounding: Here, the continuous-time version of TMLE works more or less precisely as the discrete-time version. For the setting considered in Manuscript I, on the other hand, the existing discrete-time versions of TMLE do not apply directly and we propose a new algorithm. Manuscripts III and IV both apply random forest methodology to estimate $\Psi_{t_0}(P_0)$. Manuscripts IV utilizes a weighting approach to target different variations of $\Psi_{t_0}(P_0)$ in the competing risks setting. Manuscripts III identifies the target parameter by inverse probability weighting, but proposes an estimation procedure that incorporates the (efficient) influence function. We discuss random forest methodology in the next chapter.

Recall our motivating problem from Chapter 1. Assume that we have a list of $M \in \mathbb{N}$ binary drug treatment variables $A_1, \ldots, A_M \in \{0, 1\}$, and we want to compare these drugs in terms of their effect on the time T until onset of depression. Optimally, we want to report one number for each treatment that gives us a measure of the effect of that treatment. In Manuscript IV we construct a search algorithm where we use variations over k-specific versions of the target parameter we defined Section 6.1,

$$\Psi_{t_0,k}(P) = P(T^{A_k=1} > t_0) - P(T^{A_k=0} > t_0), \tag{6.13}$$

as a measure of the effect of drug A_k. Clearly, the choice of t_0 may make a difference and estimation of the entire survival curve will give a more complete account of the effect. However, if the number of drugs $M \in \mathbb{N}$ is large, it is not optimal to compare M survival curves. Focusing on (6.13), on the other hand, we have one well-defined parameter for each drug for which we can obtain p-values and confidence intervals.

We consider only effects on the survival scale (or the absolute risk scale), and not on the hazard scale, even if there is no time-dependent confounding. As pointed out by Hernán (2010), there is a generally an unclear basis for causal interpretability of hazard ratios. The problem can be described shortly as follows; for more details we refer to Hernán and Robins (2020, Chapter 17). The hazard at any time t is defined conditional on being alive and event-free by time t. However, being alive and event-free is a post-treatment variable. Suppose the treatment has a large effect on the risk of dying; it kills all subjects in high risk of dying so that survivors in the treated group at time t are all in lower risk of dying than those in the non-treated group. The hazard ratio will then tell us that there is a beneficial effect of treatment. Effects on the absolute risk scale, like (6.3), are not subject to this issue.

Chapter 7.
Random forests

In this chapter we introduce random forests, a machine learning algorithm that uses randomized regression trees to form ensemble predictions. Our motivation for working with random forest is two-fold:

1. A random forest (Breiman, 2001) is a powerful prediction tool for regression, classification and also, as reviewed in Manuscript II, survival analysis and competing risks (Ishwaran et al., 2008, 2014). Thus, random forests are useful, for instance, to enhance super learning (Section 4.2.1).

2. Recent extensions of random forests, the causal forests (Wager and Athey, 2018) and the generalized random forests (Athey et al., 2019), provide a *targeted methodology* where, much like TMLE (Chapter 4), the estimation algorithm is tailored towards a specific target parameter. Manuscripts III and IV are concerned with adaptations of generalized random forests for right-censored data and competing risks.

We start by reviewing the original random forest algorithm (Breiman, 2001); along the way we introduce relevant terminology that we need to present the material of Manuscripts II–IV. Some of the concepts, like sample splitting and cross-validation, were also described in Section 4.2.1 and are common to many machine learning algorithms. We give a short outline of the causal forests and the generalized random forests in Section 7.2 and Section 7.3, respectively.

Throughout this chapter, we switch to the notation that we use in Manuscripts II–IV: The variable A denotes the treatment like we had earlier, but now $X \in \mathcal{X}$ is used to denote the vector of baseline covariates. Following machine learning terminology, we sometimes refer to the data as the training data and to $\mathcal{X}$ as the feature space. In Section 7.1 we consider a prediction problem, but in Sections 7.2–7.3, when introducing work by Wager and Athey (2018); Athey et al. (2019), we move on to treatment effect estimation considering again the setting of our running example from previous chapters. Wager and Athey (2018); Athey et al. (2019) provide some of the first results on the use of random forests for asymptotic statistical inference. We give a summary, but we have not considered the mathematical details in depth.

7.1 The random forest algorithm

We consider data consisting of $n \in \mathbb{N}$ iid observations of a covariate vector $X \in \mathcal{X} \subseteq \mathbb{R}^p$, $p \in \mathbb{N}$, and some outcome $Y \in \mathbb{R}$. A forest consists of $B \in \mathbb{N}$ trees (see Figure 7.1), where, generally, the bth tree defines a partitioning Π_b of the covariate space $\mathcal{X}$ into hyperrectangular cells, with each cell corresponding to a terminal node of the tree. For construction of a tree, various steps of randomization are imposed. First, only a bootstrap sample of the training data is used to grow a tree. Bootstrapping can either be done with or without replacement, and the bootstrap sample can be of size n or smaller. Second, in each step of the tree growing procedure, the covariate space is split along a single axis at a time (the splits are 'axis-aligned'); here, the axis along which to place the split is chosen among a smaller number of randomly pre-selected axis directions based on some splitting criterion. Let $n_{i,b} \geq 0$ denote the number of times unit i appears in the bootstrap sample used in tree b. Note that if bootstrap with resampling is used it may be that $n_{i,b} > 1$, otherwise $n_{i,b} \leq 1$. Consider any $x \in \mathcal{X}$. We let $\mathscr{L}_b(x) \subset \mathcal{X}$ denote the unique region of the covariate space defined by the partitioning Π_b that contains $x \in \mathcal{X}$ and let $\mathscr{N}_b(x) = \sum_{i=1}^n n_{i,b} \mathbb{1}\{X_i \in L_b(x)\}$ denote the number of units (possibly with resampling) falling in this region.[1]

The random forest algorithm has the following hyperparameters:

$B \in \mathbb{N}$: The number of trees.

$s_n \in \{1, \ldots, n\}$: The size of the subsample used to build the trees.

$p_{try} \in \{1, \ldots, p\}$: The number of axis directions pre-selected for splitting in each tree.

$k_n \geq 1$: The minimal number of unique observations in each terminal node.

Generally, the choices of the parameters above are to be made based on considerations of bias-variance trade-off, and, in particular, the choices are important for what we can infer about the asymptotic properties of a random forest estimator. Furthermore, different choices are made for different purposes.

7.1.1 Prediction using forest weights

Breiman's random forest algorithm (Breiman, 2001) forms predictions by averaging predictions across an ensemble of trees. Figure 7.2 illustrates how each tree of a forest starts with a root node and ends up with a set of terminal nodes that are to be used

[1] In the manuscripts we use $L_b(x)$ and $N_b(x)$. We have here switched notation, so that there is no confusion with the notation of counting processes and covariates from earlier chapters in this document.

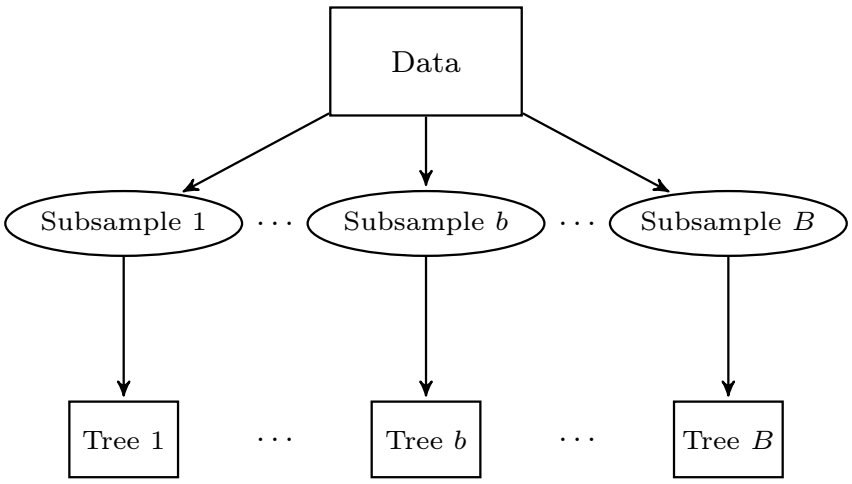


Figure 7.1: The forest is built up by trees that each uses a bootstrap subsample of the training data.

for the prediction. Suppose, for now, that we are interested in estimating,

$$f(x) = \mathbb{E}[Y \mid X = x], \qquad \forall\, x \in \mathcal{X}. \tag{7.1}$$

We distinguish between "inbag" prediction, that is based on all samples in all trees, and "out-of-bag" prediction where estimation is done for training sample j, $j = 1, \ldots, n$, using only the trees where sample j itself was left out of the bootstrap sample. Inbag prediction is used for a new independent sample point $x \in \mathcal{X}$, whereas out-of-bag prediction is used for internal validation. Out-of-bag prediction is based on the same principles as V-fold cross-validation discussed in Section 4.2.1.

Inbag prediction. The predicted value of the forest in a new observation $x \in \mathcal{X}$ may be obtained by collecting all the learning samples (with bootstrap repetition) falling in any of the $b = 1, \ldots, B$ tree terminal nodes containing x. This we can write as,

$$\hat{f}_n(x) = \frac{1}{B} \sum_{b=1}^{B} \frac{\sum_{i=1}^{n} n_{i,b}\, \mathbb{1}\{X_i \in \mathscr{L}_b(x)\}\, Y_i}{\sum_{j=1}^{n} n_{j,b} \mathbb{1}\{X_j \in \mathscr{L}_b(x)\}} = \sum_{i=1}^{n} \alpha_i(x)\, Y_i, \tag{7.2}$$

where we have defined weights $\alpha_i : \mathcal{X} \to [0, 1]$,

$$\alpha_i(x) := \frac{1}{B} \sum_{b=1}^{B} \frac{n_{i,b}\, \mathbb{1}\{X_i \in \mathscr{L}_b(x)\}}{\sum_{b=1}^{B} \sum_{j=1}^{n} n_{j,b}\, \mathbb{1}\{X_j \in \mathscr{L}_b(x)\}} = \frac{1}{B} \sum_{b=1}^{B} \frac{n_{i,b}\, \mathbb{1}\{X_i \in \mathscr{L}_b(x)\}}{\mathscr{N}_b(x)}. \tag{7.3}$$

Out-of-bag prediction. The predicted value for subject j uses only the trees where sample j is out-of-bag, corresponding to all trees b for which $n_{j,b} = 0$. This results in

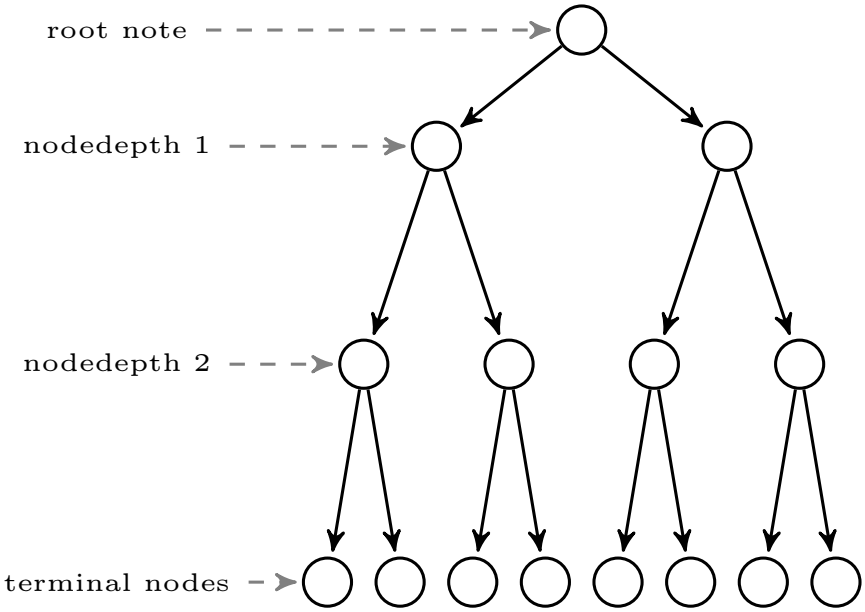


Figure 7.2: Trees sequentially partion the covariate space, starting from the root node and ending in a set of terminal nodes.

out-of-bag weights for prediction, $i \neq j$,

$$\alpha_i^{(-j)}(X_j) := \frac{1}{B}\sum_{b=1}^{B} \mathbb{1}\{n_{j,b} = 0\} \frac{n_{i,b}\,\mathbb{1}\{X_i \in \mathscr{L}_b(x)\}}{\mathscr{N}_b(X_j)}, \tag{7.4}$$

and defines the out-of-bag forest prediction $\hat{f}_n(X_j)$ which is obtained by substituting $\alpha_i^{(-j)}(X_j)$ for $\alpha_i(x)$ in (7.2).

We note how the weights measure the proximity of X_j, or x, to X_i by counting how often X_i and x belong to the same terminal node. We return to this in Section 7.1.3.

7.1.2 Splitting

Trees are grown by sequentially partitioning the covariate space. This is carried out by performing axis-aligned splits guided by a splitting criterion. Any split starts with a mother node $M \subseteq \mathcal{X}$ that is to be split into two daughter nodes D_1 and D_2, as illustrated in Figure 7.3. We consider here the standard CART regression splitting criterion (Breiman et al., 1984) as an example. We let $n_{D_1} = \#\{i \in D_1\}$ and $n_{D_2} = \#\{i \in D_2\}$ be the number of samples falling in the left and the right daughter node, respectively. The standard CART regression splitting criterion is

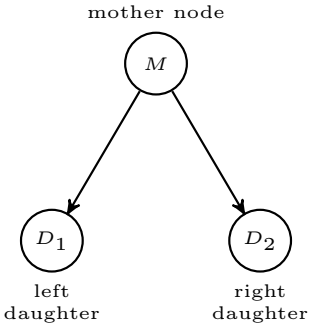


Figure 7.3: A mother node M is split into two daughter nodes D_1 and D_2.

defined as,

$$\mathcal{L}(D_1, D_2) = \sum_{i \in D_1} n_{D_1} \left(Y_i - \frac{1}{n_{D_1}} \sum_{i \in D_1} Y_i \right)^2 + \sum_{i \in D_2} n_{D_2} \left(Y_i - \frac{1}{n_{D_2}} \sum_{i \in D_2} Y_i \right)^2 \tag{7.5}$$

We note that $M = D_2 \cup D_2$ where D_1 and D_2 only differ along a single axis of the covariate space. Any potential split is on the form $(X^j \leq c)$ and $(X^j > c)$ for a value $c \in \mathbb{R}$ where j is some coordinate of X (or, if X^j is categorical, the split will collect certain values of X^j in one node and the rest in the other, since X^j is not necessarily ordered). Only a set of $p_{try} \leq p$ axes are pre-selected from which we can choose to make the split. Then j is chosen from the p_{try} axis directions such as to minimize the splitting criterion in (7.5). Note that, since splits are axis-aligned, we can always write a terminal node $\mathscr{L}_b(x) \in \mathcal{X}$ as,

$$\mathscr{L}_b(x) = \otimes_{q=1}^{p} \mathscr{I}_b(x; q), \tag{7.6}$$

where $\mathscr{I}_b(x; q)$ is some interval of the qth coordinate axis of $\mathcal{X}$.

7.1.3 Random forests as an adaptive nearest neighbor method

Consistency and other asymptotic properties of random forests have been studied and analyzed in various ways (Meinshausen, 2006; Biau, 2012; Scornet et al., 2015; Mentch and Hooker, 2016; Wager and Athey, 2018; Athey et al., 2019). A key aspect of most of the theoretical results is to analyze trees and forests as adaptive nearest neighbors (Lin and Jeon, 2006; Biau and Devroye, 2010; Meinshausen, 2006; Wager and Athey, 2018) utilizing that forests use a weighted average of nearby observations to form predictions. In this section we sketch this idea.

Define the tree-b-subject-i specific weights as,

$$\alpha_{i,b}(x) = \frac{\mathbb{1}\{X_i \in \mathscr{L}_b(x)\}}{\sum_{j=1}^{n} n_{j,b}\mathbb{1}\{X_j \in \mathscr{L}_b(x)\}}. \tag{7.7}$$

Then the (inbag) forest prediction expressed in (7.2) can be written as,

$$\hat{f}_n(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{\mathcal{T}}_b(x), \qquad \hat{\mathcal{T}}_b(x) = \sum_{i=1}^{n} \alpha_{i,b}(x)\, Y_i, \tag{7.8}$$

where $\hat{\mathcal{T}}_b(x)$ is the tree-b specific prediction. Now, consider a tree grown on a bootstrap subsample obtained without resampling. Then $n_{i,b} \in \{0, 1\}$, and the tree weight expression reduces to,

$$\alpha_{b,i}(x) = \frac{\mathbb{1}\{X_i \in \mathscr{L}_b(x)\}}{\sum_{j=1}^{n} \mathbb{1}\{X_j \in \mathscr{L}_b(x)\}}.$$

Recall that a tree defines a rectangular partitioning of $\mathcal{X}$, and that there are a minimal number of k_n samples in each terminal node.

As pointed out by Lin and Jeon (2006), we can relate random forests to so-called *potential nearest neighbors* (PNN). General potential nearest neighbors form predictions by doing a nearest neighbor search over rectangles: A point x' is a k-PNN to x if there exists fewer than k sample points other than x' in the hyperrectangle defined by x' and x (see Lin and Jeon, 2006, Definition 1 and Proposition 1). Particularly, as shown by Lin and Jeon (2006), a tree that makes axis-aligned splits and has terminal node size between k and $2k - 1$, regardless of the splitting scheme used, is a k-PNN predictor. Potential nearest neighbor predictions can generally be written as,

$$\sum_{i=1}^{n} W_i(x) Y_i,$$

where $W_i(x)$ is a variable that gives non-zero weight to the nearest neighbor samples. For tree b we have $W_i(x) = \alpha_{b,i}(x)$ (from (7.7)) which gives non-zero weight to samples falling in the same terminal node as $x \in \mathcal{X}$ and was chosen in a data-adaptive manner based on the splitting criterion. Tree weights give rise to forest weights by $\alpha_i(x) = B^{-1}\sum_{b=1}^{B} \alpha_{i,b}(x)$. Thus, random forests (with predictions on the form (7.8)) can be viewed as an adaptively weighted PNN method. Accordingly, we also refer to the forest weight $\alpha_i(x)$ as a 'neighborhood function'. Figure 7.4 and Figure 7.5 illustrate the tree-based and forest-based nearest neighbor sets for $x \in \mathcal{X}$ where $\mathcal{X} = [0, 1]^2$.

In the next section we see how work by Wager and Athey (2018) utilizes the adaptive nearest neighbor framework to propose forests for causal analysis. To gain some

intuition for this transition, we note that it is a common idea in the causal inference literature to use matching, for instance on the propensity score (Rosenbaum and Rubin, 1983), to control for confounding. Matching requires that we find samples that are "close" enough in the covariate space measured by some distance or proximity measure. A random forest defines a proximity measure in terms of its weighting function $\alpha_i(x)$. To make sense of this proximity measure, one would like to have that the terminal node $\mathscr{L}_b(x) \subset \mathcal{X}$ containing x shrinks around x when $n \to \infty$. Following Meinshausen (2006), one may show that the diameter of the terminal node $\mathscr{L}_b(x)$, defined as,

$$\operatorname{diam}(\mathscr{L}_b(x)) := \max_{q \in \{1,\dots,p\}} \left| \mathscr{I}_b(x; q) \right|,$$

converges to zero in probability under conditions largely as follows: (1) $\mathcal{X} = [0,1]^p$, (2) the proportion of samples in a terminal node is vanishing for $n \to \infty$, (3) the probability that a covariate is chosen for split is bounded away from zero, and (4) there is at least some proportion $\gamma \in (0, 0.5]$ of samples in all terminal nodes. Notably, (3) is needed to ensure that the terminal nodes become small in all directions of the covariate space $\mathcal{X}$; we refer to this as *minimum split probability*. All conditions combined with (5) $f(x)$ is Lipschitz continuous in x, can further be used to establish consistency of the forest prediction.

7.2 Causal forests

Causal forests proposed in Wager and Athey (2018) are based on Breiman's classical random forest algorithm but involves additional conditions on the tree and forest building process. We give a brief summary. The three key conditions that are imposed are: 1) Subsampling is done without replacement, 2) the subsample size s_n scales at a suitable rate relative to the sample size n, and 3) any training sample is only used to either place the splits of the tree or is part of within node estimation. The last part is referred to as *honesty* (Athey and Imbens, 2016). Honesty can be achieved by, for instance, dividing the subsample into two equally large parts, using one to build the tree and the other to do estimation.[2] Then the subsampling and the following splitting is redone over each tree.

We now suppose that the data consist of a treatment indicator $A \in \{0, 1\}$ additional to the covariates $X \in \mathcal{X}$ and the outcome $Y \in \mathbb{R}$. This corresponds to our running example from previous chapters. As in Chapter 2, let Y^1 and Y^0 be the counterfactual variables under treatment ($A = 1$) and no treatment ($A = 0$), respectively. Wager and Athey (2018) consider estimation of the treatment effect conditional on $x \in \mathcal{X}$:

$$\theta(x) = \mathbb{E}[Y^1 \mid X = x] - \mathbb{E}[Y^0 \mid X = x]. \tag{7.9}$$

[2]Similar techniques seem to be used in double/debiased machine learning, see Chernozhukov et al. (2018).

(Notice that we have changed notation from previous chapters, where we used ψ rather than θ). As in Example 1, page 13, we assume no unmeasured confounding (A2) $A \perp\!\!\!\perp (Y^1, Y^0) \,|\, X$, which allows the identification of (7.9) by,

$$\theta(x) = \mathbb{E}\left[\frac{(2A-1)Y}{g(A \,|\, x)} \,\middle|\, X = x \right], \tag{7.10}$$

where $g(a \,|\, X) = P(A = a \,|\, X)$ is the propensity of treatment. Wager and Athey (2018) use the representation in (7.10) without explicitly estimating $g(A \,|\, X)$ and propose a tree estimator for $\theta(x)$ as follows,

$$\mathcal{T}_b(x) = \frac{\sum_{i=1}^n n_{i,b}\, A_i\, \mathbb{1}\{X_i \in \mathscr{L}_b(x)\}\, Y_i}{\sum_{j=1}^n n_{j,b}\, A_j\, \mathbb{1}\{X_j \in \mathscr{L}_b(x)\}} - \frac{\sum_{i=1}^n n_{i,b}\, (1-A_i)\, \mathbb{1}\{X_i \in \mathscr{L}_b(x)\}\, Y_i}{\sum_{j=1}^n n_{j,b}\, (1-A_j)\, \mathbb{1}\{X_j \in \mathscr{L}_b(x)\}}. \tag{7.11}$$

The heuristic argument is that the terminal nodes of the trees are small enough to have "removed confounding effects". By aggregation over the tree prediction (7.11), Wager and Athey (2018) obtain a forest estimator for $\theta(x)$ as,

$$\hat{\theta}_n(x) = B^{-1} \sum_{b=1}^{B} \mathcal{T}_b(x),$$

and they show conditions under which (Wager and Athey, 2018, Theorem 4.1),

$$\frac{\hat{\theta}_n(x) - \theta(x)}{\sqrt{\mathrm{Var}(\hat{\theta}_n(x))}} \xrightarrow{\mathcal{D}} N(0,1). \tag{7.12}$$

They further propose to use the infinitesimal jackknife (Wager et al., 2014) to develop a consistent estimator for the asymptotic variance $\mathrm{Var}(\hat{\theta}_n(x))$. Their proof for (7.12) relies partly on a refinement of the proof by Meinshausen (2006) of $\mathrm{diam}(\mathscr{L}_b(x)) = o_P(1)$ briefly mentioned in Section 7.1.3, and thus requires conditions along the lines of this result. Their conditions include Lipschitz continuity of $\mathbb{E}[Y^a \,|\, X = x]$, $a = 0, 1$, $\mathcal{X} = [0,1]^p$, minimum split probability, a regularity condition on the number of samples in any node (requiring further that any node has at least a certain number of observations from each treatment group), and that the subsample size s_n scales in an appropriate way.

The splitting criterion (Section 7.1.2) is key to understanding the neighborhood function defined by the forest weights. Wager and Athey (2018) propose two different forest procedures using two different splitting schemes:

Double-sample trees. Double-sample trees use Y as outcome and a splitting criterion proposed by Athey and Imbens (2016) that maximizes the variance of $\hat{\theta}_n(x)$.

Propensity trees. Propensity trees use A as outcome and, for instance, a CART splitting criterion.

Double-sample trees produce forest weights $\alpha_i(x)$ that optimize heterogeneity in the target parameter. Propensity trees, on the other hand, produce forest weights that capture similarity in the propensity score. As pointed out by Wager and Athey (2018), the use of propensity trees can be useful in the same problems as propensity score matching (Rosenbaum and Rubin, 1983). Wager and Athey (2018) argue that the double-sample trees and the propensity trees have different strengths, with, for instance, the propensity trees being more robust to confounding. We note that this discussion is similar to that of choosing between a simple inverse probability weighted estimator or the estimator that is based entirely on g-computation. In the next section we introduce generalized random forests that have been proposed as a generalization of causal forests. Generalized random forests are formulated for much more general problems, but also apply to the estimation of conditional treatment effects where they provide a more robust (perhaps even double robust) alternative to double-sample trees and propensity trees.

7.3 Generalized random forests

Generalized random forests (Athey et al., 2019) have been proposed as an extension of the causal forests (and the original random forests), casting forests as a general kernel method with kernels estimated in a data-driven way. The idea is to use the forest weights to solve estimating equations of the form,

$$\mathbb{E}[\psi_{\theta(x),\nu(x)}(O_i) \mid X_i = x] = 0, \quad \forall x \ \in \mathcal{X},$$

where $\theta(x)$ is the parameter of interest and $\nu(x)$ is an optional nuisance parameter. Given forest weights $\alpha_i(x)$ that measure the relevance of the ith sample for the estimation of $\theta(x)$, estimators $\big(\hat{\theta}_n(x), \hat{\nu}_n(x)\big)$ are found as a solution to,

$$\sum_{i=1}^{n} \alpha_i(x)\psi_{\hat{\theta}_n(x),\hat{\nu}_n(x)}(O_i) = 0.$$

Athey et al. (2019) propose that the forest should be constructed such that the adaptive neighborhood function captures heterogeneity in the target $\theta(x)$ specifically. The key is again the splitting scheme of the trees: We should split a mother node into daughter nodes such as to improve estimation of the target parameter $\theta(x)$ in the daughter nodes as much as possible.

It is here nice to recall some concepts from Chapter 3 to relate the work of Athey et al. (2019) to the framework of previous chapters. From Chapter 3, recall that an

estimator $\hat{\theta}_n$ is an asymptotically linear estimator with influence function ϕ if,[3]

$$\hat{\theta}_n - \theta = \frac{1}{n}\sum_{i=1}^{n} \phi(O_i; \theta) + o_P(n^{-1/2}), \tag{7.13}$$

where $\mathbb{E}[\phi(O;\theta)] = 0$ and $\mathbb{E}[\phi(O;\theta)^2] < \infty$. Once the influence function is known, we can define,

$$\psi(O_i, \theta) = \phi(O_i; \theta),$$

such that θ is identified as the solution to the estimating equation,

$$\mathbb{E}[\psi(O, \theta)] = 0.$$

Now consider the split of a mother node $M \subset \mathcal{X}$ into two daughter nodes $M = D_1 \cup D_2$. In the mother node we have an estimator for the target parameter $\hat{\theta}_M$. We further have an estimator $\phi(O_i; \hat{\theta}_M)$ for its influence function. As in Section 7.1.2, let n_{D_1}, n_{D_1} be the number of samples in the first and second daughter node, respectively. Now, instead of estimating the target parameter in the daughter nodes $(\hat{\theta}_{D_1}, \hat{\theta}_{D_2})$ across all possible splits, we can use an asymptotic approximation by the influence function via (7.13):

$$\hat{\theta}_{D_1} - \hat{\theta}_M \approx n_{D_1}^{-1} \sum_{\{i: X_i \in D\}} \phi(O_i; \hat{\theta}_M),$$

and implement splits such as to maximize:

$$\tilde{\mathcal{L}}(D_1, D_2) := \sum_{j=1,2} n_{D_j}^{-1} \bigg(\sum_{\{i: X_i \in D\}} \phi(O_i; \hat{\theta}_M) \bigg)^2. \tag{7.14}$$

This is quite elegant: The influence function $\phi(O_i; \hat{\theta}_M)$ represents the rate of change in the estimator $\hat{\theta}_M$ in direction of O_i. By placing splits that maximize (7.14), we maximize heterogeneity in daughter node estimates $\hat{\theta}_{D_1}$, $\hat{\theta}_{D_2}$ relative to the estimate in the mother node $\hat{\theta}_M$. Further, it is computationally convenient since the implementation can use a standard CART splitting criterion (see (7.5) in Section 7.1.2) applied to $\rho = \phi(O_i; \hat{\theta}_M)$ as a pseudo-outcome and does not have to recompute $\hat{\theta}_{D_1}, \hat{\theta}_{D_2}$ across all possible splits.

7.3.1 Conditional average treatment effects

Athey et al. (2019, Section 6) consider conditional average partial effect estimation. This includes estimation of the conditional average treatment effect (7.9) that we

[3]Note that we have here changed the notation from Chapter 3 slightly.

(re)introduced in Section 7.2 as a special case. In their framework one starts with a random effects model,

$$Y = Ab + \varepsilon.$$

The assumption of no unmeasured confounding is expressed as $(b, \varepsilon) \perp\!\!\!\perp A \,|\, X$. The parameter of interest is identified as,

$$\theta(x) = \mathbb{E}[b \,|\, X = x] = \frac{\operatorname{cov}(Y, A \,|\, X = x)}{\operatorname{Var}(A \,|\, X = x)}, \tag{7.15}$$

and, given forest weights $\alpha_i(x)$, one may estimate $\theta(x)$ by,

$$\hat{\theta}_\alpha(x) = \frac{\sum_{i=1}^n \alpha_i(x)(A_i - \bar{A}_\alpha)^2}{\sum_{i=1}^n \alpha_i(x)(A_i - \bar{A}_\alpha)(Y_i - \bar{Y}_\alpha)},$$

where $\bar{A}_\alpha = \sum_{i=1}^n \alpha_i(x) A_i$ and $\bar{Y}_\alpha = \sum_{i=1}^n \alpha_i(x) Y_i$. Further, a splitting criterion based on pseudo-outcomes is specified as,

$$\rho_i = W_M^{-1}(A_{k,i} - \bar{A}_M)\Big(Y_i - \bar{Y}_M - \big(A_{k,i} - \bar{A}_M\big)\hat{\theta}_M\Big), \tag{7.16}$$

where,

$$W_M = \frac{1}{\#\{i \,:\, \boldsymbol{X}_i \in M\}} \sum_{\{i \,:\, X_i \in M\}} (A_{k,i} - \bar{A}_M)^2,$$

and $\bar{A}_M$, $\bar{Y}_M$ are mother node averages. Accordingly, splits are carried out such as to maximize (7.14) with ρ_i substituted for $\phi(O_i; \hat{\theta}_M)$. If we look more closely,[4] it can be noted that $\theta(x)$ in (7.15) is equal to the conditional treatment effect in (7.9) when $A \in \{0, 1\}$. Further, we see that ρ_i provides a mother node estimator for,

$$\begin{aligned}\phi(O; \theta) = \big(\operatorname{Var}(A_k \,|\, X = x)\big)^{-1}\big(A - g(A \,|\, X = x)\big)\big(Y - \mathbb{E}[Y \,|\, X = x] \\ - \big(A - g(A \,|\, X = x)\big)\theta(x)\big),\end{aligned} \tag{7.17}$$

which can be shown to equal the familiar efficient influence function for our running example (except only including the projection (3.10) on the tangent space for varying q_Y and not q_L). Supposedly, the generalized random forest inherits to some extent the double robustness properties of the efficient influence function. Athey et al. (2019, Section 6.2) present a simulation study where it seems to be the case that causal forests of Wager and Athey (2018) based on double-sample trees and propensity trees, respectively, outperform one another in different scenarios, whereas the generalized random forest works well across all scenarios.

[4] See Supplementary Material of Manuscript IV.

7.3.2 Asymptotic theory

Athey et al. (2019) provide conditions under which they can prove consistency and asymptotic normality of $\hat{\theta}_n(x)$ obtained using generalized random forests (specifically, Athey et al., 2019, Assumptions 1–6, Theorem 3, Theorem 5). Their analysis requires that $X \in \mathcal{X} = [0,1]^p$ and that the density of X is bounded away from zero and infinity. Further conditions can be categorized as follows. First, conditions on the tree construction, second, assumptions on the expected score function $x \mapsto \mathbb{E}[\psi(O) \mid X = x]$ and, third, assumptions on the score function ψ itself. Like Wager and Athey (2018), they require honesty, regularity, minimum split probability, and a specific scaling of the subsample size. They assume that $x \mapsto \mathbb{E}[\psi(O) \mid X = x]$ is Lipschitz continuous and that $(\theta, \nu) \mapsto \mathbb{E}[\psi_{\theta,\nu}(O) \mid X = x]$ is twice continuously differentiable with uniformly bounded second derivative. Lastly, their conditions on the score function ψ include convexity and certain regularity. Athey et al. (2019, Section 6) argue that all required conditions hold for the example in Section 7.3.1, if $\mathbb{E}[Y \mid X = x]$, $\mathbb{E}[A \mid X = x]$ and $\text{cov}(Y, A \mid X = x)$ are Lipschitz in x and if $\text{Var}(A \mid X = x)$ is invertible. Then their Theorem 5 applies, and,

$$\frac{\hat{\theta}(x) - \theta(x)}{\sigma_n(x)} \xrightarrow{\mathcal{D}} N(0,1),$$

for a sequence $\sigma_n(x) = \text{polylog}(n/s_n)^{-1} s_n/n$ (see Athey et al., 2019, Theorem 5). Athey et al. (2019) further derive an estimator for $\sigma_n(x)$ for constructing asymptotically valid confidence intervals for $\theta(x)$ (see Athey et al., 2019, Section 4, Theorem 6).

7.4 Final comments

Random forests are well-established for prediction purposes in survival and competing risks analysis (Ishwaran et al., 2008, 2014). However, current implementations consider estimation of the entire survival curve (or, similarly, the entire cumulative incidence function). We review the different algorithms in Manuscript II.

But what if we are interested in low-dimensional parameters? In the previous chapter we defined a target parameter $\Psi : \mathcal{M} \to \mathbb{R}$ by,

$$\Psi_{t_0}(P) = P(T^1 > t_0) - P(T^0 > t_0),$$

where T^a, for $a = 0, 1$, is the counterfactual event time had there been no censoring and had treatment A been set to a. In Chapter 4 we presented the TMLE methodology as a two-step procedure where the first step involves initial estimation, e.g., using super learning, and the second step is concerned with fluctuation of the initial estimator in an optimal way, targeted towards the specific parameter of interest. In this chapter we have briefly introduced the generalized random forest methodology,

that provides a one-step approach along the same lines. While most machine learning methods for estimation of conditional expectations, density or survival estimation in nonparametric models are used without concern for the target parameter, the splitting in generalized random forests is implemented for optimal estimation of the target parameter. Manuscript III and Manuscript IV are concerned with generalized random forests for targeted estimation of treatment effects on time-to-event outcomes.

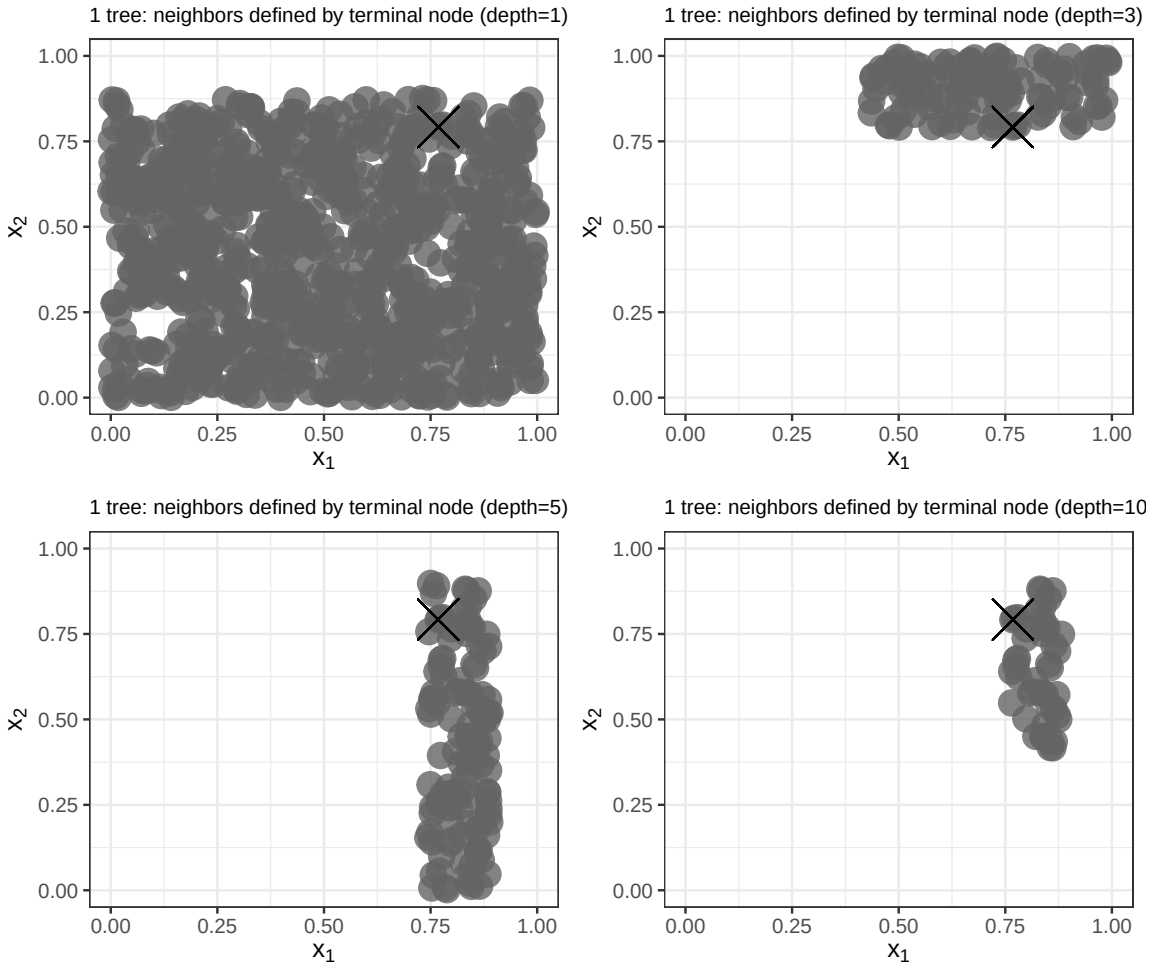


Figure 7.4: Illustration of the tree-based nearest neighbors for $x \in \mathcal{X}$ where $\mathcal{X} = (0,1)^2$. The colored circles show the training samples that fall into the same daughter node as the new test point x (marked by a large cross). The set of neighbors becomes smaller the deeper we grow the trees. **Upper left:** Shown is a tree consisting of a single split (depth 1). **Upper right:** Shown is a tree grown to depth 3 as in Figure 7.2. **Lower left:** Shown is a tree grown to depth 5. **Lower right:** Shown is a tree grown to depth 10.

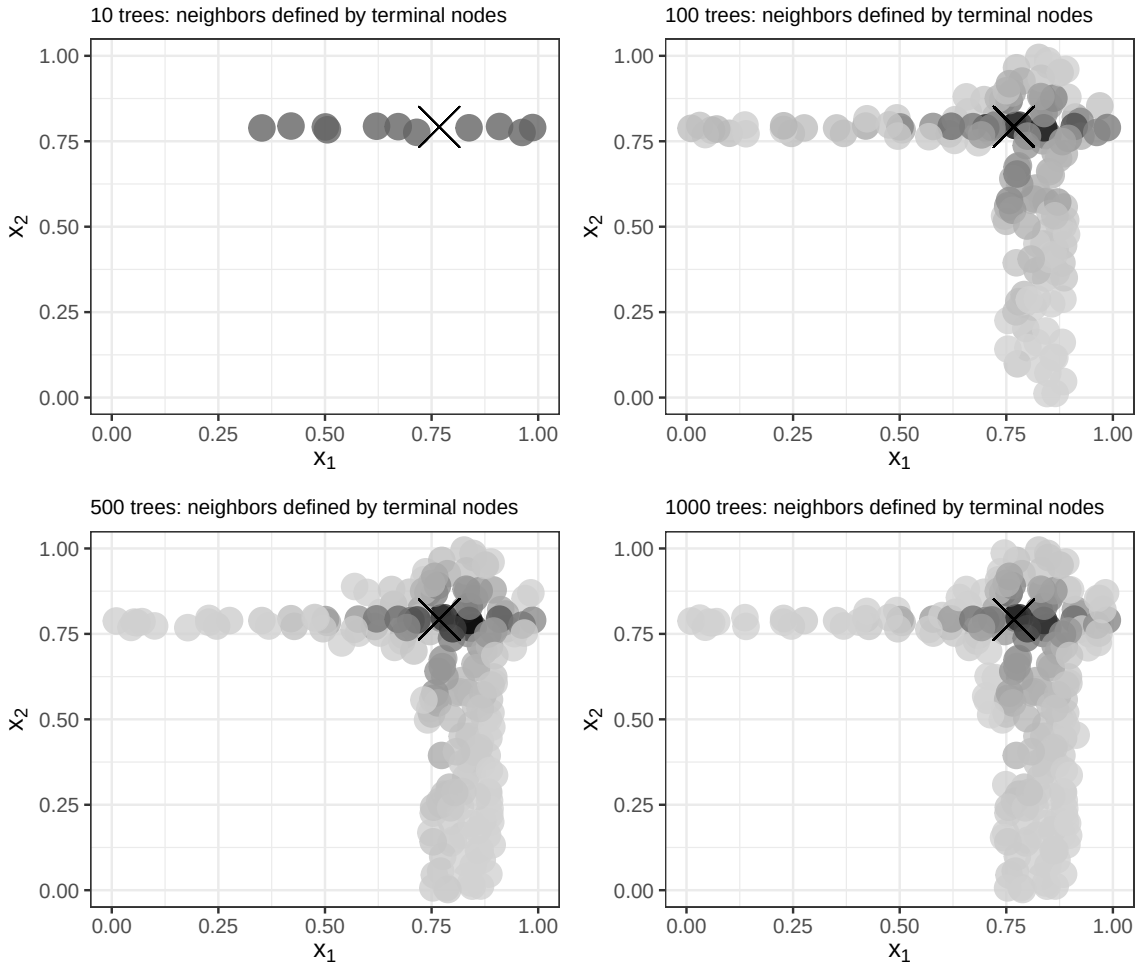


Figure 7.5: Illustration of the forest-based nearest neighbors for $x \in \mathcal{X}$ where $\mathcal{X} = (0, 1)^2$. The colored circles show the weight applied to the training samples for prediction in the new test point x (marked by a large cross): The darker the circle, the closer the weight is to 1. **Upper left:** Shown is a forest consisting of 10 trees. **Upper right:** Shown is a forest consisting of 100 trees. **Lower left:** Shown is a forest consisting of 500 trees. **Lower right:** Shown is a forest consisting of 1000 trees. The more trees we include, the more nuanced the weights become.

Chapter 8.
Summary of manuscripts

In this chapter we summarize the manuscripts of the thesis. As an overview we note that:

Manuscript I is concerned with a generalization of targeted minimum loss based estimation to continuous time.

Manuscript II is concerned with random forests as a prediction strategy for survival analysis.

Manuscript III is concerned with generalized random forest methodology for survival analysis.

Manuscript IV is concerned with an inverse probability weighted approach to the use of generalized random forests for right-censored data and competing risks.

The first section of Chapter 9 provides further summary of the results.

8.1 Manuscript I

Manuscript I builds upon the work of Robins (1986, 1987, 1989); Robins et al. (1992, 2000); Gill and Robins (2001); Bang and Robins (2005); Stitelman et al. (2011); Petersen et al. (2014) and proposes a generalization of TMLE methods from van der Laan (2010a,b); Stitelman and van der Laan (2011); van der Laan and Gruber (2012); Schnitzer et al. (2014); Petersen et al. (2014) to a continuous-time setting. The considered setting is one where changes in both treatment, covariates and outcome can happen on an arbitrarily fine time-scale. The motivation comes from the problem presented in Chapter 1: Subjects of a population are measured on a daily basis over many years and we wish to infer on treatment effects on an outcome of interest while taking into account time-dependent confounding. This setting is quite general, and with increasing accessibility of large-scale observational databases such methods become more and more relevant. As explored in (Sofrygin et al., 2019), the existing (LTMLE) methods are not directly scalable to the data when the time scale gets finer and finer (and at most time-points we do not observe anything), unless a discretization approach is taken instead.

In the manuscript we propose a unified framework based on counting processes (Andersen et al., 1993): By use of product integrals, we can treat both the discrete and

the continuous-time case simultaneously, and we can use intensities to track the individual monitoring of subjects conditional on their observed history. We note that a continuous-time approach is also considered in Lok (2008); Didelez (2008); Røysland (2011, 2012); Commenges and Gégout-Petit (2009); Aalen (1987); Aalen et al. (2012), but not extended to semiparametric efficient estimation.

We define our parameter of interest as the intervention-specific mean outcome at a fixed time-horizon. This incorporates a lot of different special cases, including the survival probability. We analyze the estimation problem and present the efficient influence function which is used to construct a targeting algorithm. We propose an algorithm that can be viewed as a combination of the existing TMLEs for longitudinal data; the one that targets the full likelihood and the one based on the sequential regression representation. It proceeds iteratively, using separate targeting steps for intensities (to deal with random monitoring times) and conditional expectations (to deal with covariate information). The targeted estimator solves the efficient influence curve equation.

We consider also a continuous-time version of Theorem 3.3 and discuss Donsker class conditions and the convergence rate of the second-order remainder. To show that estimation procedures do exist to meet these conditions, we apply the highly adaptive lasso (HAL) methodology to the continuous-time setting. Relying on this, we are able to establish asymptotic linearity and efficiency along the lines of van der Laan (2017). The discussion of HAL in the manuscript is mostly for theoretical purposes and we do not yet provide an implementation for the continuous-time setting.

The manuscript includes a simulation study which serves as a proof-of-concept.

8.2 Manuscript II

In Manuscript II (Rytgaard and Gerds, 2018) we review extensions of the random forest methodology proposed for right-censored data and competing risks. Our focus is mainly on the random survival forest algorithm (Ishwaran et al., 2008, 2014) implemented in `R` in the package `randomForestSRC` (Ishwaran and Kogalur, 2018, 2007a), but there exist other implementations and many other variations of the algorithm specifically for survival analysis (Hothorn et al., 2004, 2005, 2006; Wright and Ziegler, 2017; Wright et al., 2017).

Splitting rules for trees in a survival forest are, for the most part, based on two-sample tests for survival data (LeBlanc and Crowley, 1993). In the setting without competing risks, a common choice is the log-rank test for the null hypothesis of equal survival probabilities in the daughter nodes. Here the split is chosen such as to maximize the log-rank test statistic, and thus the survival difference across the entire time interval,

in the two daughter nodes. For competing risks analysis, one has a choice between a log-rank test of equal cause-specific hazards or the Gray's test of equal cumulative incidence.

Forest ensembles for the cumulative incidence or the survival function are, for instance, obtained by averaging over tree-specific Aalen-Johanson estimators (Aalen and Johansen, 1978) and tree-specific Kaplan-Meier estimators (Kaplan and Meier, 1958), respectively. We here consider the Kaplan-Meier estimator as proposed in Hothorn et al. (2004), where all learning samples that fall in a terminal node with x are collected and used to estimate the survival function,

$$\hat{S}_n(t \mid x) = \prod_{s \leq t} \left(1 - \frac{\sum_{i=1}^{n} \sum_{b=1}^{B} n_{i,b} \mathbb{1}\{X_i \in \mathscr{L}_b(x)\} dN_i^1(s)}{\sum_{i=1}^{n} \sum_{b=1}^{B} n_{i,b} \mathbb{1}\{X_i \in \mathscr{L}_b(x)\} R_i(s)} \right).$$

This is in contrast to Ishwaran et al. (2008) and Ishwaran et al. (2014) who propose ensembles obtained as averages over tree-specific estimators. (Note that we have used the notation from Section 5.1.1 and from Section 7.1).

Manuscript II further reviews prediction accuracy for survival forests (which includes the area under the ROC curve (Heagerty et al., 2000; Chambless and Diao, 2006) and the time-dependent Brier score (Gerds and Schumacher, 2006)) and variable importance measures. Variable importance measures allow the researcher to explore which covariates were important for constructing the forest. The so-called VIMP (Variable IMPortance) of a variable X (Breiman, 2001) targets the prediction error between running the forest with a "noised-up" version of X and running the forest with X as was observed (Liaw et al., 2002; Ishwaran and Kogalur, 2007b). If the prediction performance decreases more for variable X_1 than for variable X_2, then importance(X_1) > importance(X_2). Another measure is the minimal depth (Ishwaran et al., 2010, 2011) that is defined as the distance from the root node of a tree to the first node that is split on X.

8.3 Manuscript III

Manuscript III (Rytgaard, 2019) proposes an adaptation of the framework of generalized random forests (Athey et al., 2019) to right-censored data. In the paper we focus on the target parameter defined as the difference in absolute risk between the two counterfactual scenarios of being treated and not being treated. We provide the causal assumptions needed for identifiability and express our parameter in terms of an inverse probability weighted estimating equation for the target parameter. We propose a kernel weighted estimator for the nuisance parameters, where the kernel weights are the forest weights.

The splitting criterion in the generalized random forest framework is targeted towards the parameter of interest, and splits are implemented by using an approximation based on influence functions. In the paper, we present the influence function that is to be used in the algorithm of a generalized random forest applied to censored data.

The final estimator for the target parameter becomes a forest version of the kernel-weighted Kaplan-Meier estimator (Dabrowska, 1989; Kaplan and Meier, 1958). It is defined as,

$$\hat{\theta}_\alpha(x) = \sum_{a\in\{0,1\}} (2a-1) \int_0^\infty \mathbb{1}\{t > t_0\} \frac{d\hat{H}^1_\alpha(t,a)}{\hat{G}_\alpha(t,a)},$$

using the kernel function $\alpha_i(x)$ defined by the forest, and,

$$\begin{aligned}
\hat{H}^\delta_\alpha(t,a \,|\, x) &= \sum_{i=1}^n K(x,x_i)\, \mathbb{1}\{\tilde{T}_i \le t, \Delta_i = \delta, A_i = a\}, \quad \text{for} \quad \delta = 0,1, \\
\hat{H}_K(t,a \,|\, x) &= \sum_{i=1}^n K(x,x_i)\, \mathbb{1}\{\tilde{T}_i > t, A_i = a\}, \\
\hat{G}_\alpha(t,a \,|\, x) &= \prod_{s\le t} \left(1 - \frac{\hat{H}^0_K(ds,a \,|\, x)}{\hat{H}_K(s,a \,|\, x)}\right).
\end{aligned}$$

We note how the approach of this paper stands in contrast to the random forest methods for survival analysis reviewed in Manuscript II, where splitting rules and prediction focus on the entire survival curve rather than a low-dimensional causal parameter.

8.4 Manuscript IV

Manuscript IV uses generalized random forests (Athey et al., 2019) to construct a causal search algorithm as motivated by the problem presented in Chapter 1. The central idea is to conduct a variable importance analysis that targets causal parameters (van der Laan, 2006) in survival and competing risks settings.

The manuscript discusses two causal parameters in a competing risks setting, where interest is in the treatment effect on occurrence of a particular event type of interest. The manuscript provides the causal assumptions needed for identifiability of both parameters. The two parameters have different interpretations and are defined as follows,

$$\bar{\theta}_1 = P(T^{1,1} \le t_0) - P(T^{1,0} \le t_0), \tag{8.1}$$
$$\bar{\theta}_2 = P(T^1 \le t_0, \Delta^1 = 1) - P(T^0 \le t_0, \Delta^0 = 1), \tag{8.2}$$

We start with the latter, $\bar{\theta}_2$. This is the treatment effect in the world without censoring, with (T^a, Δ^a) being the uncensored counterfactual event time and cause indicator in the hypothetical scenario where treatment is set to $A = a$. Under causal assumptions, $P(T^a \leq t_0, \Delta^a = 1)$ corresponds to the treatment-specific cumulative incidence, c.f. (5.8) in Section 5.1.2. As was briefly noted here, this also depends on the occurrence of competing events. This means that we may have $\bar{\theta}_2 \neq 0$, even if A has no direct effect on the specific event type of interest. In contrast, the parameter $\bar{\theta}_1$ is formulated under further hypothetical reasoning, and corresponds to a treatment effect in a world where competing events have been eliminated. Specifically, the counterfactual variable $T^{1,a}$ is the latent time to the particular event of interest in the hypothetical scenario where treatment is set to $A = a$ and latent time-to-competing-event times are considered right-censoring times. This means that we have $\bar{\theta}_1 = 0$ whenever A has no direct effect on the specific event type of interest.

Identifiability of the parameters $\bar{\theta}_1, \bar{\theta}_2$ relies on a coarsening at random assumption, conditional on the observed history. Let us consider this for $\bar{\theta}_1$: The distribution of the latent times to competing events only depends on the observed history, and not on any unmeasured confounding.

In the manuscript we apply an inverse probability weighting scheme to target estimation of $\bar{\theta}_1$ and $\bar{\theta}_2$, respectively. We investigate the difference between the two through simulation and in a data application. Particularly, we consider the problem from Chapter 1, where we are interested in searching for drugs with an effect on depression; here we do not wish to conclude effects, or lack of effect, of a treatment if that effect was only due to an effect on a competing event. We implement a search algorithm based on the `R` implementation of generalized random forests (Tibshirani et al., 2018) combined with our weighting scheme. We use this to rank a list of drugs according to their protecting effect against depression relapse.

Chapter 9.
Conclusions and Perspectives

In this thesis we have investigated statistical methodology for causal inference based on non-randomized longitudinal data. In particular, we have proposed an extension of the targeted minimum loss based estimation framework to continuous-time data, and we have worked on generalized random forest methodology for survival analysis.

In Chapter 1 we motivated our research work with an applied problem from the Danish registries. We used this problem to illustrate some complications arising when analyzing longitudinal data. What have we gained so far?

- Manuscripts I–IV impose as few restrictions as possible on the data-generating distribution. Rather than a priori having to specify interactions and functional forms in a parametric model, we allow for the use of flexible machine learning tools to learn directly from the data. Thus, model misspecification becomes much less of an issue and our parameters have a sound interpretation, across a very large class of data-generating mechanisms.
- Manuscript I, III and IV are concerned with target parameters defined as an absolute risk (or survival) difference. All parameters are either defined in terms of counterfactual variables or defined directly from the g-computation formula. The parameters can, generally, be ascribed a causal interpretation under a set of assumptions. The questions we are able to address are such as: *What is the expected difference in survival beyond one year had everyone received a particular treatment compared to no one having received this treatment?*
- Manuscript I presents the main theoretical contribution of the thesis, proposing a methodology for targeted minimum loss based estimation of intervention-specific mean outcomes based on a continuous-time counting process model. Under regularity conditions, we demonstrate asymptotic linearity and efficiency under minimal conditions on the statistical model. This lays the groundwork for an improved approach for data-driven analysis of longitudinal data with time-dependent confounding.
- Manuscript III outlines an adaptation of the generalized random forest methodology to a survival analysis setting based on a kernel Kaplan-Meier estimator that uses forest weights as kernels. The forest weights are adaptively estimated in a targeted way to capture heterogeneity specifically in the absolute survival difference.

- Manuscript IV discusses two different causal parameters for the purpose of ranking treatments according to their causal effect on a time-to-event outcome in a competing risks setting. A weighting scheme is proposed to move between different interpretations, and is further used to construct an algorithm based on generalized random forests to search for causal treatment effects.

In the following, we discuss limitations and future perspectives. We start by discussing the two main topics of the thesis (TMLE and random forests) separately, and then we move on to more general issues.

9.1 TMLE in continuous time

Manuscript I presents our ideas for generalizing TMLE to a continuous-time setting. There are still a large number of aspects to consider. One important aspect that requires further discussion is the final causal interpretation of the estimates. In the manuscript, we go directly after the g-computation formula, which, obviously, can be computed in any case. However, if no causal interpretation in the end can be ascribed to the parameter, the method lacks motivation.

The current implementation used in the simulation study only deals with a very simple setting, and the simulation study itself is rather limited. We only consider data simulated on a discrete (although fine) grid to enable comparison with the existing LTMLE. Extensive evaluation on simulated and real data is necessary, before we can apply the methodology to the problem from Chapter 1.

It is our plan to make a general software implementation. For nuisance parameter estimation in the continuous-time case, we can construct a library based on proportional hazards models, Aalen models, or random forests, but a current practical limitation is that software and methods for machine learning estimation of continuous-time intensities are rather limited. In comparison, discrete-time intensity estimators can be constructed based on binary regression, for which there exists a vast amount of methods. To enhance the library for continuous-time intensities, we may consider including also discrete-time intensity estimators using different discretizations (we can even estimate an optimal grid size by cross-validation). In Manuscript I, we consider highly adaptive lasso estimation, but the treatment here is mainly theoretical. Future implementation of the highly adaptive lasso for continuous-time intensities should be added as an option in the existing HAL software.

In further regard to the implementation, we note that we in this manuscript use a full likelihood approach for initial estimation. This means that we also have to estimate the conditional density of covariates given the history of the observed data. We are working on ways to avoid this, just like LTMLE avoids it by exploiting the sequential regression representation of the target parameter. The fact is that our targeting

algorithm does not depend on this conditional density, only through conditional expectations. In future work we will address in detail how to construct estimators for the conditional expectations directly.

To deal with potential scalability issues it may be of interest to go into the online learning methods for targeting learning, see van der Laan and Lendle (2014); Benkeser et al. (2018) and also van der Laan and Rose (2018, Part V). In online learning we have "old" and "new" data; we have performed estimation on the old data and we want to know how inference changes when also taking into account the new data. Instead of recomputing the estimator using all data (old and new), an online learner uses the new data to *update* the estimator based on the old data only. For longitudinal data (that may have a long time-horizon), we can collect "old" data up to a certain time-point $\tau' \in (0, \tau)$ and compute the estimator here. Then we can update that estimator using "new" data after τ'.

In Manuscript I, we argue that our methods work for general treatment regimes, both static, dynamic and stochastic. Nonetheless, we do only consider very simple treatment regimes (a static intervention on treatment decisions). For an example of a relevant stochastic intervention, refer back to our data application from the Chapter 1 where we want to compare treatment with different drugs to one another. Consider the case where one drug only comes on the market at a later point $\tau' \in (0, \tau)$ in calendar time. Now consider an intervention on the timing of treatment decisions represented by the intensity λ^a: We may proceed as in Ryalen et al. (2018) and define $\lambda^{a,*}$ that corresponds to treatment monitoring being randomized. To ensure positivity one could further specify only a non-zero monitoring for $t \in (\tau', \tau)$. Such an intervention might make our treatment effect estimates more relevant for the applied research.

In the manuscript we demonstrate that the efficient influence function admits a double robust structure, and we argue for the use of data-adaptive algorithms for estimation of the nuisance parameters. We remark upon a general issue in this regard. As pointed out by Benkeser et al. (2017), and also studied by van der Laan (2014), *doubly robust inference* is not guaranteed. Specifically, as argued by Benkeser et al. (2017), when nuisance parameters are estimated based on data-adaptive methods, the resulting estimator for the target parameter may be biased and further not asymptotically linear if the nuisance parameters are inconsistently estimated. This is in contrast to parametric models that preserve asymptotically linearity, even under misspecification.We have not addressed this issue in our work.

9.2 Random forests for survival analysis

We plan to continue working with the generalized random forest methodology for survival analysis, both with the implementation (implementing the splitting scheme

of Manuscript III in the `R` package `grf` (Tibshirani et al., 2018)), and also the theory. In our manuscripts, we do not delve deeply into the theoretical work by Athey et al. (2019); Wager and Athey (2018) but focus on the applicability. It remains for us to fully understand and exploit the asymptotic theory of Athey et al. (2019) as was briefly summarized in Section 7.3.2.

It is generally not clear how we should identify effects on the occurrence of a particular event of interest in presence of competing risks: Either we target the cumulative incidence or the cause-specific hazard. In Manuscript IV we propose a weighting approach to target parameters with different interpretations, where one is the treatment effect in a hypothetical world with no competing risks. Recent work by Stensrud et al. (2019) propose a different approach borrowing ideas from mediation analysis (Robins and Greenland, 1992; Pearl, 2001): Positing that the treatment of interest can be divided into two active parts, one with an effect on the outcome of interest and one with an effect on the competing event, we can split the effect of the drug between a direct effect on the outcome of interest and an indirect effect through the competing event. This may be an alternative approach worth pursuing for handling competing risks in the future, as it defines treatment effects on the event type of interest without referring to a hypothetical scenario where the competing risk has been eliminated.

There is a general issue with the weighting approach used in Manuscript IV that relies on consistent estimation of the censoring and the competing event distributions. In the manuscript, we propose weights based on the Kaplan-Meier estimator which is unbiased for applications with completely independent censoring. However, when weights are needed for an effect in the hypothetical world with competing risks eliminated, a Kaplan-Meier estimator is likely biased. Furthermore, even when censoring is independent of covariates, one can improve efficiency by incorporating covariate information in the weight estimation. One idea for improvement would be to use a separate random survival forest (Ishwaran et al., 2008) to estimate the inverse probability weights in a flexible way that uses all covariates. Such an approach would require a thorough analysis of the asymptotic behavior of the proposed two-step approach. An alternative idea is the one-step approach of Manuscript III.

To further improve the generalized random forest estimators, perhaps the work of Hubbard et al. (2000), or even a targeting step, may be relevant: The estimator proposed by Hubbard et al. (2000) improves upon the inverse probability weighted estimator by adding the sample mean of the estimated efficient influence curve. This results in a double robust estimator. In a similar manner, one could incorporate a targeting step for the nuisance parameters in the forest procedure; it would be interesting to see if one could achieve double robustness gains by implementing such a targeting step.

Our work so far deals with random forests only for survival analysis with baseline covariates and baseline treatment. It is very much of interest to incorporate time-dependent covariates, both for the existing survival forests implementations as summarized in Manuscript II and for the generalized random forests. Trees and forests for time-varying covariates have been proposed in different variants (Bacchetti and Segal, 1995; Xu and Adak, 2002; Bou-Hamad et al., 2011; Wallace, 2014; Fu and Simonoff, 2016; Bertolet et al., 2016), but the applicability and interpretability in settings with time-dependent confounding should be investigated further.

9.3 Other topics

Estimation of the entire survival curve. Throughout this thesis, we have focused on estimation and efficiency theory for univariate parameters. The argument is that we are only interested in a single effect measure for each treatment (as, for example, in our variable importance analysis in Manuscript IV, and in general for our motivating problem from Chapter 1). But what if we are, in fact, interested in the entire survival curve? If we simply apply our methods for each time-point encountered in the dataset, the result is possibly a non-monotone survival function. However, recent advances in targeted learning have addressed exactly this issue and propose a so-called one-step TMLE for estimation of the entire survival curve (see Cai and van der Laan (2019), which is based on work by van der Laan and Gruber (2016)). In regard to the random forest methodology, we propose in Manuscripts III and IV forest estimation of the counterfactual survival and absolute risk difference at a single time-point t_0; as reviewed in Manuscript II, existing forest implementations target the entire survival curve by using a log-rank splitting rule, but this is with no considerations on causal interpretability.

Heterogeneous treatment effects. We have focused on average treatment effects. But what if there is an effect only for some individuals and not others? In that case we say that there is a heterogeneous effect, or an individualized effect. Various methods based on recursive partitioning such as Athey and Imbens (2016); Seibold et al. (2019) have been proposed for data-adaptively discovering groups of individuals that differ in terms of their treatment effects. The work by Athey et al. (2019); Wager and Athey (2018) target heterogeneous causal effects as well. It would further be interesting to extend our methods to estimation of optimal dynamic treatment regimes (Zhang et al., 2013, 2012; Hernán et al., 2006; Murphy, 2003; Bai et al., 2017), defined, for instance, as the dynamic regime that minimizes the absolute risk under the dynamic rule (see also van der Laan and Rose, 2018, Chapter 22).

Practical limitations. It should be noted that we have disregarded many practical challenges. Some are general problems of causal inference. We assumed a given causal direction in the observed variables determined by the time-order; we have

not discussed, for instance, reverse causality. Another common problem in causal inference arises from near positivity violations. This may be handled by considering more stable dynamic interventions (van der Laan and Petersen, 2007). Moreover, assumptions such as no unmeasured confounding pose a general problem for our methods. Is it reasonable to think that we observe everything that predicts both treatment and outcome? We need to "borrow" a lot of information between subjects which requires a lot from the observed covariates. Instrumental variables (see, e.g., Angrist et al., 1996; Hernán and Robins, 2006) provide an approach to relax some of these assumptions. For the problem discussed in Chapter 1 we could use calendar time as an instrument, which would require that calendar time is only a predictor for drug intake and not for diagnosis with depression. Generalized random forests are implemented for instrumental variable analysis (see Athey et al., 2019, Section 7), which must be adapted to right-censored data and competing risks. Additional to these more general problems, there are many other practical aspects specifically for the motivating problem in Chapter 1 that were not discussed here. For example, we assume that drugs purchased were in fact taken; the actual treatment regime followed remains unobserved. Moreover, when we compare various drugs, we should take into account that different drugs work very differently; some drugs have an instant effect, others a lagged effect. Some of these problems have been addressed by, e.g., Nielsen et al. (2008, 2009).

Bibliography

Aalen, O. O. (1987). Dynamic modelling and causality. *Scandinavian Actuarial Journal 1987*(3-4), 177–190.

Aalen, O. O. and S. Johansen (1978). An empirical transition matrix for non-homogeneous markov chains based on censored observations. *Scandinavian Journal of Statistics 5*, 141–150.

Aalen, O. O., K. Røysland, J. M. Gran, and B. Ledergerber (2012). Causality, mediation and time: a dynamic viewpoint. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 175*(4), 831–861.

Andersen, P. K., O. Borgan, R. D. Gill, and N. Keiding (1993). *Statistical models based on counting processes*. Springer, New York.

Angrist, J. D., G. W. Imbens, and D. B. Rubin (1996). Identification of causal effects using instrumental variables. *Journal of the American statistical Association 91*(434), 444–455.

Athey, S. and G. Imbens (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences 113*(27), 7353–7360.

Athey, S., J. Tibshirani, and S. Wager (2019). Generalized random forests. *The Annals of Statistics 47*(2), 1148–1178.

Bacchetti, P. and M. R. Segal (1995). Survival trees with time-dependent covariates: application to estimating changes in the incubation period of aids. *Lifetime data analysis 1*(1), 35–47.

Bai, X., A. A. Tsiatis, W. Lu, and R. Song (2017). Optimal treatment regimes for survival endpoints using a locally-efficient doubly-robust estimator from a classification perspective. *Lifetime data analysis 23*(4), 585–604.

Bang, H. and J. M. Robins (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics 61*(4), 962–973.

Benkeser, D. (2015). *Data-adaptive Estimation in Longitudinal Data Structures with Applications in Vaccine Efficacy Trials*. Ph. D. thesis, University of Washington.

Benkeser, D., M. Carone, M. J. van der Laan, and P. B. Gilbert (2017). Doubly robust nonparametric inference on the average treatment effect. *Biometrika 104*(4), 863–880.

Benkeser, D., C. Ju, S. Lendle, and M. J. van der Laan (2018). Online cross-validation-based ensemble learning. *Statistics in medicine 37*(2), 249–260.

Benkeser, D. and M. van der Laan (2016). The highly adaptive lasso estimator. In *Proceedings of the... International Conference on Data Science and Advanced Analytics. IEEE International Conference on Data Science and Advanced Analytics*, Volume 2016, pp. 689. NIH Public Access.

Bertolet, M., M. M. Brooks, and V. Bittner (2016). Tree-based identification of subgroups for time-varying covariate survival data. *Statistical methods in medical research 25*(1), 488–501.

Biau, G. (2012). Analysis of a random forests model. *Journal of Machine Learning Research 13*(Apr), 1063–1095.

Biau, G. and L. Devroye (2010). On the layered nearest neighbour estimate, the bagged nearest neighbour estimate and the random forest method in regression and classification. *Journal of Multivariate Analysis 101*(10), 2499–2518.

Bickel, P. J., C. A. J. Klaassen, Y. Ritov, and J. A. Wellner (1993). *Efficient and adaptive inference in semiparametric models*. Johns Hopkins University Press, Baltimore.

Bou-Hamad, I., D. Larocque, H. Ben-Ameur, et al. (2011). A review of survival trees. *Statistics Surveys 5*, 44–71.

Breiman, L. (2001). Random forests. *Machine learning 45*(1), 5–32.

Breiman, L., J. Friedman, R. Olshen, and C. Stone (1984). Classification and regression trees. wadsworth int. *Group 37*(15), 237–251.

Breslow, N. (1974). Covariance analysis of censored survival data. *Biometrics 30*, 89–99.

Cai, W. and M. J. van der Laan (2019). One-step targeted maximum likelihood estimation for time-to-event outcomes. *Biometrics*.

Chakraborty, B. and E. E. Moodie (2013). *Statistical methods for dynamic treatment regimes*. Springer.

Chambaz, A. and M. J. van der Laan (2014). Inference in targeted group-sequential covariate-adjusted randomized clinical trials. *Scandinavian Journal of Statistics 41*(1), 104–140.

Chambless, L. E. and G. Diao (2006). Estimation of time-dependent area under the ROC curve for long-term risk prediction. *Statistics in medicine 25*(20), 3474–3486.

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

Commenges, D. and A. Gégout-Petit (2009). A general dynamical statistical model with causal interpretation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 71*(3), 719–736.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological) 34*(2), 187–202.

Dabrowska, D. M. (1989). Uniform consistency of the kernel conditional kaplan-meier estimate. *The Annals of Statistics 17*, 1157–1167.

Daniel, R. M., S. N. Cousens, B. L. De Stavola, M. G. Kenward, and J. A. C. Sterne (2013). Methods for dealing with time-dependent confounding. *Statistics in medicine 32*(9), 1584–1618.

Dawid, A. P. and V. Didelez (2010). Identifying the consequences of dynamic treatment strategies: A decision-theoretic overview. *Statistics Surveys 4*, 184–231.

Didelez, V. (2008). Graphical models for marked point processes based on local independence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70*(1), 245–264.

Fu, W. and J. S. Simonoff (2016). Survival trees for left-truncated and right-censored data, with application to time-varying covariate data. *Biostatistics 18*(2), 352–369.

Gerds, T. and M. Schumacher (2006). Consistent estimation of the expected Brier score in general survival models with right-censored event times. *Biometrical Journal 48*(6), 1029–1040.

Gill, R. D. and J. M. Robins (2001). Causal inference for complex longitudinal data: the continuous case. *Annals of Statistics 29*, 1–27.

Gill, R. D., M. J. van der Laan, and J. M. Robins (1997). Coarsening at random: Characterizations, conjectures, counter-examples. In *Proceedings of the First Seattle Symposium in Biostatistics*, pp. 255–294. Springer.

Gill, R. D., M. J. van der Laan, and J. A. Wellner (1995). *Inefficient estimators of the bivariate survival function for three models*, Volume 31. Annales de l'Institut Henri Poincaré.

Grøn, R., T. A. Gerds, and P. K. Andersen (2016). Misspecified poisson regression models for large-scale registry data: inference for 'large n and small p'. *Statistics in medicine 35*(7), 1117–1129.

Heagerty, P. J., T. Lumley, and M. S. Pepe (2000). Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics 56*(2), 337–344.

Heitjan, D. F. and D. B. Rubin (1991). Ignorability and coarse data. *The annals of statistics 19*, 2244–2253.

Hernán, M. A. (2010). The hazards of hazard ratios. *Epidemiology (Cambridge, Mass.) 21*(1), 13.

Hernán, M. A., E. Lanoy, D. Costagliola, and J. M. Robins (2006). Comparison of dynamic treatment regimes via inverse probability weighting. *Basic & clinical pharmacology & toxicology 98*(3), 237–242.

Hernán, M. A. and J. M. Robins (2006). Instruments for causal inference: an epidemiologist's dream? *Epidemiology 17*, 360–372.

Hernán, M. A. and J. M. Robins (2020). *Causal Inference*. Boca Raton, Fl: Chapman & Hall/CRC.

Hothorn, T., P. Bühlmann, S. Dudoit, A. Molinaro, and M. J. van der Laan (2005). Survival ensembles. *Biostatistics 7*(3), 355–373.

Hothorn, T., K. Hornik, and A. Zeileis (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical statistics 15*(3), 651–674.

Hothorn, T., B. Lausen, A. Benner, and M. Radespiel-Tröger (2004). Bagging survival trees. *Statistics in medicine 23*(1), 77–91.

Hubbard, A. E., M. J. van Der Laan, and J. M. Robins (2000). Nonparametric locally efficient estimation of the treatment specific survival distribution with right censored data and covariates in observational studies. In *Statistical Models in Epidemiology, the Environment, and Clinical Trials*, pp. 135–177. Springer.

Ishwaran, H., T. A. Gerds, U. B. Kogalur, R. D. Moore, S. J. Gange, and B. M. Lau (2014). Random survival forests for competing risks. *Biostatistics 15*(4), 757–773.

Ishwaran, H. and U. B. Kogalur (2007a, October). Random survival forests for r. *R News 7*(2), 25–31.

Ishwaran, H. and U. B. Kogalur (2007b). Random survival forests for r. *R news 7*(2), 25–31.

Ishwaran, H. and U. B. Kogalur (2018). *Random Forests for Survival, Regression, and Classification (RF-SRC)*. R package version 2.7.0.

Ishwaran, H., U. B. Kogalur, E. H. Blackstone, and M. S. Lauer (2008). Random survival forests. *The annals of applied statistics 2*(3), 841–860.

Ishwaran, H., U. B. Kogalur, X. Chen, and A. J. Minn (2011). Random survival forests for high-dimensional data. *Statistical analysis and data mining 4*(1), 115–132.

Ishwaran, H., U. B. Kogalur, E. Z. Gorodeski, A. J. Minn, and M. S. Lauer (2010). High-dimensional variable selection for survival data. *Journal of the American Statistical Association 105*(489), 205–217.

Jacobsen, M. and N. Keiding (1995). Coarsening at random in general sample spaces and random censoring in continuous time. *The Annals of Statistics 23*(3), 774–786.

Kaplan, E. L. and P. Meier (1958). Nonparametric estimation from incomplete observations. *Journal of the American statistical association 53*(282), 457–481.

Kessing, L. V., H. C. Rytgaard, T. A. Gerds, M. Berk, C. T. Ekstrøm, and P. K. Andersen (2019a). New drug candidates for bipolar disorder – a nation-wide population-based study. *Bipolar disorders 21*, 410–418.

Kessing, L. V., H. C. Rytgaard, T. A. Gerds, M. Berk, C. T. Ekstrøm, and P. K. Andersen (2019b). New drug candidates for depression – a nationwide population-based study. *Acta Psychiatrica Scandinavica 139*(1), 68–77.

LeBlanc, M. and J. Crowley (1993). Survival trees by goodness of split. *Journal of the American Statistical Association 88*(422), 457–467.

Lendle, S. D., J. Schwab, M. L. Petersen, and M. J. van der Laan (2017). ltmle: an r package implementing targeted minimum loss-based estimation for longitudinal data. *Journal of Statistical Software 81*(1), 1–21.

Liaw, A., M. Wiener, et al. (2002). Classification and regression by randomforest. *R news 2*(3), 18–22.

Lin, Y. and Y. Jeon (2006). Random forests and adaptive nearest neighbors. *Journal of the American Statistical Association 101*(474), 578–590.

Lok, J. J. (2008). Statistical modeling of causal effects in continuous time. *The Annals of Statistics 36*(3), 1464–1507.

Meinshausen, N. (2006). Quantile regression forests. *Journal of Machine Learning Research 7*(Jun), 983–999.

Mentch, L. and G. Hooker (2016). Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. *The Journal of Machine Learning Research 17*(1), 841–881.

Moore, K. L. and M. J. van der Laan (2009a). Application of time-to-event methods in the assessment of safety in clinical trials. *Design and Analysis of Clinical Trials with Time-to-Event Endpoints. Taylor & Francis*, 455–482.

Moore, K. L. and M. J. van der Laan (2009b). Covariate adjustment in randomized trials with binary outcomes: targeted maximum likelihood estimation. *Statistics in medicine 28*(1), 39–64.

Moore, K. L. and M. J. van der Laan (2009c). Increasing power in randomized trials with right censored outcomes through covariate adjustment. *Journal of biopharmaceutical statistics 19*(6), 1099–1131.

Muñoz, I. D. and M. van der Laan (2012). Population intervention causal effects based on stochastic interventions. *Biometrics 68*(2), 541–549.

Murphy, S. A. (2003). Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 65*(2), 331–355.

Neyman, J. (1923). Sur les applications de la theorie des probabilites aux experiences agricoles: Essai des principes (in polish). english translation by dm dabrowska and tp speed (1990). *Statistical Science 5*, 465–480.

Nielsen, L. H., E. Løkkegaard, A. H. Andreasen, Y. A. Hundrup, and N. Keiding (2009). Estimating the effect of current, previous and never use of drugs in studies based on prescription registries. *Pharmacoepidemiology and drug safety 18*(2), 147–153.

Nielsen, L. H., E. Løkkegaard, A. H. Andreasen, and N. Keiding (2008). Using prescription registries to define continuous drug use: how to fill gaps between prescriptions. *Pharmacoepidemiology and drug safety 17*(4), 384–388.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika 82*(4), 669–688.

Pearl, J. (2001). Direct and indirect effects. In *Proceedings of the seventeenth conference on uncertainty in artificial intelligence*, pp. 411–420. Morgan Kaufmann Publishers Inc.

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

Petersen, M., J. Schwab, S. Gruber, N. Blaser, M. Schomaker, and M. van der Laan (2014). Targeted maximum likelihood estimation for dynamic and static longitudinal marginal structural working models. *Journal of causal inference 2*(2), 147–185.

Polley, E. C., S. Rose, and M. J. van der Laan (2011). Super learning. In *Targeted Learning*, pp. 43–66. Springer.

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.

Robins, J. M. (1987). Addendum to "a new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect". *Computers & Mathematics with Applications 14*(9-12), 923–945.

Robins, J. M. (1989). The analysis of randomized and non-randomized aids treatment trials using a new approach to causal inference in longitudinal studies. *Health service research methodology: a focus on AIDS*, 113–159.

Robins, J. M. (2000). Robust estimation in sequentially ignorable missing data and causal inference models. In *Proceedings of the American Statistical Association*, Volume 1999, pp. 6–10.

Robins, J. M., D. Blevins, G. Ritter, and M. Wulfsohn (1992). G-estimation of the effect of prophylaxis therapy for pneumocystis carinii pneumonia on the survival of aids patients. *Epidemiology*, 319–336.

Robins, J. M. and S. Greenland (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, 143–155.

Robins, J. M. and M. A. Hernán (2008). Estimation of the causal effects of time-varying exposures. In *Longitudinal data analysis*, pp. 547–593. Chapman and Hall/CRC.

Robins, J. M., M. A. Hernan, and B. Brumback (2000). Marginal structural models and causal inference in epidemiology.

Robins, J. M., M. A. Hernán, and U. Siebert (2004). Effects of multiple interventions. *Comparative quantification of health risks: global and regional burden of disease attributable to selected major risk factors 1*, 2191–2230.

Robins, J. M. and A. Rotnitzky (1992). Recovery of information and adjustment for dependent censoring using surrogate markers. In *AIDS epidemiology*, pp. 297–331. Springer.

Robins, J. M., A. Rotnitzky, and D. O. Scharfstein (2000). Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In *Statistical models in epidemiology, the environment, and clinical trials*, pp. 1–94. Springer.

Rosenbaum, P. R. and D. B. Rubin (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika 70*(1), 41–55.

Røysland, K. (2011). A martingale approach to continuous-time marginal structural models. *Bernoulli 17*(3), 895–915.

Røysland, K. (2012). Counterfactual analyses with graphical models based on local independence. *The Annals of Statistics 40*(4), 2162–2194.

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

Ryalen, P. C., M. J. Stensrud, S. Fosså, and K. Røysland (2018). Causal inference in continuous time: an example on prostate cancer therapy. *Biostatistics*.

Rytgaard, H. C. (2019). Application of generalized random forests for survival analysis. In *European Young Statisticians Meeting*, pp. 102.

Rytgaard, H. C. and T. A. Gerds (2018). Random forests for survival analysis. *Wiley StatsRef: Statistics Reference Online*, 1–8.

Schnitzer, M. E., E. E. M. Moodie, M. J. van der Laan, R. W. Platt, and M. B. Klein (2014). Modeling the impact of hepatitis c viral clearance on end-stage liver disease in an hiv co-infected cohort with targeted maximum likelihood estimation. *Biometrics 70*(1), 144–152.

Scornet, E., G. Biau, and J. Vert (2015). Consistency of random forests. *The Annals of Statistics 43*(4), 1716–1741.

Seibold, H., A. Zeileis, and T. Hothorn (2019). model4you: An r package for personalised treatment effect estimation. *Journal of Open Research Software 7*(1).

Sofrygin, O. and M. J. van der Laan (2017). Semi-parametric estimation and inference for the mean outcome of the single time-point intervention in a causally connected population. *Journal of causal inference 5*(1).

Sofrygin, O., Z. Zhu, J. A. Schmittdiel, A. S. Adams, R. W. Grant, M. J. van der Laan, and R. Neugebauer (2019). Targeted learning with daily ehr data. *Statistics in medicine 38*(16), 3073–3090.

Spirtes, P., C. N. Glymour, R. Scheines, D. Heckerman, C. Meek, G. Cooper, and T. Richardson (2000). *Causation, prediction, and search*. MIT press.

Stensrud, M. J., J. G. Young, V. Didelez, J. M. Robins, and M. A. Hernán (2019). Separable effects for causal inference in the presence of competing risks. *arXiv preprint arXiv:1901.09472*.

Stitelman, O. M., V. De Gruttola, C. W. Wester, and M. J. van der Laan (2011). Rcts with time-to-event outcomes and effect modification parameters. In *Targeted Learning*, pp. 271–298. Springer.

Stitelman, O. M. and M. J. van der Laan (2011). Targeted maximum likelihood estimation of time-to-event parameters with time-dependent covariates. Technical report, Technical Report, Division of Biostatistics, University of California, Berkeley.

Stitelman, O. M., C. W. Wester, V. De Gruttola, and M. J. van der Laan (2011). Targeted maximum likelihood estimation of effect modification parameters in survival analysis. *The international journal of biostatistics 7*(1), 1–34.

Tibshirani, J., S. Athey, S. Wager, R. Friedberg, L. Miner, and M. Wright (2018). grf: Generalized random forests. *R package*.

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

Tsiatis, A. (2007). *Semiparametric theory and missing data*. Springer Science & Business Media.

van der Laan, M. and S. Gruber (2016). One-step targeted minimum loss-based estimation based on universal least favorable one-dimensional submodels. *The international journal of biostatistics 12*(1), 351–378.

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

van der Laan, M. J. (2010a). Targeted maximum likelihood based causal inference: Part I. *The International Journal of Biostatistics 6*(2).

van der Laan, M. J. (2010b). Targeted maximum likelihood based causal inference: Part II. *The international journal of biostatistics 6*(2).

van der Laan, M. J. (2014). Targeted estimation of nuisance parameters to obtain valid statistical inference. *The international journal of biostatistics 10*(1), 29–57.

van der Laan, M. J. (2015). A generally efficient targeted minimum loss based estimator.

van der Laan, M. J. (2017). A generally efficient targeted minimum loss based estimator based on the highly adaptive lasso. *The International Journal of Biostatistics 13*(2).

van der Laan, M. J., A. Chambaz, and S. Lendle (2018). Online targeted learning for time series. In *Targeted Learning in Data Science*, pp. 317–346. Springer.

van der Laan, M. J. and S. Dudoit (2003). Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator: Finite sample oracle inequalities and examples.

van der Laan, M. J., S. Dudoit, and A. W. van der Vaart (2006). The cross-validated adaptive epsilon-net estimator. *Statistics & Decisions 24*(3), 373–395.

van der Laan, M. J. and S. Gruber (2012). Targeted minimum loss based estimation of causal effects of multiple time point interventions. *The international journal of biostatistics 8*(1).

van der Laan, M. J. and S. D. Lendle (2014). Online targeted learning.

van der Laan, M. J. and A. R. Luedtke (2014). Targeted learning of an optimal dynamic treatment, and statistical inference for its mean outcome.

van der Laan, M. J. and M. L. Petersen (2007). Causal effect models for realistic individualized treatment and intention to treat rules. *The international journal of biostatistics 3*(1).

van der Laan, M. J., E. C. Polley, and A. E. Hubbard (2007). Super learner. *Statistical applications in genetics and molecular biology 6*(1).

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

van der Laan, M. J. and S. Rose (2011). *Targeted learning: Causal inference for observational and experimental data*. Springer Science & Business Media.

van der Laan, M. J. and S. Rose (2018). *Targeted learning in data science: causal inference for complex longitudinal studies*. Springer.

van der Laan, M. J. and D. Rubin (2006). Targeted maximum likelihood learning. *The International Journal of Biostatistics 2*(1).

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

van der Vaart, A. W., S. Dudoit, and M. J. van der Laan (2006). Oracle inequalities for multi-fold cross validation. *Statistics & Decisions 24*(3), 351–371.

van der Vaart, A. W. and J. A. Wellner (1996). Weak convergence. In *Weak convergence and empirical processes*, pp. 16–28. Springer.

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

Wager, S., T. Hastie, and B. Efron (2014). Confidence intervals for random forests: The jackknife and the infinitesimal jackknife. *The Journal of Machine Learning Research 15*(1), 1625–1651.

Wallace, M. L. (2014). Time-dependent tree-structured survival analysis with unbiased variable selection through permutation tests. *Statistics in medicine 33*(27), 4790–4804.

Wright, M. N., T. Dankowski, and A. Ziegler (2017). Unbiased split variable selection for random survival forests using maximally selected rank statistics. *Statistics in medicine 36*(8), 1272–1284.

Wright, M. N. and A. Ziegler (2017). ranger: A fast implementation of random forests for high dimensional data in c++ and r. *Journal of Statistical Software* (77), 1–17.

Xu, R. and S. Adak (2002). Survival analysis with time-varying regression effects using a tree-based approach. *Biometrics 58*(2), 305–315.

Zhang, B., A. A. Tsiatis, E. B. Laber, and M. Davidian (2012). A robust method for estimating optimal treatment regimes. *Biometrics 68*(4), 1010–1018.

Zhang, B., A. A. Tsiatis, E. B. Laber, and M. Davidian (2013). Robust estimation of optimal dynamic treatment regimes for sequential treatment decisions. *Biometrika 100*(3), 681–694.

Manuscript I

Continuous-time targeted minimum-loss based estimation of intervention-specific mean outcomes

Helene C. Rytgaard, Thomas A. Gerds and Mark J. van der Laan

Details:

The manuscript is in revision for the Annals of Statistics.

The current version is partly changed according to reviewers' comments.

Submitted to the Annals of Statistics

CONTINUOUS-TIME TARGETED MINIMUM LOSS-BASED ESTIMATION OF INTERVENTION-SPECIFIC MEAN OUTCOMES

By Helene C. Rytgaard*, Thomas A. Gerds* and Mark J. van der Laan†

*University of Copenhagen** *and University of California, Berkeley*†

This paper studies the generalization of the targeted minimum loss-based estimation (TMLE) framework to estimation of effects of time-varying interventions in settings where both interventions, covariates and outcome can happen at subject-specific time-points on an arbitrarily fine time-scale. TMLE is a general template for constructing asymptotically linear substitution estimators for smooth low-dimensional parameters in infinite-dimensional models. Existing longitudinal TMLE methods are developed for data where observations are made on a discrete time-grid.

We consider a continuous-time counting process model where intensity measures track the monitoring of subjects, and focus on a low-dimensional target parameter defined as the intervention-specific mean outcome at the end of follow-up. To construct our TMLE algorithm for the given statistical estimation problem we derive an expression for the efficient influence curve and represent the target parameter as a functional of intensities and conditional expectations. The high-dimensional nuisance parameters of our model are estimated and updated in an iterative manner according to separate targeting steps for the involved intensities and conditional expectations.

The resulting estimator solves the efficient influence curve equation. We state a general efficiency theorem and describe a highly adaptive lasso estimator for nuisance parameters that allows us to establish asymptotic linearity and efficiency of our estimator under minimal conditions on the underlying statistical model.

1. Introduction. We consider a continuous-time longitudinal data structure of $n \in \mathbb{N}$ independent and identically distributed observations of a multivariate counting process on a bounded interval of time $[0, \tau]$ with distribution P_0 belonging to a semiparametric statistical model $\mathcal{M}$. We are interested in assessing the effect of interventions on an outcome of interest under interventions that can happen at arbitrary points in time and are sub-

Keywords and phrases: targeted minimum loss-based estimation (TMLE), time-varying confounding, continuous-time interventions, semiparametric model, efficient estimation, causal inference.

ject to time-dependent confounding. Our focus is on the construction of an asymptotically efficient substitution estimator of intervention-specific mean outcomes represented as a parameter $\Psi : \mathcal{M} \to \mathbb{R}$.

In all that follows, we use the words 'intervention' and 'treatment' synonymously. Recent developments in the field of causal inference have produced numerous methods to deal with effects of time-varying treatments in presence of time-varying confounding (Robins, 1986, 1987, 1989a,b, 1992, 1998, 2000a,b; Bang and Robins, 2005; Robins, Orellana and Rotnitzky, 2008; van der Laan, 2010a,b; Petersen et al., 2014). These methods deal with settings with a fixed number of time-points at which subjects of a population are all measured and can be intervened upon.

In the present work we consider a continuous-time model, utilizing a counting process framework (Andersen et al., 1993) where intensity processes define the rate of a finite number of continuous monitoring times for each subject conditional on their observed history, see Figure 1. Our approach is closely related to the work of Lok (2008) and of Røysland (2011, 2012), who propose continuous-time versions of structural nested models and marginal structural models (Robins, 1989a,b, 1992, 1998, 2000b,a; Robins, Orellana and Rotnitzky, 2008), respectively, using counting processes and martingale theory. Our parameter $\Psi(P_0)$ is defined via the g-computation formula (Robins, 1986) in terms of interventions on the product integral representing the data-generating distribution.

To estimate $\Psi(P_0)$, we proceed on the basis of the targeted minimum loss-based estimation (TMLE) framework (van der Laan and Rubin, 2006; van der Laan and Rose, 2011, 2018). TMLE is a general methodology for constructing regular and asymptotically linear substitution estimators for smooth low-dimensional parameters in infinite-dimensional models, combining flexible ensemble learning and semiparametric efficiency theory in a two-step procedure. The earliest of the TMLE developments for estimation of effects of time-varying treatments in longitudinal data structures (LTMLE) involve full likelihood estimation and targeting (van der Laan, 2010a,b; Stitelman, De Gruttola and van der Laan, 2012) whereas the later (van der Laan and Gruber, 2012; Petersen et al., 2014) are based on the techniques of sequential regression originating from Bang and Robins (2005). All these methods rely on a discrete time-scale.

To construct our continuous-time TMLE algorithm we derive an expression for the efficient influence curve. The efficient influence curve is the canonical gradient of the functional Ψ and is a central component in the construction of locally efficient estimators of a target parameter in general (Bickel et al., 1993). Semiparametric efficiency theory yields that, given a statistical

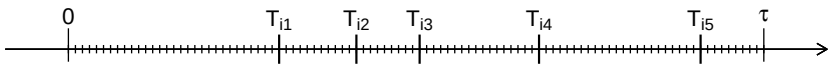


FIG 1. *Observations measured on a continuous scale, with five random time-points* $T_{i1}, \ldots, T_{i5}$ *between time* 0 *and time* τ*, referred to as monitoring times, where actual changes for the given subject* i *are measured.*

model and a target parameter, a regular and asymptotically linear estimator is efficient if and only if its influence curve is equal to the efficient influence curve. Our estimation procedure is based on a representation of the target parameter in terms of intensities and conditional expectations for which we need initial estimators and a targeting updating algorithm. We propose such a targeting algorithm based on separate targeting steps for the intensities and for the conditional expectations that are iterated until convergence. The resulting estimator solves the efficient influence curve equation.

Our TMLE relies on initial estimators for the intensities and the conditional expectations that constitute our nuisance parameters. The TMLE framework allows us to take advantage of flexible and data-adaptive nuisance parameter estimation through super learning (van der Laan, Polley and Hubbard, 2007). A super learner is based on a library of candidate estimators for each nuisance parameter, and uses cross-validation to select the best combination of estimators. The general oracle inequality for cross-validation shows that the super learner performs asymptotically as well as the best combination of estimators in the library (van der Laan and Dudoit, 2003; van der Vaart, Dudoit and van der Laan, 2006). In particular, we discuss the highly adaptive lasso (HAL) (van der Laan, 2017) for all likelihood components. If this HAL estimator is included in the library of the super learner used for initial estimation, we can show that our TMLE estimator is asymptotically linear and efficient under minimal conditions on the model $\mathcal{M}$.

Our methods extend the existing longitudinal TMLE (LTMLE) to the continuous-time case. While the theory and asymptotic performance of LTMLE remain valid on any arbitrarily fine time-scale as long as it is discrete, we

note the following. Sequential regression based LTMLE works by iterating through a sequence of regressions across all time-points and has been a popular choice over the full likelihood-based LTMLE that requires modeling of densities of potentially high-dimensional covariates. Our framework provides a unified methodology that covers both the continuous and the discrete-time case, where intensities of monitoring times are modeled separately and the regression approach can be used to deal with high-dimensional covariates. The substantial difference between LTMLE and our continuous-time TMLE lies in the limitation of LTMLE that one needs sufficiently many events at each time-point to fit the regressions. In contrast, our method can be applied when there are monitoring times with few events, or just one event.

1.1. *Motivating applications.* The methods developed here are applicable to a large variety of problems in pharmacoepidemiology. In this field of research, hazard ratios are often used as measures of the association of time-dependent exposure with time-to-event outcomes (see, e.g., Andersen et al., 2013; Karim et al., 2016; Kessing et al., 2019). However, the interpretation of hazard ratios as the measure of causal treatment effects is hampered for many reasons (Hernán, 2010; Martinussen, Vansteelandt and Andersen, 2018). Furthermore, the time-dependent Cox model cannot be used in presence of time-dependent confounding (Keiding, 1999). In many applications, it is thus of great interest to formulate and estimate statistical parameters under hypothetical treatment interventions that have a causal interpretation under the right assumptions.

Specifically, it may be of interest to assess the effect of a (dynamic) drug treatment regime on the τ-year risk of death. As an example, let $N^a(t)$ denote the process counting visits to a medical doctor who can prescribe the drug. Let the process $A(t)$ be information on the type and dose level of the drug prescribed at patient-specific times $T_1^a < T_2^a < \ldots < T_{N^a(\tau)}^a$ during the study period $[0, \tau]$. Further, let $N^\ell(t)$ denote the process counting, e.g., hospital admissions, and let $L(t)$ be the information collected at, including the reason for, the hospital admission. At visit T_k^a, the doctor considers the baseline characteristics L_0 of the patient, the treatment so far $\{(N^a(t), A(t)) : t \in [0, T_{k-1}^a]\}$, and evaluates changes of covariates $\{(N^\ell(t), L(t)) : t \in [0, T_k^a)\}$ to decide to continue or to change the treatment. We are interested in the result of a hypothetical experiment that we would have liked to have conducted but did not. In our example, the causal effect resulting from such a hypothetical experiment could be the difference of the τ-year risk of death under different (dynamic) treatment regimes.

A treatment regime is any a priori defined rule which can be applied at

doctor visits during the study period to decide if the treatment should be changed or continued. For example, the intervention could state that the patient stays on an initially randomized drug throughout the entire study period. A dynamic treatment regime is an example of an adaptive intervention which allows the decision to depend on the current history of the patient. The rule could be that whenever the measurement of a blood marker value exceeds a certain threshold the dose level of the drug should be adapted. By comparison of the effects of different regimes we may further learn optimal strategies for treatment interventions based on patients' medical past. In this article, we focus on interventions of the treatment process keeping the doctor visits where they naturally occur.

The data analysis is complicated by the fact that the observed data are subject to time-dependent confounding: At any point in time, doctors make treatment decisions for a reason, and past treatment may further influence future values of covariates (biomarkers and diagnoses) and future treatment changes. Additionally, a proportion of subjects in the population may be lost to follow-up (censored) which can likewise depend on prior treatment decisions and covariate values. Our estimation framework allows us to control for the history of all observed time-varying covariates and treatment choices that we believe could be predictive of both treatment decisions and censoring, and of the risk of death. Importantly, in the case where neither the treatment decision nor the censoring mechanism depend on any unrecorded health information, i.e., the observed history at any point in time is sufficient to predict the next treatment decision and the censoring at that time, the sequential randomization assumption holds, and the estimated effects of hypothetical interventions based on the observed data can be interpreted causally.

1.2. *Organization of article.* The article is organized as follows. We first define the statistical estimation problem. Section 2 introduces the continuous-time longitudinal setting and presents the model of a single subject along with the likelihood. In particular, Section 2.2 defines interventions on the likelihood, Section 2.4 defines the target parameter, and Section 2.5 provides the efficient influence function for the estimation problem. In Section 2.4.1, we briefly review the assumptions under which the target parameter identifies the causal effect of interest. Section 3 presents a representation of the target parameter in terms of intensities and conditional expectations for which we need to construct an estimation procedure. Section 4 states a general theorem for asymptotically efficient substitution estimation. Section 5 introduces targeted minimum loss-based estimation (TMLE) for the considered estimation

problem. Section 6 describes initial estimation of the likelihood components needed for the targeting updating steps and for estimation of the target parameter that fulfills the criteria of the efficiency theorem. Section 7 presents a particular TMLE algorithm for estimation of the target parameter in the continuous-time setting that involves separate targeting for conditional expectations and intensities, pooled over time. Section 8 reviews inference for the TMLE estimator. Section 9 presents the results of a simulation study as a demonstration of the methods and a proof-of-concept. Section 10 closes with a discussion.

2. Formulation of the statistical estimation problem. We start out defining the data structure, the statistical estimation problem and the target parameter.

2.1. *Notation and setup.* We represent the subject-specific information in terms of a counting process model (Andersen et al., 1993). Suppose $n \in \mathbb{N}$ subjects of a population are followed in a bounded interval of time $[0, \tau]$. Let (N^a, N^ℓ, N^c, N^d) be a multivariate counting process generating random times at which treatment, covariates, censoring status and survival status may change. At jump times of N^a we observe changes in a treatment regime $A(t)$ taking values in finite set $\mathcal{A}$ and at jump times of N^ℓ we observe changes of a covariate vector $L(t)$ with values in a compact subset of $\mathbb{R}^d$. The processes N^c and N^d generate changes in censoring status and death status, respectively. Furthermore, L_0 denotes a vector of baseline covariates measured at time 0, and Y a real valued outcome variable measured at time τ. We assume that there are no events at time zero, and that realizations of all processes are càdlàg functions on $[0, \tau]$. Let $(\mathcal{F}_t)_{t\geq 0}$ denote the filtration generated by the history of the observed processes up to time t. Specifically, we have $\mathcal{F}_0 = \sigma(L_0)$.

Let $T_1^a < T_2^a < \ldots < T_{N^a(\tau)}^a$ and $T_1^\ell < T_2^\ell < \ldots < T_{N^\ell(\tau)}^\ell$ be the random times at which the treatment regime A and the covariate process L may change, and let T^c and T^d be the right-censoring time and the survival time, respectively. By definition, $N^a(t)$ and $N^\ell(t)$ are the subject-specific numbers of treatment and covariate monitoring times in $[0, t]$. The actual end of follow-up for a subject is $\min(T^c, T^d, \tau)$ and the total number of unique event times before time t is,

$$K_t = \#\big\{\{T_1^a, \ldots, T_{N^a(t)}^a\} \cup \{T_1^\ell, \ldots, T_{N^\ell(t)}^\ell\} \cup \{T^c, T^d\}\big\}. \tag{1}$$

For each subject, the number of change points of the multivariate counting process on $[0, \tau]$, $K = K_\tau$, is finite.

For $x = a, \ell, d, c$, we denote by Λ_0^x the cumulative intensity that characterizes the compensator of N^x. We further denote by $M^x = N^x - \Lambda_0^x$ the corresponding martingale. Heuristically, we have that,

$$P(N^x(dt) = 1 \,|\, \mathcal{F}_{t-}) = \mathbb{E}[N^x(dt) \,|\, \mathcal{F}_{t-}] = d\Lambda_0^x(t \,|\, \mathcal{F}_{t-}),$$

where the increment $N^x(dt)$ is non-zero and equal to 1 if and only if there is a jump of N^x in the infinitesimal interval $[t, t + dt)$ (Andersen et al., 1993; Gill, 1994).

We assume that the processes A and L only change at monitoring times. The distribution of the treatment $A(t)$ at any time t where $N^a(t)$ jumps is denoted $\pi_{0,t}(a \,|\, \mathcal{F}_{t-}) = P(A(t) = a \,|\, \mathcal{F}_{t-})$, $a \in \mathcal{A}$, and the distribution of covariates $L(t)$ at any time t where $N^\ell(t)$ jumps is characterized by a conditional density $\mu_{0,t}(\ell \,|\, \mathcal{F}_{t-})$ with respect to a dominating measure ν_L. We assume that $A(t) = A(t-)$ and $L(t) = L(t-)$ otherwise. Lastly, the probability distribution of baseline covariates L_0 is characterized by the conditional density μ_{0,L_0} with respect to a dominating measure ν_{L_0}.

2.1.1. *Observations.* For subject i, with $i = 1, \ldots, n$ independent subjects, let $\{T_{i,k}\}_{k=1}^{K_{i,t}}$ denote the ordered set of unique event times up to time t. The observed data for subject i in a bounded interval $[0, t]$, $t \leq \tau$, is given by,

(2)
$$\bar{O}_i(t) = \left\{ \left(L_{0,i}, s, N_i^a(s), A_i(s), N_i^\ell(s), L_i(s), N_i^d(s), N_i^c(s) \right) \,:\, s \in \{T_{i,k}\}_{k=1}^{K_{i,t}} \right\}.$$

We also use the shorthand notation $O_i = \bar{O}_i(\tau)$ and denote by $\mathcal{O}$ the space where O_i takes its values. Let $\mathbb{P}_n$ denote the empirical distribution of the data $\{O_i\}_{i=1}^n$. For a dataset with n observations, we use,

$$0 = t_0 < t_1 < \cdots < t_{\bar{K}_n}, \tag{3}$$

with $\bar{K}_n = \sum_{i=1}^n K_i$, to denote the ordered sequence of unique times of changes $\cup_{i=1}^n \{T_{i,k}\}_{k=1}^{K_i}$.

The distribution P_0 of the observed data O factorizes according to the time-ordering, going from one infinitesimal time interval to the next with $d\Lambda_0^x(t \,|\, \mathcal{F}_{t-})$ representing the conditional probability of an event of N^x in $[t, t + dt)$, for $x = a, \ell, d, c$ (Andersen et al., 1993, Section II.7). Accordingly,

we express the density p_0 of P_0 as,

(4)
$$p_0(O) = \mu_{0,L_0}(L_0) \prod_{t\in(0,\tau]} \left(d\Lambda_0^{\ell}(t \mid \mathcal{F}_{t-})\, \mu_{0,t}(L(t) \mid \mathcal{F}_{t-})\right)^{N^{\ell}(dt)} \left(1 - d\Lambda_0^{\ell}(t \mid \mathcal{F}_{t-})\right)^{1-N^{\ell}(dt)} \prod_{t\in(0,\tau]} \left(d\Lambda_0^{a}(t \mid \mathcal{F}_{t-})\, \pi_{0,t}(A(t) \mid \mathcal{F}_{t-})\right)^{N^{a}(dt)} \left(1 - d\Lambda_0^{a}(t \mid \mathcal{F}_{t-})\right)^{1-N^{a}(dt)} \prod_{t\in(0,\tau]} \left(d\Lambda_0^{c}(t \mid \mathcal{F}_{t-})\right)^{N^{c}(dt)} \left(1 - d\Lambda_0^{c}(t \mid \mathcal{F}_{t-})\right)^{1-N^{c}(dt)} \prod_{t\in(0,\tau]} \left(d\Lambda_0^{d}(t \mid \mathcal{F}_{t-})\right)^{N^{d}(dt)} \left(1 - d\Lambda_0^{d}(t \mid \mathcal{F}_{t-})\right)^{1-N^{d}(dt)},$$

with $\prod$ denoting the product integral (Gill and Johansen, 1990; Andersen et al., 1993, Section II.6). A particularly nice aspect of the product-integral representation is that it gives a unified presentation for the discrete and the continuous time case.

2.2. *Interventions.* We are interested in estimating the effect of dynamic treatment regimes (Robins, 2002; Murphy et al., 2001; Hernán et al., 2006; van der Laan and Petersen, 2007) corresponding to hypothetical experiments, under hypothetical interventions, where data had been generated differently. An intervention defines a rule specifying treatment at each intervention time point given the data so far. In our setting, we allow the number and schedule of the intervention time-points to be subject-specific and to occur in continuous time. We thus distinguish between interventions that control the treatment, but not the schedule of the intervention time points, and interventions that control both the treatment and the schedule of the intervention time points. In addition to the treatment regimes of interest, interventions always control the censoring mechanism, such that, in the hypothetical experiment, all subjects are followed for the entire study period $[0, \tau]$.

The observed data are generated by the distribution P_0 which factorizes as displayed in (4). We define interventions directly on P_0, by replacing a subset of its components by an intervention-specific choice. To formulate this, we decompose the observed data distribution P_0 into two parts which we refer to as the interventional part (G_0) and the non-interventional part

(Q_0), respectively. We parametrize P_0 accordingly,

$$dP_0 = dP_{Q_0,G_0} = \prod_{t\in[0,\tau]} dQ_{0,t}\, dG_{0,t},$$

with $dG_{0,t}(o)$ and $dQ_{0,t}(o)$ denoting the conditional measures, given the observed history, corresponding to the interventional part and the non-interventional part at time t, respectively. We use $g_{0,t}$ to denote the density of $G_{0,t}$ and likewise $q_{0,t}$ to denote the density of $Q_{0,t}$, both with respect to appropriate dominating measures.

Now, an intervention involves replacing G_0 by some G^* encoding how treatment and censoring is generated conditional on the available history in the hypothetical experiment. In its generality, this is what is referred to as a *randomized plan* in Gill and Robins (2001, Sections 6 and 7), or a *stochastic* intervention (Robins, Hernán and Siebert, 2004; Dawid and Didelez, 2010), but it includes *static* and *dynamic* interventions (Hernán et al., 2006; Chakraborty and Moodie, 2013) plans as special cases as we explain below. Which components we include in G_0, and thus intervene upon, depends on what kinds of effects we are interested in and thus what scientific question we wish to address. Consider the following options.

DEFINITION 1 (Interventions on treatment assigned). *An intervention on treatment assigned involves replacing $\pi_{0,t}$ by some choice π_t^*. Thus, the interventional part includes the treatment and the censoring mechanism:*

$$dG_{0,t}(O) = \big(\pi_{0,t}(A(t)\,|\,\mathcal{F}_{t-})\big)^{N^a(dt)}\big(d\Lambda_0^c(t\,|\,\mathcal{F}_{t-})\big)^{N^c(dt)} \big(1 - d\Lambda_0^c(t\,|\,\mathcal{F}_{t-})\big)^{1-N^c(dt)}. \tag{5}$$

The intervention prevents censoring and specifies the treatment regime π_t^*, such that,

$$dG_t^*(O) = \big(1 - N^c(t)\big)\,\pi_t^*(A(t)\,|\,\mathcal{F}_{t-}).$$

The distribution of the intervention times is not intervened upon, i.e., Λ_0^a is included in the non-interventional part.

When the interventional treatment distributions are degenerated and, for example, set to a single value $a^* \in \mathcal{A}$ throughout the entire study period,

$$\pi_t^*(A(t)\,|\,\mathcal{F}_{t-}) = \big(\mathbb{1}\{A(t) = a^*\}\big)^{N^a(dt)}. \tag{6}$$

we refer to π_t^* as a static intervention since it deterministically sets $A(t) = a^*$ (at treatment monitoring times). A dynamic intervention, on the other hand, defines π_t^* that assigns $A(t)$, still deterministically, to some value based on the subject's past.

DEFINITION 2 (Intervention on treatment and schedule). *The interventional part includes the treatment, the schedule of intervention times, and the censoring mechanism:*

$$dG_{0,t}(O) = \big(d\Lambda_0^a(t \mid \mathcal{F}_{t-})\, \pi_{0,t}(A(t) \mid \mathcal{F}_{t-})\big)^{N^a(dt)} \big(1 - d\Lambda_0^a(t \mid \mathcal{F}_{t-})\big)^{1-N^a(dt)} \\ \big(d\Lambda_0^c(t \mid \mathcal{F}_{t-})\big)^{N^c(dt)} \big(1 - d\Lambda_0^c(t \mid \mathcal{F}_{t-})\big)^{1-N^c(dt)}.$$

The intervention prevents censoring and specifies a treatment regime by π_t^* and $\Lambda_0^{a,*}$, such that,

$$dG_{0,t}^*(O) = \big(d\Lambda_0^{a,*}(t \mid \mathcal{F}_{t-})\, \pi_{0,t}^*(A(t) \mid \mathcal{F}_{t-})\big)^{N^a(dt)} \big(1 - d\Lambda_0^{a,*}(t \mid \mathcal{F}_{t-})\big)^{1-N^a(dt)}.$$

An intervention on the schedule of treatment decisions involves replacing the intensity Λ_0^a by some choice $\Lambda_0^{a,*}$. In the context of the applications described in Section 1.1, this could be to decrease or increase the frequency of doctor visits, for instance to ensure at least a monthly visit.

To focus our presentation, we continue with $dG_{0,t}$ as defined by (5) according to Definition 1, considering interventions only on the treatment decision π_t and on the censoring mechanism Λ^c. Note that this defines the non-interventional part as,

$$dQ_{0,t}(O) = \big(d\Lambda_0^a(t \mid \mathcal{F}_{t-})\big)^{N^a(dt)} \big(1 - d\Lambda_0^a(t \mid \mathcal{F}_{t-})\big)^{1-N^a(dt)} \\ \big(d\Lambda_0^\ell(t \mid \mathcal{F}_{t-})\, \mu_{0,t}(L(t) \mid \mathcal{F}_{t-})\big)^{N^\ell(dt)} \big(1 - d\Lambda_0^\ell(t \mid \mathcal{F}_{t-})\big)^{1-N^\ell(dt)} \\ \big(d\Lambda_0^d(t \mid \mathcal{F}_{t-})\big)^{N^d(dt)} \big(1 - d\Lambda_0^d(t \mid \mathcal{F}_{t-})\big)^{1-N^d(dt)},$$

for $t > 0$.

2.3. *Statistical model.* Let $\mathcal{Q}$ denote the parameter set for the non-interventional part Q_0 and $\mathcal{G}$ the parameter set for the interventional part G_0. We consider a statistical model $\mathcal{M}$ as follows:

$$\mathcal{M} = \left\{ P \,:\, dP = dP_{Q,G} = \prod_{t \in [0,\tau]} dQ_t dG_t,\ G \in \mathcal{G},\, Q \in \mathcal{Q} \right\}. \tag{7}$$

In Section 6 we consider Q, G to be contained in the set of càdlàg functions with bounded variation norm. For now we leave $\mathcal{Q}, \mathcal{G}$ unspecified.

2.4. *Target parameter.* Suppose an intervention G^* is given. We define the post-interventional distribution for any $P \in \mathcal{M}$ by replacing G by G^* in $P_{Q,G}$. The resulting P_{Q,G^*} is commonly referred to as the *g-computation*

formula (Robins, 1986; Gill and Robins, 2001). We will also denote this by P^{G^*}. Based on the data $O = \bar{O}(\tau)$, our overall aim is to estimate the expectation of the outcome of interest Y under the g-computation formula. That is, we are interested in the parameter $\Psi^{G^*} : \mathcal{M} \to \mathbb{R}$ given by,

$$\Psi^{G^*}(P) = \mathbb{E}_{P^{G^*}}\left[Y\right] = \int_{\mathcal{O}} y \prod_{t\in[0,\tau]} dQ_t(o)\, dG_t^*(o), \tag{8}$$

where the notation $\mathbb{E}_{P^{G^*}}$ refers to the expectation operator with respect to the post-interventional measure $P^{G^*} = P_{Q,G^*}$.

In this paper we focus on an all-cause mortality outcome, $Y = N^d(\tau)$, so that (8) is the expected risk of dying by time τ under the distribution defined by the g-computation formula. We emphasize, however, that in principle Y can be defined as any mapping of the observed past $\mathcal{F}_\tau$ as long as Y takes value in a compact set. We denote by $\psi_0 := \Psi^{G^*}(P_0)$, the true value of the target parameter. The target parameter is identifiable from the observed data under the following positivity assumption.

ASSUMPTION 1 (Positivity). *We assume absolute continuity of* $d\bar{G}_\tau^* = \prod_{s<\tau} dG_s^*$ *with respect to* $d\bar{G}_\tau = \prod_{s<\tau} dG_s$, *i.e.,* $d\bar{G}_\tau^* \ll d\bar{G}_\tau$. *This implies existence of the Radom-Nikodym derivative* $d\bar{G}_\tau^*/d\bar{G}_{0,\tau}$.

The g-computation formula arising from replacing the observed G_0 by G^* and the resulting target parameter in (8) are well-defined statistical quantities by Assumption 1.

2.4.1. *Causal parameter and causal interpretability.* Causal interpretability of the g-computation formula in the setting of fully discrete data is provided by Robins' work (Robins, 1986, 1987, 1989a) under assumptions of sequential randomization (SRA) and positivity (for a nice review, see also, Hernan and Robins, 2020, Part III), and is further generalized by Gill and Robins (2001) to continuously varying covariates and treatments. Our setting differs from their considered setting in that subjects are measured at different random times that take place continuously in time: Our g-computation formula is represented as a product integral over times where something actually happens, whereas the g-computation formula of Gill and Robins (2001) consists of a finite product over times of a discrete grid. Nevertheless, note that the counting processes only have finitely many changes in the compact time interval $[0, \tau]$, and that interventions on the treatment decision are in fact only applied at a finite number of (random) times. Thus, the 'traditional' causal assumptions as stated by Gill and Robins (2001), applied now

at the random rather than at fixed times, ensure the causal interpretation of the g-computation formula (Gill and Robins, 2001, Theorem 2). We consider here the SRA assumption for the treatment mechanism in our setting. A more detailed discussion can be found in Gill and Robins (2001).

Particularly, let O^{G^*} be the counterfactual random variable representing the data that would have been observed, had G^* been adhered to from the beginning of follow-up rather than the factual G. The causal parameter of interest is $\mathbb{E}[Y^{G^*}]$, the mean causal effect on Y of imposing the intervention G^*. For our setting with random treatment monitoring times, $T_1^a < T_2^a < \ldots < T_{N^a(\tau)}^a$, we formulate the sequential randomization assumption as follows:

ASSUMPTION 2 (SRA). $Y^{G^*} \perp\!\!\!\perp A(T_k^a) \,|\, \bar{O}(T_k^a-)$, *for all* $k = 1, \ldots, N^a(\tau)$.

Under Assumptions 1 and Assumption 2, P^{G^*} identifies the distribution of O^{G^*} and our target parameter the expectation of the counterfactual outcome, i.e., $\Psi^{G^*}(P) = \mathbb{E}[Y^{G^*}]$.

2.5. *Canonical gradient.* The canonical gradient of the pathwise derivative of the target parameter characterizes the information bound of the estimation problem relative to the statistical model $\mathcal{M}$ (Bickel et al., 1993). It is also known as the efficient influence curve. The following theorem provides a representation of the canonical gradient for our statistical model $\mathcal{M}$ and target parameter $\Psi^{G^*} : \mathcal{M} \to \mathbb{R}$. We will use this representation to construct an asymptotically efficient estimator.

We derive our expression for the canonical gradient as the projection of the influence curve of an inverse probability of action weighted (IPAW) estimator in the submodel of $\mathcal{M}$ that takes the interventional part G as known onto the tangent space $\mathcal{T}_Q$ of the non-interventional part Q (van der Laan and Robins, 2003). The derivations can be found in Appendix A.

THEOREM 1 (Canonical gradient). *The canonical gradient* $D^*(P)$ *at* P

in $\mathcal{M}$ can be represented as follows,

$$
\begin{aligned}
D^*(P) &= \mathbb{E}_{PG^*}[Y \mid \mathcal{F}_0] - \Psi^{G^*}(P) \\
&+ \int_0^\tau \prod_{s<t} \frac{dG_s^*}{dG_s} \Big(\mathbb{E}_{PG^*}[Y \mid L(t), N^\ell(t), \mathcal{F}_{t-}] - \mathbb{E}_{PG^*}[Y \mid N^\ell(t), \mathcal{F}_{t-}] \Big) N^\ell(dt) \\
&+ \int_0^\tau \prod_{s<t} \frac{dG_s^*}{dG_s} \Big(\mathbb{E}_{PG^*}[Y \mid \Delta N^\ell(t) = 1, \mathcal{F}_{t-}] - \mathbb{E}_{PG^*}[Y \mid \Delta N^\ell(t) = 0, \mathcal{F}_{t-}] \Big) M^\ell(dt) \\
&+ \int_0^\tau \prod_{s<t} \frac{dG_s^*}{dG_s} \Big(\mathbb{E}_{PG^*}[Y \mid \Delta N^a(t) = 1, \mathcal{F}_{t-}] - \mathbb{E}_{PG^*}[Y \mid \Delta N^a(t) = 0, \mathcal{F}_{t-}] \Big) M^a(dt) \\
&+ \int_0^\tau \prod_{s<t} \frac{dG_s^*}{dG_s} \Big(1 - \mathbb{E}_{PG^*}[Y \mid N^d(t) = 0, \mathcal{F}_{t-}] \Big) M^d(dt).
\end{aligned}
$$

We here follow notation of Andersen et al. (1993) and use Δ to denote the difference operator defined by $\Delta X = X - X_-$ for a càdlàg process X.

3. Representation of the target parameter by iterated expectations. Estimation of the target parameter requires evaluation of a large integral. We here present a parametrization of the target parameter in terms of a nested sequence of conditional expectations which is central for our estimation procedure. As for the discrete-time analogue (Bang and Robins, 2005; Robins, 2000b), the parametrization is defined backwards through time, starting at the end of the study period τ. The main difference to the discrete-time representation is that the time-points where events happen are random.

The notation that we present in this section will be used throughout the remainder of the paper. For $P \in \mathcal{M}$, a fixed regime G^* according to Definition 1 and $t \in [0, \tau]$, we denote by:

(9)
$$
Z_t^{G^*} := \mathbb{E}_{PG^*}\left[Y \,\middle|\, L(t), N^\ell(t), N^a(t), N^d(t), \mathcal{F}_{t-} \right] = \int Y \prod_{s \geq t} dG_s^* \prod_{s>t} dQ_s.
$$

Note that the conditioning set of $Z_t^{G^*}$ excludes the interventional part at time t. We further denote the conditional expectation where $L(t)$ has been integrated out by:

(10)
$$
\begin{aligned}
Z_{t,L(t)}^{G^*} &:= \mathbb{E}_{PG^*}\left[Y \,\middle|\, N^\ell(t), N^a(t), N^d(t), \mathcal{F}_{t-} \right] \\
&= \mathbb{E}_{PG^*}\left[Z_t^{G^*} \,\middle|\, N^\ell(t), N^a(t), N^d(t), \mathcal{F}_{t-} \right] \\
&= \int \left(\int Y \prod_{s \geq t} dG_s^* \prod_{s>t} dQ_s \right) d\mu_{0,t}(L(t) \mid \mathcal{F}_{t-}).
\end{aligned}
$$

The notation using the subscript '$L(t)$' in $Z^{G^*}_{t,L(t)}$ to refer to $L(t)$ being integrated out follows the notation of van der Laan and Gruber (2012).

LEMMA 1. *Define, for any $P \in \mathcal{M}$:*

$$\boldsymbol{Z} = \boldsymbol{Z}(P) := \left(Z^{G^*}_t, Z^{G^*}_{t,L(t)}, \Lambda^\ell(t), \Lambda^a(t), \Lambda^d(t) \,:\, t \in [0,\tau]\right). \tag{11}$$

Particularly, let $\boldsymbol{Z}_0 = \boldsymbol{Z}(P_0)$. The target parameter Ψ^{G^*} defined in (8) can be represented as a functional of P only through $\boldsymbol{Z}$.

PROOF. See Appendix A.6. □

In line with the sequential regression representation of Bang and Robins (2005), we may thus utilize that $\boldsymbol{Z}$ rather than P itself can be used to evaluate the target parameter. Our aim is to estimate $\boldsymbol{Z}$ in an optimal way. The canonical gradient presented in Theorem 1 can be represented as a functional of $\boldsymbol{Z}$ and G; this will guide the construction of estimators for $\boldsymbol{Z}$ as a result of our efficiency theorem in Section 4.

4. Efficiency theorem for substitution estimation. In the previous section we have demonstrated that the target parameter can be represented as a functional of $P \in \mathcal{M}$ through $\boldsymbol{Z}$. We define a substitution estimator of the target parameter $\psi_0 = \Psi(P_0)$ based on an estimator $\hat{\boldsymbol{Z}}_n$ for $\boldsymbol{Z}_0$. The following theorem states the general conditions for asymptotic efficiency of such substitution estimation of ψ_0. The proof is short and relies on an analysis of the separate terms of a von Mises expansion of the target parameter. In Section 6.1 we state sufficient smoothness conditions on $\mathcal{M}$ and describe initial estimators that meet the regularity conditions. Throughout we use the notation $Pf = \int f dP$ and $\|f\|_P = \sqrt{Pf^2}$.

We show in Appendix A.5 that the difference $\Psi^{G^*}(P) - \Psi^{G^*}(P_0)$ admits the following presentation,

$$\Psi^{G^*}(P) - \Psi^{G^*}(P_0) = -P_0 D^*(P) + R_2(P, P_0), \qquad P \in \mathcal{M}, \tag{12}$$

where $D^*(P)$ is the canonical gradient and the second-order remainder $R_2(P, P_0)$ is given by,

$$\begin{aligned} R_2(P,P_0) &= \Psi^{G^*}(P) - \Psi^{G^*}(P_0) + P_0 D^*(P) \\ &= \int_0^\tau \int_{\mathcal{O}} Y \frac{1}{\bar{g}_t}(\bar{g}_{0,t} - \bar{g}_t) \prod_{s \le \tau} dG^*_s \prod_{s<t} dQ_{0,s} \left(dQ_{0,t} - dQ_t\right) \prod_{s>t} dQ_s. \end{aligned} \tag{13}$$

Here we have used the notation,

$$\bar{g}_t = \prod_{s<t} g_s, \quad \text{and,} \quad \bar{g}_{0,t} = \prod_{s<t} g_{0,s},$$

with g_t being the density of the interventional part G_t as defined in Section 2.2. Note that an estimator $\hat{P}_n^*$ is characterized by $(\boldsymbol{Z}_n^*, \hat{G}_n)$.

THEOREM 2. *Consider an estimator $\hat{P}_n^*$ for P_0, such that*

$$\mathbb{P}_n D^*(\hat{P}_n^*) = o_P(n^{-1/2}). \tag{14}$$

If the following conditions 1 and 2 hold true,

Condition 1) $R_2(\hat{P}_n^*, P_0) = o_P(n^{-1/2})$,

Condition 2) $D^*(\hat{P}_n^*)$ *belongs to a Donsker class, and* $P_0\big(D^*(\hat{P}_n^*) - D^*(P_0)\big)^2$ *converges to zero in probability,*

then,

$$\Psi^{G^*}(P_n^*) - \Psi^{G^*}(P_0) = \mathbb{P}_n D^*(P_0) + o_P(n^{-1/2}),$$

that is, $\Psi^{G^*}(\hat{P}_n^*)$ is asymptotically linear at P_0 with influence curve $D^*(P_0)$ and is thus asymptotically efficient among all locally regular estimators at P_0.

PROOF. The proof of the theorem relies on the expansion (12). Applying the representation at $\hat{P}_n^*$ and using Equation (14) from the theorem now yields,

$$\Psi^{G*}(\hat{P}_n^*) - \Psi^{G*}(P_0) = (\mathbb{P}_n - P_0)D^*(\hat{P}_n^*) + R_2(\hat{P}_n^*, P_0) + o_P(n^{-1/2}).$$

It is now a result of empirical process theory (see, e.g., van der Vaart, 2000, Lemma 19.24) that condition 2 implies,

$$(\mathbb{P}_n - P_0)\big(D^*(\hat{P}_n^*) - D^*(P_0)\big) = o_P(n^{-1/2}),$$

which finishes the proof. □

If we are able to construct estimators $\hat{P}_n^*$ that meet the conditions of Theorem 2 and solve Equation (14), the so-called efficient influence curve equation, then the resulting substitution estimator $\Psi^{G^*}(\hat{P}_n^*)$ is asymptotically linear and efficient. In the following, we comment on conditions 1 and 2.

REMARK 1 (Second-order remainder). *The second-order remainder expressed in (13) displays a double robustness structure (van der Laan and Robins, 2003) in the sense that,*

$$R_2(P, P_0) = 0 \quad \textit{if} \quad \bar{g}_t = \bar{g}_{0,t} \quad \textit{or} \quad q_t = q_{0,t}.$$

When $\bar{g}_t$ is bounded away from zero, the product structure of the remainder $R_2(P, P_0)$ yields, by use of the Cauchy-Schwartz inequality, an upper bound in terms of the $L_2(P_0)$-norm of $(\bar{g}_{0,t} - \bar{g}_t)$ and the $L_2(P_0)$-norm of $(q_{0,t} - q_t)$,

$$\int_0^\tau \|\bar{g}_{0,t} - \bar{g}_t\|_{P_0} \, \|q_{0,t} - q_t\|_{P_0} \, dt.$$

The required convergence rate $R_2(\hat{P}_n^*, P_0) = o_P(n^{-1/2})$ will for example be achieved if we estimate both parts at a rate faster than $o_P(n^{-1/4})$. In Section 6.1 we apply recent results on highly adaptive lasso (HAL) estimation (van der Laan, 2017) to show that such estimators do exist.

REMARK 2 (Donsker class conditions). *In Section 6.1 we give conditions under which $D^*(\hat{P}_n^*) \in \mathscr{F}$ where $\mathscr{F} = \{D^*(P) : P\}$ is a Donsker class. The key is that there is a finite number of monitoring times per subject, so that any subject contributes with finitely many terms to the likelihood. Then the canonical gradient can be written as a well-behaved mapping of the nuisance parameters such that Donsker properties of the former will rely on Donsker properties of the latter. An important Donsker class is the class of càdlàg functions with finite variation.*

Theorem 2 tells us what we need for efficient estimation of our target parameter. The next sections deal with construction of such an estimator.

5. Targeted minimum loss-based estimation (TMLE). We present a TMLE procedure for construction of an estimator that will satisfy the conditions of Theorem 2. In summary, this consists of, first, constructing initial estimators for the components of $\boldsymbol{Z}$ and G, and, second, setting up an algorithm for performing an update of the collection of initial estimators that guarantees that it solves the efficient influence curve equation (14).

We also refer to the second step as the targeting step or the targeting algorithm. It involves for each component of $\boldsymbol{Z}$ a choice of a loss function and a corresponding path indexed by $\varepsilon \in \mathbb{R}$ through the initial estimator of that component such that the generated score at $\varepsilon = 0$ gives a desired part of the canonical gradient. The general targeting algorithm involves iterative updating steps that are repeated until convergence, at which point the efficient influence curve equation (14) is solved.

A certain amount of extra notation is needed. For $t \in (0, \tau]$, we define the "clever weights",

$$h_t^{G^*} = \prod_{s<t} \frac{dG_s^*}{dG_s}, \tag{15}$$

that only depends on the G-part of the likelihood, and the "clever covariates",

$$h_t^{\ell} = \mathbb{E}_{P^{G^*}}[Y \mid \Delta N^{\ell}(t) = 1, \mathcal{F}_{t-}] - \mathbb{E}_{P^{G^*}}[Y \mid \Delta N^{\ell}(t) = 0, \mathcal{F}_{t-}], \tag{16}$$

$$h_t^{a} = \mathbb{E}_{P^{G^*}}[Y \mid \Delta N^{a}(t) = 1, \mathcal{F}_{t-}] - \mathbb{E}_{P^{G^*}}[Y \mid \Delta N^{a}(t) = 0, \mathcal{F}_{t-}], \tag{17}$$

$$h_t^{d} = 1 - \mathbb{E}_{P^{G^*}}[Y \mid N^{d}(t) = 0, \mathcal{F}_{t-}], \tag{18}$$

that only depend on $\boldsymbol{Z}$. This now allows us to write the canonical gradient as,

$$\begin{aligned} D^*(P) &= Z_{t=0}^{G^*} - \Psi^{G^*}(P) \\ &+ \int_0^{\tau} h_t^{G^*} \big(Z_t^{G^*} - Z_{t,L(t)}^{G^*}\big) dN^{\ell}(t) + \sum_{x \in \{a,\ell,d\}} \int_0^{\tau} h_t^{G^*}\, h_t^{x}\, dM^{x}(t). \end{aligned} \tag{19}$$

In the following, we define loss functions and one-dimensional parametric submodels for each component of $\boldsymbol{Z}$ such that the scores are equal to the respective terms of (19). These will be used to construct our targeting algorithm in Section 7.

5.1. *Loss function and submodel for $Z_{t,L(t)}^{G^*}$.* We need a loss function and a submodel for $Z_{t,L(t)}^{G^*}$ such that the score of the submodel equals the first term of (19). This term consists of an integral over a difference between $Z_t^{G^*}$ and $Z_{t,L(t)}^{G^*}$. Thus, we will define our loss function and submodel for $Z_{t,L(t)}^{G^*}$ for a given $Z_t^{G^*}$. First, we define the time-point specific logarithmic loss for $Z_{t,L(t)}^{G^*}$, indexed by $Z_t^{G^*}$,

$$\mathcal{L}_{t,Z_t^{G^*}}\big(Z_{t,L(t)}^{G^*}\big) = Z_t^{G^*} \log Z_{t,L(t)}^{G^*} + \big(1 - Z_t^{G^*}\big) \log \big(1 - Z_{t,L(t)}^{G^*}\big).$$

With this loss function, the parametric submodel,

$$\operatorname{logit} Z_{t,L(t)}^{G^*}(\varepsilon) = \operatorname{logit} Z_{t,L(t)}^{G^*} + \varepsilon h_t^{G^*},$$

has the desired property that,

$$\frac{d}{d\varepsilon} \mathcal{L}_{t,Z_t^{G^*}}\big(Z_{t,L(t)}^{G^*}(\varepsilon)\big)\Big|_{\varepsilon=0} = h_t^{G^*}\big(Z_t^{G^*} - Z_{t,L(t)}^{G^*}\big).$$

Correspondingly, we define the integrated loss for $\bar{Z}^{G^*}_{\tau,L(\tau)} = (Z^{G^*}_{s,L(s)} : 0 \leq s \leq \tau)$, indexed by $\bar{Z}^{G^*}_{\tau} = (Z^{G^*}_{s} : 0 \leq s \leq \tau)$,

$$\bar{\mathcal{L}}_{\bar{Z}^{G^*}_{\tau}}\left(\bar{Z}^{G^*}_{\tau,L(\tau)}(\varepsilon)\right) = \int_0^\tau \mathcal{L}_{t,Z^{G^*}_t}\left(Z^{G^*}_{t,L(t)}(\varepsilon)\right) dN^\ell(t),$$

for which we have that,

$$\begin{aligned}\frac{d}{d\varepsilon}\bar{\mathcal{L}}_{\bar{Z}^{G^*}_{\tau}}\left(\bar{Z}^{G^*}_{\tau,L(\tau)}(\varepsilon)\right)\Big|_{\varepsilon=0} &= \int_0^\tau \frac{d}{d\varepsilon}\mathcal{L}_{t,Z^{G^*}_t}\left(Z^{G^*}_{t,L(t)}(\varepsilon)\right)\Big|_{\varepsilon=0} dN^\ell(t) \\ &= \int_0^\tau h^{G^*}_t\left(Z^{G^*}_t - Z^{G^*}_{t,L(t)}\right) dN^\ell(t),\end{aligned}$$

which, as we wanted, equals the first term of the part of the canonical gradient expressed in (19).

5.2. *Loss function and submodel for the intensities* $\Lambda^x(t)$. We consider the case that Λ^x is absolutely continuous and denote by λ^x the intensity rate such that $\Lambda^x(t \,|\, \mathcal{F}_{t-}) = \int_0^t \lambda^x(s \,|\, \mathcal{F}_{s-})ds$. For each $x = a, \ell, d$, we specify a proportional hazard type submodel for Λ^x with time-dependent covariates as follows,

$$d\Lambda^x(t;\varepsilon) = d\Lambda^x(t)\exp(\varepsilon h^{G^*}_t h^x_t), \tag{20}$$

together with the partial log-likelihood loss function,

$$\begin{aligned}\mathcal{L}_x(\Lambda^x) &= \log\left(\prod_{t\in[0,\tau]}\left(\lambda^x(t)\right)^{N^x(dt)}\left(1-d\Lambda^x(t)\right)^{1-N^x(dt)}\right) \\ &= \int_0^\tau \log\lambda^x(t)\, dN^x(t) - \int_0^\tau d\Lambda^x(t).\end{aligned}$$

For this pair of submodel and loss function we have the desired property that,

$$\begin{aligned}\frac{d}{d\varepsilon}\mathcal{L}_x(\Lambda^x(\cdot\,;\varepsilon))\Big|_{\varepsilon=0} &= \frac{d}{d\varepsilon}\left(\int_0^\tau\left(\log\lambda^x(t) + \varepsilon h^{G^*}_t h^x_t\right)dN^x(t) - \right. \\ &\qquad \left.\int_0^\tau \exp\left(\varepsilon h^{G^*}_t h^x_t\right)d\Lambda^x(t)\right)\Big|_{\varepsilon=0} \\ &= \int_0^\tau h^{G^*}_t h^x_t\, dN^x(t) - \int_0^\tau h^{G^*}_t h^x_t\, d\Lambda^x(t) \\ &= \int_0^\tau h^{G^*}_t h^x_t\left(dN^x(t) - d\Lambda^x(t)\right),\end{aligned}$$

is equal to the corresponding terms of the canonical gradient as expressed in (19).

6. Initial estimation of nuisance parameters. To carry out our targeting algorithm, we need initial estimators for the following time-sequence of conditional densities, conditional expectations and intensities,

$$
\begin{aligned}
G &= \Big(\pi_t, \Lambda^c(t) \,:\, t \in [0,\tau]\Big), \qquad \text{and,} \\
\boldsymbol{Z} &= \Big(Z_t^{G^*}, Z_{t,L(t)}^{G^*}, \Lambda^\ell(t), \Lambda^a(t), \Lambda^d(t) \,:\, t \in [0,\tau]\Big).
\end{aligned} \tag{21}
$$

In this section, we describe HAL estimation for each of the quantities displayed in (21). In the present work the aim of this is mainly theoretical; we can prove formal convergence results for the HAL estimator. In Appendix C we provide some preliminaries on implementing our HAL estimators.

Although we focus on describing the HAL estimator, we note that in order to optimize the estimation of the quantities in (21) we would use a super learner that selects the best weighted combination of a prespecified library of candidate algorithms by minimizing the cross-validated empirical risk. See also Table 3 in Appendix C. The super learner performs asymptotically no worse than any algorithm included in its library, a property known as the oracle inequality (van der Laan and Dudoit, 2003; van der Vaart, Dudoit and van der Laan, 2006). Accordingly, by proving the convergence results for the HAL estimator we also prove it for a super learner that includes the HAL estimator in its library.

6.1. *Highly adaptive lasso (HAL).* In the remaining part of this section we describe HAL estimation for the quantities in (21) and state a set of conditions on our model $\mathcal{M}$ that are sufficient to establish asymptotic efficiency of our TMLE estimator. Appendix B provides supplementary details. The key is to restrict attention to the Donsker class of functions that are càdlàg (right-continuous with left limits) with finite variation norm (van der Laan, 2017). We define the Banach space $\mathbb{D}^k([0,\eta])$ of k-variate càdlàg functions on $[0,\eta] \subset \mathbb{R}^k$, $\eta \in \mathbb{R}^k_+$, and the variation norm of $f \in \mathbb{D}^k([0,\eta])$ as (Gill, van der Laan and Wellner, 1995),

$$
\|f\|_{\mathrm{v}} = |f(0)| + \sum_{s \in \mathcal{P}(\{1,\ldots,k\})} \int_{(0_s,\eta_s]} |f(dx_s, 0_{-s})|,
$$

where $\mathcal{P}(\{1,\ldots,k\})$ denotes the power set of $\{1,\ldots,k\}$, $x_s = (x_j \,:\, j \in s)$, $x_{-s} = (x_j \,:\, j \notin s)$ and $x_s \to f(x_s, 0_{-s})$ is the s-specific section of f that sets the coordinates in the complement of the index set s equal to zero. Then we let,

$$
\mathcal{J}_{\mathrm{v},M} = \big\{ f \in \mathbb{D}^k([0,\eta]) \,:\, \|f\|_v \le M \big\},
$$

denote the subset of càdlàg functions with variation norm bounded by a constant $M < \infty$. Any $f \in \mathcal{J}_{\mathrm{v},M}$ admits an integral representation in terms of the measures generated by its s-specific sections (Gill, van der Laan and Wellner, 1995) which is used to define a HAL estimator.

6.2. *Fixed-dimensional representation of the observed data.* It is essential to our analysis that we at any point in time t can represent the observed data $\bar{O}(t)$ in terms of a finite-dimensional vector. For this purpose, we define,

$$\bar{O}_k = \left\{ \left(L_0, s, dN^a(s), A(s), dN^\ell(s), L(s), dN^d(s), dN^c(s) \right) : s \in \{T_j\}_{j=0}^k \right\},$$

where $\{T_k\}_{k=1}^K$ denotes the ordered set of unique event times of one subject. Then $\bar{O}_k \in \mathbb{R}^{kp'+d_0}$ where $p' \in \mathbb{N}$ denotes the dimension of an observation at a monitoring time T_k, i.e.,

$$\left(T_k, dN^a(T_k), A(T_k), dN^\ell(T_k), L(T_k), dN^d(T_k), dN^c(T_k) \right) \in \mathbb{R}^{p'},$$

and $d_0 \in \mathbb{N}$ is the dimension of L_0. Following the notation from the previous Section 6.1, where $x_s = (x_j : j \in s)$ denotes the s-specific coordinates of a vector x, we use $\bar{O}_{k,s}$ to denote the s-specific coordinates of $\bar{O}_k$.

6.3. *Estimation of conditional expectations.* Our targeting algorithm requires initial estimators for the time-sequence of conditional expectations, $Z_t^{G^*}$, $Z_{t,L(t)}^{G^*}$. We will proceed by estimating the conditional density μ_t directly. Then we construct substitution estimators for $Z_t^{G^*}$ and $Z_{t,L(t)}^{G^*}$ based on estimators for $\mu_t, \Lambda^\ell(t), \Lambda^a(t), \Lambda^d(t)$, by evaluating the g-computation formula.

6.4. *Parametrizations and loss functions.* We parametrize the conditional density μ_t of $L(t)$ in terms of a function $f^L(t, \bar{O}(t))$, the conditional distribution π_t of $A(t)$ in terms of $f^A(t, \bar{O}(t))$ and the intensities $\Lambda^x(t)$ in terms of $f^x(t, \bar{O}(t))$, $x = \ell, a, c, d$.

In the following, x runs through $\{L, A, c, a, \ell, d\}$. Let $(O, f^x) \mapsto \mathcal{L}_x(f^x)(O)$ be the log-likelihood loss. We denote by $f_0^x = \operatorname{argmin}_{f^x} P_0 \mathcal{L}_x(f^x)$, the minimizer of the true risk. We define the sum loss function $O \mapsto \mathcal{L}_Q(f^Q)(O)$, where $f^Q = (f^x : x = L, \ell, a, d)$, for the Q-factor of the likelihood, and, equivalently, the sum loss function $O \mapsto \mathcal{L}_G(f^G)(O)$, where $f^G = (f^x : x = A, c)$, for the G-factor of the likelihood.

Assumption 3 (Bounded loss functions). *We assume that the loss functions are uniformly bounded in the sense that* $\sup_{f^x, O} \mathcal{L}_x(f^x)(O) < \infty$ *a.s.*

In addition, we assume that,

$$\sup_{f^G} \frac{\|\mathcal{L}_G(f^G) - \mathcal{L}_G(f_0^G)\|_{P_0}^2}{d_G(f^G, f_0^G)} < \infty, \quad \sup_{f^Q} \frac{\|\mathcal{L}_Q(f^Q) - \mathcal{L}_Q(f_0^Q)\|_{P_0}^2}{d_Q(f^Q, f_0^Q)} < \infty \tag{22}$$

where $d(f, f_0) := P_0\mathcal{L}(f) - P_0\mathcal{L}(f_0)$ *denotes the loss-based dissimilarity measure.*

Assumption 3 guarantees the oracle properties of the cross-validation selector (van der Laan and Dudoit, 2003; van der Vaart, Dudoit and van der Laan, 2006). We note that (22) holds for most common loss functions.

The next assumption is central: It will allow us to define HAL estimators and it provides sufficient conditions for condition 2 of Theorem 2 to hold (see Lemma 2). To formulate the assumption, we recall that, for any subject, $\bar{O}(t)$ only changes at the observed event times $T_1, \ldots, T_{K_t}$. Accordingly, we may now further parametrize f^x in terms of a fixed-dimensional real-valued function $(t, \bar{O}_k) \mapsto f_k^x(t, \bar{O}_k)$,

$$f^x(t, \bar{O}(t)) = \sum_{k=0}^{K} \mathbb{1}\{K_t = k\}\, f_k^x(t, \bar{O}_k). \tag{23}$$

Assumption 4 is formulated for the fixed-dimensional f_k^x, but translates to f^x because the sum in (23) is finite.

ASSUMPTION 4 (Càdlàg and finite variation). *We assume that* $f_k^x \in \mathcal{J}_{\mathrm{v},M^x}$ *for all* $k \le K$ *and all* $x = L, A, c, a, \ell, d$, *that is,* f_k^x *is càdlàg and has variation norm bounded by a constant* $M^x < \infty$.

As detailed in Appendix B.2, the representation of a càdlàg function with finite variation norm becomes a finite sum over indicator basis functions when the function is defined on a discrete support. The HAL estimator for f^x is defined as the minimizer of the empirical risk over all discrete measures with a particular support defined by the actual observations $\{O_i\}_{i=1}^n$.

6.5. *HAL representation and HAL estimation.* To define the HAL estimator for f^x, consider the representation for $f_k^x \in \mathcal{J}_{\mathrm{v},M^x}$ on the support

defined by the n observations $\{O_i\}_{i=1}^n$ as follows:

(24)
$$f^x_{k,\beta}(t,\bar{O}_k) = \sum_{r=0}^{\bar{K}_n} \mathbb{1}\{t_r \leq t\}\left(\beta^x_{r,k,0} + \sum_{s\in\mathcal{P}(\{1,\ldots,kp'+d_0\})}\sum_{j\in\mathcal{I}_{k,s}} \phi^x_{k,s,j}(\bar{O}_k)\beta^x_{r,k,s,j}\right).$$

In this representation, $\mathcal{I}_{k,s}$ is used to denote the index set for the unique observed values of $\bar{O}_{k,s}$. The time-points $t_0 < \cdots < t_{\bar{K}_n}$ are the elements of the ordered unique sequence of all observed times of changes from Display (3). Further, $\beta_{r,k,s,j}$ is the measure that $f^x_{k,\beta}$ assigns to cube j for $t_r \leq t$ and $\phi_{k,s,j}(\bar{O}_k) = \mathbb{1}\{m_{k,s,j} \leq \bar{O}_{k,s}\}$ is the indicator that the support point $m_{k,s,j}$ is smaller than or equal to $\bar{O}_{k,s}$.

Corresponding to the representation $f^x_{k,\beta}$ for f^x_k in (24) we have, by (23), an equivalent representation f^x_β for f^x. The variation norm of the finite sum representation (24) equals the sum of the absolute values of the coefficients. A HAL estimator for each f^x, $x = L, A, \ell, a, d, c$, can now be defined as the solution to the following minimization problem:

(25)
$$\hat{f}^x_n = \underset{\beta}{\operatorname{argmin}}\ \mathbb{P}_n\mathcal{L}_x(f^x_\beta) \qquad \text{s.t.} \quad \|\beta\|_1 = \sum_{r,k,s,j} |\beta^x_{r,k,s,j}| \leq M^x,$$

corresponding to L_1-penalized (Lasso) regression (Tibshirani, 1996) with the indicator functions $\phi^x_{k,s,j}(\bar{O}_k)$ as covariates and $\beta^x_{r,k,s,j}$ as corresponding coefficients. For the log-likelihood loss and the squared error loss functions, standard software can be used to find the estimators (25), including estimation of M^x by cross-validation. See Table 4 in Appendix C.

6.6. *Theoretical results for HAL.* We collect here the theoretical results for HAL estimation. These results rely on Assumptions 3 and 4.

LEMMA 2. *The canonical gradient can be written as* $D^*(P) = D^*(f^x : x = L, A, c, a, \ell, d)$. *If* $d\bar{G}^*_t/d\bar{G}_t$ *is uniformly bounded for all* t, *then*

$$\left\{D^*(f^x : x = L, A, c, a, \ell, d)\right\}$$

is a Donsker class.

PROOF. See Appendix B.3. □

THEOREM 3. *For the given loss functions* $O \mapsto \mathcal{L}_x(f^x)(O)$, *for* $x = Q, G$, *we have that, under Assumptions 3 and 4,*

$$d_G(\hat{f}^G_n, f^G_0) = P_0\mathcal{L}_G(\hat{f}^G_n) - P_0\mathcal{L}_G(f^G_0) = o_P(n^{-1/2}),$$
$$d_Q(\hat{f}^Q_n, f^Q_0) = P_0\mathcal{L}_Q(\hat{f}^Q_n) - P_0\mathcal{L}_Q(f^Q_0) = o_P(n^{-1/2}),$$

from which it follows that $\|\hat{f}_n^G - f_0^G\|_{P_0} = o_P(n^{-1/4})$ *and* $\|\hat{f}_n^Q - f_0^Q\|_{P_0} = o_P(n^{-1/4})$.

PROOF. See Appendix B.4. □

In summary, Theorem 3 combined with Remark 1 implies that using HAL for initial estimation fulfills condition 1 of Theorem 2. Moreover, under our nonparametric smoothness assumptions (Assumption 4), condition 2 of Theorem 2 holds by Lemma 2.

7. Targeting algorithm. Based on the loss functions and submodels defined in Sections 5.1–5.2, we here present a targeting algorithm for updating the collection of initial estimators for $\boldsymbol{Z}_0$ for a given estimator $\hat{G}_n$ for G_0 on $[0, \tau]$. We index the initial estimator for $\boldsymbol{Z}$ by $k = 0$ as follows,

$$\hat{\boldsymbol{Z}}_n^{k=0} = \left(\hat{Z}_{t,k=0}^{G^*}, \hat{Z}_{t,L(t),k=0}^{G^*}, \hat{\Lambda}_{k=0}^{\ell}(t), \hat{\Lambda}_{k=0}^{a}(t), \hat{\Lambda}_{k=0}^{d}(t)\right)_{t\in[0,\tau]}.$$

Our targeting algorithm involves separate updating steps for current estimators $\hat{Z}_{t,L(t),k}^{G^*}$ and $\hat{\Lambda}_k^x(t)$, $x = a, \ell, d$, from k to $k+1$, $k = 0, 1, \ldots$, carried out simultaneously for all time-points. Our algorithm is overall centered around the representation of the target parameter by iterated expectations that we introduced in Section 3, with the separate updating steps ensuring that we solve the individual terms of the efficient influence curve equation. To describe the algorithm, recall that, as introduced in Display (3),

$$0 = t_0 < t_1 < \cdots < t_{\bar{K}_n},$$

denotes the ordered sequence of unique times of changes $\cup_{i=1}^n \{T_{i,k}\}_{k=1}^{K_i}$ across all subjects of the dataset. Evaluated at time t_r, the conditional expectations $Z_t^{G^*}$ and $Z_{t,L(t)}^{G^*}$ from Section 3 can be written:

$$Z_{t_r}^{G^*} = \mathbb{E}_{P^{G^*}}[Y \mid L(t_r), N^{\ell}(t_r), N^a(t_r), N^d(t_r), \mathcal{F}_{t_{r-1}}], \tag{26}$$

$$Z_{t_r,L(t_r)}^{G^*} = \mathbb{E}_{P^{G^*}}[Y \mid N^{\ell}(t_r), N^a(t_r), N^d(t_r), \mathcal{F}_{t_{r-1}}]. \tag{27}$$

Further, at each t_r, we need to estimate the sequence of conditional expectations,

$$\mathbb{E}_{P^{G^*}}[Y \mid N^a(t_r), N^d(t_r), \mathcal{F}_{t_{r-1}}], \tag{28}$$

$$\mathbb{E}_{P^{G^*}}[Y \mid N^d(t_r), \mathcal{F}_{t_{r-1}}], \tag{29}$$

$$\mathbb{E}_{P^{G^*}}[Y \mid \mathcal{F}_{t_{r-1}}]. \tag{30}$$

Based on current estimators $\hat{\boldsymbol{Z}}_n^k$, we construct estimators for (26)–(30) for all $r = 1, \ldots, \bar{K}_n$, which further yield estimators for $\hat{h}^{\ell}_{t_r,k}, \hat{h}^{a}_{t_r,k}, \hat{h}^{d}_{t_r,k}$ for the clever covariates $h^{\ell}_{t_r}, h^{a}_{t_r}, h^{d}_{t_r}$. A more detailed description of this procedure is given in Appendix C.2. Estimators $\hat{h}^{G^*}_{t_r}$, $r = 1, \ldots, \bar{K}_n$, for the clever weights are obtained by substituting the estimator $\hat{G}_n$ for G in (15).

7.1. *Updating the estimator for $Z^{G^*}_{t_r,L(t_r)}$.* The updating step for the time-sequence $(\hat{Z}^{G^*}_{t_r,L(t_r),k+1})_{1\leq r\leq \bar{K}_n}$ of estimators for the conditional expectations (27) is defined according to the loss function and submodel from Section 5.1, given the estimated clever weights $(\hat{h}^{G^*}_{t_r})_{1\leq r\leq \bar{K}_n}$, as:

$$\hat{\bar{Z}}^{G^*}_{\tau,L(\tau),k+1} := \hat{\bar{Z}}^{G^*}_{\tau,L(\tau),k}(\hat{\varepsilon}_n).$$

Here, $\hat{\varepsilon}_n$ is estimated from the data by,

$$\hat{\varepsilon}_n := \operatorname*{argmin}_{\varepsilon} \mathbb{P}_n \, \bar{\mathcal{L}}_{\hat{\bar{Z}}^{G^*}_{\tau,k}}\big(\hat{\bar{Z}}^{G^*}_{t_r,L(t_r),k}(\varepsilon)\big),$$

and the now updated estimators $\hat{\bar{Z}}^{G^*}_{\tau,L(\tau),k+1} = \big(\hat{Z}^{G^*}_{t_r,L(t_r),k+1}\big)_{1\leq r\leq \bar{K}_n}$ solves the desired part of the efficient influence curve equation,

$$\mathbb{P}_n \int_0^{\tau} \hat{h}^{G^*}_t \big(\hat{Z}^{G^*}_{t,k} - \hat{Z}^{G^*}_{t,L(t),k+1}\big)\, dN^{\ell}(t) = 0.$$

Notably, this updating step is carried out only at subject-specific time-points $t_r \in \{T^{\ell}_1, \ldots, T^{\ell}_{N^{\ell}(\tau)}\}$, where changes in $L(t)$ are observed.

7.2. *Updating the estimators for the intensities.* For each of $x = \ell, a, d$, the updating step for the current estimator $\hat{\Lambda}^x_k$ for the intensity Λ^x uses the estimated time-sequence of clever weights $(\hat{h}^{G^*}_{t_r})_{1\leq r\leq \bar{K}_n}$ and the time-sequence $(\hat{h}^{x}_{t_r,k})_{1\leq r\leq \bar{K}_n}$ of current estimators for the relevant clever covariate. Based on the loss function and submodel as defined in Section 5.2, $\hat{\varepsilon}_x$ is now estimated from the observed data by:

$$\hat{\varepsilon}_x := \operatorname*{argmax}_{\varepsilon} \mathbb{P}_n \, \mathcal{L}_x\big(\hat{\Lambda}^x_k(\cdot\,; \varepsilon)\big).$$

We denote the corresponding updated intensity by $\hat{\Lambda}^x_{k+1} = \hat{\Lambda}^x_k(\cdot\,; \hat{\varepsilon}_x)$, which now solves,

$$\mathbb{P}_n \int_0^{\tau} \hat{h}^{G^*}_t \hat{h}^{x}_{t,k} \big(dN^x(t) - d\hat{\Lambda}^x_{k+1}(t)\big) = 0,$$

the equation of interest.

7.3. *Iterating the targeting steps.* The updated estimators for $Z^{G^*}_{t,L(t)}$ and the intensities across time yield updated estimators for the sequence of conditional expectations (26)–(30) and thus for the clever covariates. This process constitutes the targeting iteration from k to $k+1$, corresponding to updating $\hat{\boldsymbol{Z}}^k_n$ into $\hat{\boldsymbol{Z}}^{k+1}_n$. The process is now repeated iteratively for $k = 0, 1, 2, \ldots$, moving from a current collection of estimators $\hat{\boldsymbol{Z}}^k_n$ to an updated collection of estimators $\hat{\boldsymbol{Z}}^{k+1}_n$. At each step k, the efficient score equation is evaluated,

$$\mathbb{P}_n D^*(\hat{\boldsymbol{Z}}^k_n, \hat{G}_n) = \mathbb{P}_n \left(\hat{Z}^{G^*}_{0,k} + \int_0^\tau \hat{h}_t \left(\hat{Z}^{G^*}_{t,k} - \hat{Z}^{G^*}_{t,L(t),k} \right) dN^\ell(t) \right.$$
$$\left. + \sum_{x \in \{a,\ell,d\}} \int_0^\tau \hat{h}_t \, \hat{h}^x_{t,k} \left(dN^x(t) - d\hat{\Lambda}^x_k(t) \right) \right) - \Psi^{G^*}(\hat{\boldsymbol{Z}}^k_n),$$

and the iterations from k to $k+1$ are continued until,

$$| \, \mathbb{P}_n D^*(\hat{\boldsymbol{Z}}^k_n, \hat{G}_n) \, | < s_n,$$

for some choice of stopping criterion $s_n = o_P(n^{-1/2})$. We can for example use $s_n = \sigma/(n^{-1/2} \log n)$, where σ^2 is the variance of the efficient influence function which we can estimate by substituting the current estimators $\hat{\boldsymbol{Z}}^k_n$. Letting $k^* := \min(k : | \, \mathbb{P}_n D^*(\hat{\boldsymbol{Z}}^k_n, \hat{G}_n) \, | < s_n)$, we denote the final estimator by $\hat{\boldsymbol{Z}}^*_n = \hat{\boldsymbol{Z}}^{k^*}_n$, where,

$$\hat{\boldsymbol{Z}}^{k^*}_n = \left(\hat{Z}^{G^*}_{t,k^*}, \hat{Z}^{G^*}_{t,L(t),k^*}, \hat{\Lambda}^\ell_{k^*}(t), \hat{\Lambda}^a_{k^*}(t), \hat{\Lambda}^d_{k^*}(t) \right)_{t \in [0,\tau]}.$$

8. Inference for the targeted substitution estimator. We have shown in Section 6.1 that there exist initial estimators that fulfill the regularity conditions of Theorem 2. In Section 7 we have presented a targeting algorithm that maps the initial estimator $\hat{\boldsymbol{Z}}^{k=0}_n$ into a targeted estimator $\hat{\boldsymbol{Z}}^*_n$ that solves the efficient influence curve equation (14). In the end we estimate the target parameter by,

$$\hat{\psi}^{G^*}_n = \mathbb{P}_n \, \hat{Z}^{G^*}_{t=0,k^*},$$

but we note that $\hat{\psi}^{G^*}_n$ is equal to the substitution estimator $\Psi^{G^*}(\hat{P}^*_n)$, where $\hat{P}^*_n$ is characterized by $(\boldsymbol{Z}^*_n, \hat{G}_n)$. Specifically, the components of $\hat{\boldsymbol{Z}}^*_n$ are compatible with a probability distribution $\hat{P}^*_n$ whose conditional expectations of Y given the relevant histories coincide with those of $\hat{\boldsymbol{Z}}^*_n$.

Theorem 2 implies asymptotic efficiency of $\hat{\psi}^{G^*}_n$, and we can use the asymptotic normal distribution,

$$\sqrt{n} \left(\Psi^{G^*}(\hat{P}^*_n) - \Psi^{G^*}(P_0) \right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, P_0 D^*(P_0)^2),$$

to provide an approximate two-sided confidence interval. Here,

$$\hat{\sigma}_n^2 := \mathbb{P}_n\{D^*(\hat{P}_n^*)\}^2, \tag{31}$$

can be used to estimate the variance of the TMLE estimator.

9. Empirical study. We demonstrate our methods by describing the application to simulated data, using the proposed algorithm to estimate the contrast between a treatment rule that sets $A(t) = 1$ and a treatment rule that sets $A(t) = 0$. One could think of a randomized trial where subjects in the study population are initially randomized to receive treatment or no treatment, but may switch treatment over time depending on the value of time-varying covariates. This is a setting where a standard Cox regression, as mentioned in Section 1.1, does not apply. Our focus is here to confirm our theory by evaluating the targeting algorithm, and, further, to compare our new algorithm to the existing longitudinal TMLE (LTMLE).

We consider a setting where subjects of a population are followed for τ days of follow-up time. On any given day, any subject may change treatment, covariates, may be lost to follow-up (right-censored) or may experience the outcome of interest. Both the treatment and the censoring mechanisms are subject to time-dependent confounding. The data are simulated such that the number of monitoring times per subject are approximately the same across different τ. Thus, the larger τ is, the less events are observed at single monitoring times. Throughout we let $n = 1,000$.

9.1. *Setup.* In all simulations, we generate data from a sequence of logistic regressions allowing time-varying treatment and covariates to affect one another, keeping effects time-constant. Throughout, we let $L_0 \in \{1, \ldots, 6\}$, $A_0 \in \{0, 1\}$, $A(t) \in \{0, 1\}$ and $L(t) \in \{0, 1\}$. We draw observations L_0 such that large values increase the probability of treatment with A_0. Subjects for which $A_0 = 0$ and with current covariate value equal to 1 are more likely to begin treatment by time t. Further, current treatment increases the probability of $L(t) = 1$. Both the baseline treatment A_0 and the time-varying treatment $A(t)$ have a negative main effect on the outcome process $N^d(t)$. Moreover, A_0 has a different effect within levels of $L(t)$, and $L(t)$ has a positive main effect. At last, the censoring process $N^c(t)$ depends on both A_0 and current covariates. Our code for the simulations is available on github (https://github.com/helenecharlotte/continuousTMLE).

9.2. *Parameter of interest.* Our parameter of interest is the contrast between the intervention-specific mean outcomes under the regime that imposes $A(t) = 1$ (subjects adhering to treatment) to the regime that imposes

$A(t) = 0$ (subjects adhering to no treatment). Both regimes also impose no censoring throughout follow-up. Thus, our interventions are specified as follows, following Definition 1:

$$\text{Intervention 0:} \qquad dG_t^0(O) = \big(1 - A(t)\big)^{N^a(dt)}\big(1 - N^c(t)\big),$$
$$\text{Intervention 1:} \qquad dG_t^1(O) = \big(A(t)\big)^{N^a(dt)}\big(1 - N^c(t)\big).$$

We let ψ^1 denote the intervention-specific mean outcome under Intervention 1 and ψ^0 the intervention-specific mean outcome under Intervention 0. Our target parameter is the corresponding difference:

$$\Psi(P) = \int Y dP_{Q,G^1} - \int Y dP_{Q,G^0} = \psi^1 - \psi^0.$$

As usual, the true value is denoted by $\psi_0 = \Psi(P_0)$. There is no unmeasured confounding, so that our parameter can be interpreted causally as the treatment effect we would see in the real world if subjects had between treated $(A(t) = 1)$ compared to not $(A(t) = 0)$. Note that $\psi_0 < 0$ reflects a protecting effect of the treatment. Throughout, we estimate ψ_0^1 and ψ_0^0 separately and report results for the estimated difference $\psi_0 = \psi_0^1 - \psi_0^0$.

9.3. *Simulations.* Overall, we seek to investigate the following:

1. The distribution of $\sqrt{n}\,(\hat{\psi}_n^* - \psi_0)$ under correctly specified parametric models for initial estimation of the nuisance parameters.
2. Comparison to LTMLE estimation when data are given on a discrete grid with small τ.
3. Robustness properties of our TMLE estimator: What happens when we misspecify the distribution of, for example, the outcome process.

We use σ_{ic}^2 to denote the variance of the canonical gradient and $\hat{\sigma}_{ic}^2$ its estimator. We are generally interested in the coverage of confidence intervals based on $\hat{\sigma}_{ic}^2$ and further in the mean squared error (MSE) of our estimator across the simulation repetitions relative to the variance of the canonical gradient, or, equivalently, $\sqrt{MSE}/\sigma_{ic}$.

Comparison to LTMLE. For small τ, we can compare our estimation procedure to the results from the existing LTMLE implementation (Lendle et al., 2017). We note that the interventions that we consider correspond to dynamic rules in the LTMLE framework that only intervene on $A(t)$ at jumps of $N^a(t)$. Table 1 shows the results of estimating the contrast $\psi_0 = \psi_0^1 - \psi_0^0$ when $\tau = 5$ and when $\tau = 50$, using LTMLE and using our algorithm. The table illustrates the increasing bias of LTMLE when τ increases. Moreover,

for the simulations with $\tau = 50$, the estimated standard error from LTMLE is completely off.

LTMLE	$\tau = 5$	$\tau = 30$	$\tau = 50$
ψ_0	-0.135	-0.138	-0.148
$\hat{\psi}^{\text{ltmle}}_{n,M}$	-0.137	-0.133	-0.117
Bias	-0.002	0.005	0.031
Cov (95%)	0.970	0.968	1.000
$\sqrt{\text{MSE}}$	0.044	0.092	0.080
$\hat{\sigma}_{n,M}$	0.049	0.108	0.648

contTMLE	$\tau = 5$	$\tau = 30$	$\tau = 50$
ψ_0	-0.135	-0.138	-0.148
$\hat{\psi}^{*}_{n,M}$	-0.137	-0.138	-0.147
Bias	-0.002	0.000	0.001
Cov (95%)	0.942	0.940	0.942
$\sqrt{\text{MSE}}$	0.041	0.060	0.099
$\hat{\sigma}_{n,M}$	0.039	0.058	0.101

TABLE 1

Presented are results from a simulation study with $\tau = 5$, $\tau = 30$ and $\tau = 50$ days of follow-up. In the left table, results from applying existing LTMLE software are shown. In the right table, results from applying our algorithm (contTMLE) are shown. Note that $\hat{\psi}^{\text{tmle}}_{n,M}$ and $\hat{\sigma}_{n,M}$ denote averages of estimates across the $M = 500$ simulation repetitions. For $\tau = 50$, LTMLE did not run at all for 7% of the simulations.

Performance of estimation and double robustness. We investigate the distribution of $\sqrt{n}\,(\hat{\psi}^{*}_{n} - \psi_0)$ under correctly specified and misspecified parametric models used for the initial estimation. For the latter, we leave out all time-varying variables when estimating the outcome distribution. Table 2 shows the results of estimating the contrast $\psi_0 = \psi^1_0 - \psi^0_0$ with $\tau = 30, 50, 100$ days of follow-up. In the last setting, LTMLE cannot be applied due to too few events at single monitoring times. The results in the table illustrate that our algorithm achieve appropriate coverage across τ, and that the targeting step of our algorithm removes bias from the initial estimation.

9.4. *Conclusions on the empirical findings.* Based on our simulations we draw the following conclusions:

1. The confidence intervals based on the estimate of the efficient influence curve have valid coverage in the simulations when the initial estimation is correctly specified.
2. Our new estimation algorithm produces similar results to the existing LTMLE for the settings where LTMLE applies. When τ gets larger, and thus observation sparsity increases, LTMLE breaks down.
3. Misspecification of the Q-part of the likelihood, such as the distribution of the outcome process, leads to biased initial estimation. This bias is corrected for by our targeting procedure.

Correctly specified initial estimation				Misspecified initial estimation			
	$\tau = 30$	$\tau = 50$	$\tau = 100$		$\tau = 30$	$\tau = 50$	$\tau = 100$
ψ_0	-0.138	-0.148	-0.130	ψ_0	-0.138	-0.148	-0.130
$\hat{\psi}^{\text{init}}_{n,M}$	-0.127	-0.141	-0.129	$\hat{\psi}^{\text{init}}_{n,M}$	-0.086	-0.125	-0.119
$\hat{\psi}^{*}_{n,M}$	-0.138	-0.147	-0.129	$\hat{\psi}^{*}_{n,M}$	-0.138	-0.147	-0.128
Bias (init)	0.011	0.007	0.001	Bias (init)	0.052	0.023	0.011
Bias (tmle)	0.000	0.001	0.001	Bias (tmle)	0.000	0.001	0.002
Cov (95%)	0.940	0.942	0.956	Cov (95%)	0.946	0.956	0.976
$\sqrt{\text{MSE}}$	0.060	0.099	0.032	$\sqrt{\text{MSE}}$	0.060	0.099	0.030
$\hat{\sigma}_{n,M}$	0.058	0.101	0.035	$\hat{\sigma}_{n,M}$	0.058	0.101	0.035

TABLE 2

Presented are results from a simulation study with $\tau = 30$, $\tau = 50$ and $\tau = 100$ days of follow-up. The initial estimator is denoted $\hat{\psi}^{\text{init}}_n$ and the targeted estimator is denoted $\hat{\psi}^{*}_n$. In the left table, results from applying our algorithm with correctly specified initial estimation are shown. In the right table, results from applying our algorithm with misspecified initial estimation are shown. As desired, the targeting algorithm corrects the bias of the initial estimator. Note that $\hat{\psi}^{*}_{n,M}$, $\hat{\psi}^{\text{init}}_{n,M}$ and $\hat{\sigma}_{\text{ic},M}$ denote averages of estimates across the $M = 500$ simulation repetitions.

10. Discussion. The current paper lays the groundwork for targeted minimum loss based estimation of intervention-specific mean outcomes based on a continuous-time counting process model. Our generalization of the TMLE methodology is motivated by the problems arising when estimating interventional effects from observational data, where observations are made without a schedule and potentially over very long time horizons. We have developed our TMLE based on a continuous-time model to cover any time-scale. The advantage is that we can track the information of changes in continuous time, hereby preserving the original time-order of treatment and covariates.

We have derived the efficient influence function for the estimation problem, and we have proposed a particular targeting algorithm to construct an estimator that solves the efficient influence curve equation. Contrary to the existing discrete-time longitudinal TMLE method (LTMLE) that involves a regression step separately for every time-point, even if nothing is observed at that time-point, our proposed TMLE algorithm allows us to smooth information across time. In a sense, the estimation procedure we have presented amounts to doing a sequential regression for each infinitesimal time interval, but simultaneously by pooling over all t.

A substantial advantage of the proposed setting for estimation of interventional effects is that it provides a framework for studying various types of interventions. In this paper, we have focused on interventions on censoring

(Λ^c) and treatment decision (π_t), and in our simulations we considered only an intervention that imposed 'no treatment'. As discussed in Section 1.1 and Section 2.2, other interventions on treatment decision and interventions on the treatment monitoring intensity Λ^a are possible within the same framework: The targeting procedure remains the same, although, for the latter, Λ^a is taken out of the non-interventional part, and thus out of the targeting procedure. In future work we will consider stochastic and optimal interventions (Murphy, 2003, 2005; Hernán et al., 2006; Zhang et al., 2013, 2012) both on the treatment decision and on the treatment monitoring. One intervention of interest could be to replace Λ^a by $\Lambda^{a,*}$ that only depends, for example, on the time since last treatment monitoring to ensure a regime where subjects are monitored regularly.

It is evident from the general analysis of TMLE that the performance of the final estimator depends on how well the nuisance parameters are estimated. In the continuous-time setting, the nuisance parameters comprise regressions of the intensity of changes and regressions of the time-dependent means. We have shown that we can construct HAL estimators for all required components such as to fulfill the conditions needed for asymptotic efficiency. To improve the estimation of nuisance parameters, it is desirable to set up large libraries consisting of flexible models and estimators for the conditional means and intensities that can then be fed into the super learner. In general, one should adapt the choices to the application at hand, considering the length of the time interval, the marginal intensity rates of the changes of action and covariates as well as the sample size and outcome prevalence.

We further point out that our presented method to get $\boldsymbol{Z}$ is to use the conditional density of $L(t)$ to estimate the conditional expectations. However, we are not committed to using the conditional density. The fact is that our targeting algorithm does not depend on μ_t, and it can be a big advantage to avoid estimation of μ_t altogether. In future work we discuss in more detail how to construct an estimator based on inverse probability weighted regression which we map into an estimator that fulfill the representation in terms of iterated expectations from Section 3.

Future work will deal with a general implementation of our TMLE procedure, beyond the simple settings of our simulation study. In addition to the applications outlined in Section 1.1, another important area of application is that of randomized trials where subjects crossover, start additional treatment and drop out. Here our methods can be applied to supplement the intention-to-treat analysis.

References.

Andersen, P. K., Borgan, O., Gill, R. D. and Keiding, N. (1993). *Statistical models based on counting processes*. Springer, New York.

Andersen, J. T., Petersen, M., Jimenez-Solem, E., Broedbaek, K., Andersen, E. W., Andersen, N. L., Afzal, S., Torp-Pedersen, C., Keiding, N. and Poulsen, H. E. (2013). Trimethoprim use in early pregnancy and the risk of miscarriage: a register-based nationwide cohort study. *Epidemiology & Infection* **141** 1749–1755.

Bang, H. and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics* **61** 962–973.

Bickel, P. J., Klaassen, C. A. J., Ritov, Y. and Wellner, J. A. (1993). Efficient and adaptive inference in semiparametric models.

Chakraborty, B. and Moodie, E. E. (2013). *Statistical methods for dynamic treatment regimes*. Springer.

Dawid, A. P. and Didelez, V. (2010). Identifying the consequences of dynamic treatment strategies: A decision-theoretic overview. *Statistics Surveys* **4** 184–231.

Gill, R. D. (1994). Lectures on survival analysis. In *Lectures on Probability Theory* 115–241. Springer.

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

Gill, R. D. and Robins, J. M. (2001). Causal inference for complex longitudinal data: the continuous case. *Annals of Statistics* 1785–1811.

Gill, R. D., van der Laan, M. J. and Wellner, J. A. (1995). *Inefficient estimators of the bivariate survival function for three models* **31**. Annales de l'Institut Henri Poincaré.

Hernán, M. A. (2010). The hazards of hazard ratios. *Epidemiology (Cambridge, Mass.)* **21** 13.

Hernan, M. A. and Robins, J. M. (2020). *Causal Inference*. Chapman & Hall/CRC, Boca Raton, Fl.

Hernán, M. A., Lanoy, E., Costagliola, D. and Robins, J. M. (2006). Comparison of dynamic treatment regimes via inverse probability weighting. *Basic & clinical pharmacology & toxicology* **98** 237–242.

Karim, M. E., Gustafson, P., Petkau, J., Tremlett, H., the Long-Term Benefits and of Beta-Interferon for Multiple Sclerosis (BeAMS) Study Group, A. E. (2016). Comparison of statistical approaches for dealing with immortal time bias in drug effectiveness studies. *American journal of epidemiology* **184** 325–335.

Keiding, N. (1999). Event history analysis and inference from observational epidemiology. *Statistics in Medicine* **18** 2353–2363.

Kessing, L. V., Rytgaard, H. C., Gerds, T. A., Berk, M., Ekstrøm, C. T. and Andersen, P. K. (2019). New drug candidates for depression – a nationwide population-based study. *Acta Psychiatrica Scandinavica* **139** 68–77.

Lendle, S. D., Schwab, J., Petersen, M. L. and van der Laan, M. J. (2017). ltmle: An R Package Implementing Targeted Minimum Loss-Based Estimation for Longitudinal Data. *Journal of Statistical Software* **81** 1–21.

Lok, J. J. (2008). Statistical modeling of causal effects in continuous time. *The Annals of Statistics* **36** 1464–1507.

Martinussen, T., Vansteelandt, S. and Andersen, P. K. (2018). Subtleties in the interpretation of hazard ratios. *arXiv preprint arXiv:1810.09192*.

Murphy, S. A. (2003). Optimal dynamic treatment regimes. *Journal of the Royal Sta-*

tistical Society: Series B (Statistical Methodology) **65** 331–355.

Murphy, S. A. (2005). An experimental design for the development of adaptive treatment strategies. *Statistics in medicine* **24** 1455–1481.

Murphy, S. A., van der Laan, M. J., Robins, J. M. and Group, C. P. P. R. (2001). Marginal mean models for dynamic regimes. *Journal of the American Statistical Association* **96** 1410–1423.

Petersen, M., Schwab, J., Gruber, S., Blaser, N., Schomaker, M. and van der Laan, M. (2014). Targeted maximum likelihood estimation for dynamic and static longitudinal marginal structural working models. *Journal of causal inference* **2** 147–185.

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** 1393–1512.

Robins, J. M. (1987). Addendum to "a new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect". *Computers & Mathematics with Applications* **14** 923–945.

Robins, J. M. (1989a). The analysis of randomized and non-randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies. *Health service research methodology: a focus on AIDS* 113–159.

Robins, J. (1989b). The control of confounding by intermediate variables. *Statistics in medicine* **8** 679–701.

Robins, J. (1992). Estimation of the time-dependent accelerated failure time model in the presence of confounding factors. *Biometrika* **79** 321–334.

Robins, J. M. (1998). Marginal structural models. 1997 proceedings of the American Statistical Association, section on Bayesian statistical science (pp. 1–10). *Retrieved from*.

Robins, J. M. (2000a). Marginal structural models versus structural nested models as tools for causal inference. In *Statistical models in epidemiology, the environment, and clinical trials* 95–133. Springer.

Robins, J. M. (2000b). Robust estimation in sequentially ignorable missing data and causal inference models. In *Proceedings of the American Statistical Association* **1999** 6–10.

Robins, J. M. (2002). Analytic methods for estimating HIV-treatment and cofactor effects. In *Methodological Issues in AIDS Behavioral Research* 213–288. Springer.

Robins, J. M., Hernán, M. A. and Siebert, U. (2004). Effects of multiple interventions. *Comparative quantification of health risks: global and regional burden of disease attributable to selected major risk factors* **1** 2191–2230.

Robins, J., Orellana, L. and Rotnitzky, A. (2008). Estimation and extrapolation of optimal treatment and testing strategies. *Statistics in medicine* **27** 4678–4721.

Røysland, K. (2011). A martingale approach to continuous-time marginal structural models. *Bernoulli* **17** 895–915.

Røysland, K. (2012). Counterfactual analyses with graphical models based on local independence. *The Annals of Statistics* **40** 2162–2194.

Stitelman, O. M., De Gruttola, V. and van der Laan, M. J. (2012). A general implementation of TMLE for longitudinal data applied to causal inference in survival analysis. *The international journal of biostatistics* **8**.

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

Tsiatis, A. (2007). *Semiparametric theory and missing data*. Springer Science & Business Media.

VAN DER LAAN, M. J. (2010a). Targeted maximum likelihood based causal inference: Part I. *The International Journal of Biostatistics* **6**.

VAN DER LAAN, M. J. (2010b). Targeted maximum likelihood based causal inference: Part II. *The international journal of biostatistics* **6**.

VAN DER LAAN, M. J. (2017). A Generally Efficient Targeted Minimum Loss Based Estimator based on the Highly Adaptive Lasso. *The International Journal of Biostatistics* **13**.

VAN DER LAAN, M. J. and DUDOIT, S. (2003). Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator: Finite sample oracle inequalities and examples.

VAN DER LAAN, M. J. and GRUBER, S. (2012). Targeted minimum loss based estimation of causal effects of multiple time point interventions. *The international journal of biostatistics* **8**.

VAN DER LAAN, M. J. and PETERSEN, M. L. (2007). Causal effect models for realistic individualized treatment and intention to treat rules. *The international journal of biostatistics* **3**.

VAN DER LAAN, M. J., POLLEY, E. C. and HUBBARD, A. E. (2007). Super learner. *Statistical applications in genetics and molecular biology* **6**.

VAN DER LAAN, M. J. and ROBINS, J. M. (2003). *Unified methods for censored longitudinal data and causality*. Springer Science & Business Media.

VAN DER LAAN, M. J. and ROSE, S. (2011). *Targeted learning: causal inference for observational and experimental data*. Springer Science & Business Media.

VAN DER LAAN, M. J. and ROSE, S. (2018). *Targeted learning in data science: causal inference for complex longitudinal studies*. Springer.

VAN DER LAAN, M. J. and RUBIN, D. (2006). Targeted maximum likelihood learning. *The International Journal of Biostatistics* **2**.

VAN DER VAART, A. W. (2000). *Asymptotic statistics* **3**. Cambridge university press.

VAN DER VAART, A. W., DUDOIT, S. and VAN DER LAAN, M. J. (2006). Oracle inequalities for multi-fold cross validation. *Statistics & Decisions* **24** 351–371.

VAN DER VAART, A. W. and WELLNER, J. A. (1996). Weak convergence. In *Weak convergence and empirical processes* 16–28. Springer.

ZHANG, B., TSIATIS, A. A., LABER, E. B. and DAVIDIAN, M. (2012). A robust method for estimating optimal treatment regimes. *Biometrics* **68** 1010–1018.

ZHANG, B., TSIATIS, A. A., LABER, E. B. and DAVIDIAN, M. (2013). Robust estimation of optimal dynamic treatment regimes for sequential treatment decisions. *Biometrika* **100** 681–694.

SUPPLEMENTARY MATERIAL

APPENDIX A: ANALYSIS OF THE ESTIMATION PROBLEM

This appendix is concerned with the derivation of the canonical gradient (Bickel et al., 1993; van der Vaart, 2000; Tsiatis, 2007) as presented in Theorem 1.

In the following, we let $\mathcal{M}_G$ be the submodel of $\mathcal{M}$ with known G-part, and we let $\mathscr{T}_Q(P)$ be the tangent space at P in $\mathcal{M}_G$. We derive the canonical gradient of Ψ^{G^*} in the full model $\mathcal{M}$ by projecting the influence curve of any regular and asymptotically linear (RAL) estimator of Ψ^{G^*} in the smaller model $\mathcal{M}_G$ onto $\mathscr{T}_Q(P)$. Indeed, the factorization of any $P \in \mathcal{M}$ into the G-part and the Q-part implies that the tangent space of the G-part in $\mathcal{M}$ is orthogonal to the tangent space of the smaller model $\mathcal{M}_G$ (van der Laan and Robins, 2003).

Noting the following representation of the target parameter, for any $P \in \mathcal{M}$,

$$\begin{aligned}\Psi^{G^*}(P) = \mathbb{E}_{P^{G^*}}\left[Y\right] &= \int y \prod_{t\in[0,\tau]} dQ_t(o)\, dG^*_t(o) = \int y\, d\bar{G}^*_\tau(o)\, \frac{d\bar{G}_\tau(o)}{d\bar{G}_\tau(o)}\, d\bar{Q}_\tau(o) \\ &= \int \left(y\, \frac{d\bar{G}^*_\tau(o)}{d\bar{G}_\tau(o)}\right) d\bar{G}_\tau(o)\, d\bar{Q}_\tau(o) = \mathbb{E}\left[Y\, \frac{d\bar{G}^*_\tau(O)}{d\bar{G}_\tau(O)}\right],\end{aligned}$$

we can estimate $\Psi^{G^*} : \mathcal{M}_G \to \mathbb{R}$ with an inverse probability of action weighted (IPAW) estimator,

$$\hat{\psi}^{G^*}_{n,\mathrm{IPAW}}(o) := \mathbb{P}_n\, Y dG^*_0/dG_0 = \frac{1}{n}\sum_{i=1}^{n} y_i\, d\bar{G}^*_\tau(o_i)/d\bar{G}_\tau(o_i). \tag{32}$$

The IPAW estimator as written in (32) is a sample mean. It follows then directly that it is a linear (and thereby also an asymptotically linear) estimator at any $P \in \mathcal{M}_G$ with influence curve,

$$D^{G^*}_{\mathrm{IPAW}}(P)(O) = \frac{d\bar{G}^*_\tau(O)}{d\bar{G}_\tau(O)} Y - \Psi^{G^*}(P). \tag{33}$$

Since $D^{G^*}_{\mathrm{IPAW}}(P)$ is an influence curve of a RAL estimator of Ψ^{G^*} in $\mathcal{M}_Q$, we can derive the canonical gradient by projecting $D^{G^*}_{\mathrm{IPAW}}(P)$ onto $\mathscr{T}_Q(P)$.

A.1. Tangent space. In the following we characterize the tangent space $\mathscr{T}_Q(P)$. Again we use the fact that the probability distribution of the observed data factorizes. Indeed, the log-likelihood of (4) consists of separate

terms for each of μ_{L_0}, $(\mu_t)_{t\geq 0}$ and Λ^x, $x = a, \ell, d$, and the tangent space $\mathscr{T}_Q$ is thus given as the orthogonal sum of the individual tangent spaces,

$$\mathscr{T}_Q = \mathscr{T}_{\mu_0} \oplus \mathscr{T}_{\mu} \oplus \mathscr{T}_{\Lambda^a} \oplus \mathscr{T}_{\Lambda^\ell} \oplus \mathscr{T}_{\Lambda^d}. \tag{34}$$

In the following, $L_2(P)$ is used to denote the Hilbert space of measurable functions $h : \mathcal{O} \to \mathbb{R}$ with $Ph^2 < \infty$. The following lemmas characterize the relevant tangent spaces. The notation $h \in \sigma(X)$, for some random variable X, means that h is measurable with respect to the σ-algebra generated by X, and can thus be written as a function of X.

LEMMA 3. *The tangent space $\mathscr{T}_{\mu_{L_0}}$ associated with the density μ_{L_0} of L_0 is given by,*

$$\mathscr{T}_{\mu_{L_0}}(P) = \big\{h \in \mathcal{F}_0 \,:\, \mathbb{E}_P\big[h\big] = 0\big\} \cap L_2(P).$$

PROOF. We put no restrictions on the density μ_{L_0}, so the corresponding tangent space $\mathscr{T}_{\mu_{L_0}}$ is the entire Hilbert space of measurable functions of L_0 with mean zero (van der Vaart (2000), Example 25.16). □

LEMMA 4. *The tangent space $\mathscr{T}_{\mu}$ associated with the conditional density μ_t of $L(t)$ is given by,*

$$\mathscr{T}_{\mu}(P) = \left\{ \int_0^\tau h_t \, N^\ell(dt) \,:\, h_t \in \sigma(L(t), \mathcal{F}_{t-}) \wedge \mathbb{E}_P\big[h_t \,\big|\, \mathcal{F}_{t-}\big] = 0 \right\} \cap L_2(P),$$

that is, the tangent space consists of functions that can be represented as stochastic integrals with respect to N^ℓ over functions of $\sigma(L(t), \mathcal{F}_{t-})$ that have mean zero conditional on $\mathcal{F}_{t-}$.

PROOF. Consider the parametric submodel,

$$\mu_{t,\varepsilon,h_t}(o) = \big(1 + \varepsilon h_t(o)\big)\, \mu_t(o),$$

for $\varepsilon \geq 0$, with $h_t : \mathcal{O} \to \mathbb{R}$ such that $h_t \in \sigma(L(t), \mathcal{F}_{t-})$. To ensure that $\mu_{t,\varepsilon,h}$ for all $\varepsilon > 0$ actually gives rise to a well-defined probability measure, note that,

$$\int_{\mathcal{O}} \mu_{t,\varepsilon,h_t} \, d\nu^\ell = \underbrace{\int_{\mathcal{O}} \mu_t \, d\nu^\ell}_{=1} + \varepsilon \int_{\mathcal{O}} h_t \, \mu_t \, d\nu^\ell = 1 \quad \text{if and only if}$$

$$\int_{\mathcal{O}} h_t \, \mu_t \, d\nu^\ell = \mathbb{E}[h_t \,|\mathcal{F}_{t-}] = 0.$$

The contribution to the log-likelihood is,

$$\log dP_{\varepsilon,h_t}(o) = \int_0^\tau \log\Big(\big(1+\varepsilon h_t(o)\big)\,\mu_t(o)\Big)N^\ell(dt),$$

and we derive the score accordingly,

$$\begin{aligned}\frac{d}{d\varepsilon}\log dP_{\varepsilon,h_t}(o)\Big|_{\varepsilon=0} &= \frac{d}{d\varepsilon}\bigg(\int_0^\tau \log\Big(\big(1+\varepsilon h_t(o)\big)\,\mu_t(o)\Big)N^\ell(dt)\bigg)\Big|_{\varepsilon=0}\\ &= \int_0^\tau h_t(o)\,N^\ell(dt).\end{aligned}$$

Thus, the tangent space is given by,

$$\mathscr{T}_\mu(P) = \bigg\{\int_0^\tau h_t\,N^\ell(dt)\ :\ h_t\in\sigma(L(t),\mathcal{F}_{t-})\ \wedge\ \mathbb{E}_P\big[h_t\,\big|\,\mathcal{F}_{t-}\big]=0\bigg\}\cap L_2(P).$$

□

Lemma 5. *The tangent space $\mathscr{T}_{\Lambda^x}$ associated with the intensity Λ^x of N^x is given by,*

$$\mathscr{T}_{\Lambda^x}(P) = \bigg\{\int_0^\tau h_t\,\big(N^x(dt)-\Lambda^x(dt)\big)\ :\ h_t\in\mathcal{F}_{t-}\bigg\},\qquad x=a,\ell,d,$$

that is, the space consisting of stochastic integrals over functions of $\mathcal{F}_{t-}$ with respect to the martingale $M^x = N^x - \Lambda^x$.

Proof. For $x=a,\ell,d$, let $h_t : \mathcal{O}\to\mathbb{R}$ be such that $h_t\in\mathcal{F}_{t-}$ and consider the class of submodels,

$$\lambda^x_{\varepsilon,h_t}(t\,|\,\mathcal{F}_{t-}) = \exp\big(\varepsilon h_t(o)\big)\lambda^x(t\,|\,\mathcal{F}_{t-}).$$

These are valid submodels since the truth is contained at $\varepsilon=0$ and since $\lambda^x_{\varepsilon,h_t}(t\,|\,\mathcal{F}_{t-})>0$ for all t. The contribution to the log-likelihood is,

$$\begin{aligned}\log dP_{\varepsilon,h_t}(o) = \int_0^\tau \big(\varepsilon h_t(o) + \log\lambda^x(t\,|\,\mathcal{F}_{t-})\big)N^x(dt) -\\ \int_0^\tau \exp\big(\varepsilon h_t(o)\big)\,\lambda^x(t\,|\,\mathcal{F}_{t-})dt,\end{aligned}$$

and we derive the score,

$$\begin{aligned}\frac{d}{d\varepsilon}\log dP_{\varepsilon,h_t}(o)\Big|_{\varepsilon=0} &= \frac{d}{d\varepsilon}\bigg(\int_0^\tau \big(\varepsilon h_t(o)+\log\lambda^x(t\,|\,\mathcal{F}_{t-})dt\big)N^x(dt) -\\ &\qquad\qquad \int_0^\tau \exp\big(\varepsilon h_t(o)\big)\,\lambda^x(t\,|\,\mathcal{F}_{t-})dt\bigg)\Big|_{\varepsilon=0}\\ &= \int_0^\tau h_t(o)\,N^x(dt) - \int_0^\tau h_t(o)\,\lambda^x(t\,|\,\mathcal{F}_{t-})dt.\end{aligned}$$

Thus, the tangent space is given by,

$$\mathscr{T}_{\Lambda^x}(P) = \left\{ \int_0^\tau h_t \left(N^x(dt) - \lambda^x(t \,|\, \mathcal{F}_{t-})dt \right) \,:\, h_t \in \mathcal{F}_{t-} \right\},$$

for $x = a, \ell, d$. □

A.2. Projection onto tangent spaces. We denote the projection onto a tangent space $\mathscr{T}$ by $\Pi(\cdot \,|\, \mathscr{T})$. We need to project $D^{G^*}_{\text{IPAW}}(P)$ as defined in (33) onto $\mathscr{T}_Q$. The orthogonal sum decomposition of $\mathscr{T}_Q$ in (34) implies that,

(35)
$$\begin{aligned}\Pi(D^{G^*}_{\text{IPAW}}(P) \,|\, \mathscr{T}_Q(P)) = \Pi(D^{G^*}_{\text{IPAW}}(P) \,|\, \mathscr{T}_{\mu_{L_0}}(P)) + \Pi(D^{G^*}_{\text{IPAW}}(P) \,|\, \mathscr{T}_\mu(P)) \\ + \sum_{x=a,\ell,d} \Pi(D^{G^*}_{\text{IPAW}}(P) \,|\, \mathscr{T}_{\Lambda^x}(P)).\end{aligned}$$

The projections are given in the following lemmas.

LEMMA 6. *The projection of $D^{G^*}_{IPAW}(P)$ onto the tangent space $\mathscr{T}_{\mu_{L_0}}(P)$ is given by,*

$$\Pi(D^{G^*}_{IPAW}(P) \,|\, \mathscr{T}_{\mu_{L_0}}(P)) = \mathbb{E}_{P^{G^*}}[Y \,|\, \mathcal{F}_0] - \Psi^{G^*}(P).$$

PROOF. The projection of $D^{G^*}_{\text{IPAW}}$ onto $\mathscr{T}_{\mu_{L_0}}(P)$ is obtained by,

$$\Pi(D^{G^*}_{\text{IPAW}} \,|\, \mathscr{T}_{\mu_{L_0}}) = \mathbb{E}[D^{G^*}_{\text{IPAW}} \,|\, L_0] - \mathbb{E}[D^{G^*}_{\text{IPAW}}] = \mathbb{E}_{P^{G^*}}[Y \,|\, L_0] - \Psi^{G^*}(P),$$

noting that $\mathcal{F}_0 = \sigma(L_0)$. □

LEMMA 7. *The projection of $D^{G^*}_{IPAW}(P)$ onto the tangent space $\mathscr{T}_\mu(P)$ is given by,*

$$\begin{aligned}&\Pi(D^{G^*}_{IPAW}(P) \,|\, \mathscr{T}_\mu(P)) \\ &\quad = \int_0^\tau \Big(\mathbb{E}[D^{G^*}_{IPAW}(P) \,|\, L(t), \mathcal{F}_{t-}] - \mathbb{E}[D^{G^*}_{IPAW}(P) \,|\, \mathcal{F}_{t-}] \Big) N^\ell(dt).\end{aligned}$$

PROOF. We note that $\Pi\big(D^{G^*}_{\text{IPAW}}(P) \,|\, \mathscr{T}_\mu(P)\big) \in \mathscr{T}_\mu$. What remains to be shown is that,

$$\mathbb{E}\left[\Big(D^{G^*}_{\text{IPAW}}(P) - \Pi\big(D^{G^*}_{\text{IPAW}}(P) \,|\, \mathscr{T}_\mu(P)\big) \Big) S \right] = 0, \tag{36}$$

for any $S \in \mathscr{T}_\mu(P)$. Since $S \in \mathscr{T}_\mu(P)$, we can write,

$$S = \int_0^\tau h_t^S \, N^\ell(dt),$$

for some $h_t^S \in \sigma(L(t), \mathcal{F}_{t-})$ such that $\mathbb{E}_P\big[h_t^S \,\big|\, \mathcal{F}_{t-}\big] = 0$. Now, the second term of the covariance in (36) can be written as,

$$\begin{aligned}
&\mathbb{E}\left[\Pi\big(D^{G^*}_{\text{IPAW}}(P) \,|\, \mathscr{T}_\mu(P)\big)\, S\right] \\
&\quad = \mathbb{E}\left[\int_0^\tau \mathbb{E}\Big[\Big(\mathbb{E}[D^{G^*}_{\text{IPAW}}(P) \,|\, L(t), \mathcal{F}_{t-}] - \mathbb{E}[D^{G^*}_{\text{IPAW}}(P) \,|\, \mathcal{F}_{t-}]\Big)\, h_t^S \,\Big|\, \mathcal{F}_{t-}\Big] N^\ell(dt)\right] \\
&\quad = \mathbb{E}\left[\int_0^\tau \Big(\mathbb{E}\big[\mathbb{E}[h_t^S D^{G^*}_{\text{IPAW}}(P) \,|\, L(t), \mathcal{F}_{t-}]\big] - \right. \\
&\qquad\qquad \left. \mathbb{E}\big[\mathbb{E}[D^{G^*}_{\text{IPAW}}(P) \,|\, \mathcal{F}_{t-}]\, \underbrace{\mathbb{E}[h_t^S \,|\, \mathcal{F}_{t-}]}_{=0}\big]\Big) N^\ell(dt)\right] \\
&\quad = \mathbb{E}\left[\int_0^\tau h_t^S D^{G^*}_{\text{IPAW}}(P)\, N^\ell(dt)\right] = \mathbb{E}\left[D^{G^*}_{\text{IPAW}}(P) S\right],
\end{aligned}$$

where we have used the law of iterated expectations. □

LEMMA 8. *The projection of $D^{G^*}_{IPAW}(P)$ onto the tangent space $\mathscr{T}_{\Lambda^x}(P)$ is given by,*

$$\begin{aligned}
\Pi(D^{G^*}_{IPAW}(P) \,|\, \mathscr{T}_{\Lambda^x}(P)) = \int_0^\tau &\Big(\mathbb{E}[D^{G^*}_{IPAW}(P) \,|\, \Delta N^x(t) = 1, \mathcal{F}_{t-}] - \\
&\mathbb{E}[D^{G^*}_{IPAW}(P) \,|\, \Delta N^x(t) = 0, \mathcal{F}_{t-}]\Big)\big(N^x(dt) - \Lambda^x(dt)\big).
\end{aligned}$$

for $x = a, \ell, d$.

PROOF. We note that $\Pi\big(D^{G^*}_{\text{IPAW}}(P) \,|\, \mathscr{T}_{\Lambda^x}(P)\big) \in \mathscr{T}_{\Lambda^x}$. What remains to be shown is that $D^{G^*}_{\text{IPAW}} - \Pi\big(D^{G^*}_{\text{IPAW}}(P) \,|\, \mathscr{T}_{\Lambda^x}(P)\big)$ is orthogonal to all $S \in \mathscr{T}_{\Lambda^x}$, i.e.,

$$\mathbb{E}\left[\Big(D^{G^*}_{\text{IPAW}}(P) - \Pi\big(D^{G^*}_{\text{IPAW}}(P) \,|\, \mathscr{T}_{\Lambda^x}(P)\big)\Big) S\right] = 0. \tag{37}$$

Since $S \in \mathscr{T}_{\Lambda^x}(P)$, we can write,

$$S = \int_0^\tau h_t^S \, M^x(dt),$$

for some $h_t^S \in \mathcal{F}_{t-}$, and we see that the second term of (37) involves the covariance of martingale integrals. First note that,

$$\begin{aligned}
&\Big(\mathbb{E}[D^{G^*}_{\mathrm{IPAW}}(P) \,|\, \Delta N^x(t) = 1, \mathcal{F}_{t-}] - \\
&\qquad\qquad \mathbb{E}[D^{G^*}_{\mathrm{IPAW}}(P) \,|\, \Delta N^x(t) = 0,\, \mathcal{F}_{t-}]\Big)\Big(N^x(dt) - \Lambda^x(dt)\Big) \\
&= \Big(\mathbb{E}[D^{G^*}_{\mathrm{IPAW}}(P) \,|\, \Delta N^x(t) = 1, \mathcal{F}_{t-}] - \\
&\qquad\qquad \mathbb{E}[D^{G^*}_{\mathrm{IPAW}}(P) \,|\, \Delta N^x(t) = 0,\, \mathcal{F}_{t-}]\Big)\Big(N^x(dt) - \Lambda^x(dt)\Big) \\
&\qquad\qquad + \mathbb{E}[D^{G^*}_{\mathrm{IPAW}}(P) \,|\, \Delta N^x(t) = 0,\, \mathcal{F}_{t-}] \\
&\qquad\qquad - \mathbb{E}[D^{G^*}_{\mathrm{IPAW}}(P) \,|\, \Delta N^x(t) = 0,\, \mathcal{F}_{t-}] \\
&= \Big(\mathbb{E}[D^{G^*}_{\mathrm{IPAW}}(P) \,|\, \Delta N^x(t), \mathcal{F}_{t-}] - \mathbb{E}[D^{G^*}_{\mathrm{IPAW}}(P) \,|\, \mathcal{F}_{t-}]\Big),
\end{aligned}$$

so that,

$$\begin{aligned}
&\mathbb{E}\left[\Pi\big(D^{G^*}_{\mathrm{IPAW}}(P) \,|\, \mathscr{T}_{\Lambda^x}(P)\big) S\right] \\
&\quad= \mathbb{E}\left[\int_0^\tau \Big(\mathbb{E}[D^{G^*}_{\mathrm{IPAW}}(P) \,|\, \Delta N^x(t), \mathcal{F}_{t-}] - \mathbb{E}[D^{G^*}_{\mathrm{IPAW}}(P) \,|\, \mathcal{F}_{t-}]\Big)\, h_t^S\, M^x(dt)\right] \\
&\quad= \int_0^\tau \mathbb{E}\big[\mathbb{E}[D^{G^*}_{\mathrm{IPAW}}(P) \,|\, \Delta N^x(t), \mathcal{F}_{t-}]\, h_t^S\, M^x(dt)\big] - \\
&\qquad\qquad\qquad\qquad \mathbb{E}\Big[\mathbb{E}\big[\mathbb{E}[D^{G^*}_{\mathrm{IPAW}}(P) \,|\, \mathcal{F}_{t-}]\, h_t^S\, M^x(dt) \,\big|\, \mathcal{F}_{t-}\big]\Big] \\
&\quad= \int_0^\tau \mathbb{E}\big[\mathbb{E}[D^{G^*}_{\mathrm{IPAW}}(P)\, h_t^S\, M^x(dt) \,|\, \Delta N^x(t), \mathcal{F}_{t-}]\big] - \\
&\qquad\qquad\qquad\qquad \mathbb{E}\Big[\mathbb{E}[D^{G^*}_{\mathrm{IPAW}}(P) \,|\, \mathcal{F}_{t-}]\, h_t^S\, \underbrace{\mathbb{E}\big[M^x(dt) \,\big|\, \mathcal{F}_{t-}\big]}_{=0}\Big] \\
&\quad= \int_0^\tau \mathbb{E}\big[D^{G^*}_{\mathrm{IPAW}}(P)\, h_t^S\, M^x(dt)\big] = \mathbb{E}\left[\int_0^\tau D^{G^*}_{\mathrm{IPAW}}(P)\, h_t^S\, M^x(dt)\right].
\end{aligned}$$

On the other hand,

$$\begin{aligned}
\mathbb{E}\left[D^{G^*}_{\mathrm{IPAW}}(P)\, S\right] &= \mathbb{E}\left[D^{G^*}_{\mathrm{IPAW}}(P) \int_0^\tau h_t^S\, M^x(dt)\right] \\
&= \mathbb{E}\left[\int_0^\tau D^{G^*}_{\mathrm{IPAW}}(P)\, h_t^S\, M^x(dt)\right],
\end{aligned}$$

thus verifying (37). □

A.3. Proof of Theorem 1.

PROOF. (Theorem 1). Consider first,

$$\begin{aligned}
\mathbb{E}\left[\frac{d\bar{G}^*_\tau}{d\bar{G}_\tau} Y \,\middle|\, \mathcal{F}_{t-}\right] &= \mathbb{E}\left[\left(\prod_{s<t} \frac{dG^*_s}{dG_s}\right)\left(\prod_{s\geq t} \frac{dG^*_s}{dG_s}\right) Y \,\middle|\, \mathcal{F}_{t-}\right] \\
&= \prod_{s<t} \frac{dG^*_s}{dG_s}\, \mathbb{E}\left[\prod_{s\geq t} \frac{dG^*_s}{dG_s} Y \,\middle|\, \mathcal{F}_{t-}\right] \\
&= \prod_{s<t} \frac{dG^*_s}{dG_s} \int Y \prod_{s\geq t} \frac{dG^*_s}{dG_s} \prod_{s\geq t} dQ_s dG_s \\
&= \prod_{s<t} \frac{dG^*_s}{dG_s} \int Y \prod_{s\geq t} dG^*_s dQ_s \\
&= \prod_{s<t} \frac{dG^*_s}{dG_s}\, \mathbb{E}_{P^{G^*}}\left[Y \,\middle|\, \mathcal{F}_{t-}\right].
\end{aligned}$$

To derive the canonical gradient, we sum the projections of $D^{G^*}_{\text{IPAW}}(P)$ onto the individual tangent spaces as stated in (35). The expression in the previous display together with Lemma 6–8 finishes the proof. □

A.4. Corollary 1. The following corollary provides an alternative representation of the canonical gradient, that is utilized to analyze the estimation problem in Section A.5. Another utility of this representation is the fairly direct resemblance to the discrete time counterpart.

COROLLARY 1 (Canonical gradient). *We can rewrite the canonical gradient from Theorem 1 for* $\Psi^{G^*} : \mathcal{M} \to \mathbb{R}$ *as follows,*

$$D^*(P) = \mathbb{E}_{P^{G^*}}[Y \mid \mathcal{F}_0] - \Psi^{G^*}(P) + \int_0^\tau \left(\prod_{s<t} \frac{dG^*_s}{dG_s}\right) Z^{G^*}(dt),$$

where $Z^{G^*}(dt) := \mathbb{E}_{P^{G^*}}[Y \mid L(t), N^\ell(dt), N^a(dt), N^d(dt), \mathcal{F}_{t-}] - \mathbb{E}_{P^{G^*}}[Y \mid \mathcal{F}_{t-}]$.

PROOF.

$$
\begin{aligned}
&\mathbb{E}_{PG^*}[Y \mid L(t), N^\ell(dt), N^a(dt), N^d(dt), \mathcal{F}_{t-}] - \mathbb{E}_{PG^*}[Y \mid \mathcal{F}_{t-}] \\
&\quad = \mathbb{E}_{PG^*}[Y \mid N^\ell(dt)L(t), N^\ell(dt), N^a(dt), N^d(dt), \mathcal{F}_{t-}] - \\
&\qquad\qquad \mathbb{E}_{PG^*}[Y \mid N^\ell(dt), N^a(dt), N^d(dt), \mathcal{F}_{t-}] \\
&\qquad + \mathbb{E}_{PG^*}[Y \mid N^\ell(dt), N^a(dt), N^d(dt), \mathcal{F}_{t-}] - \mathbb{E}_{PG^*}[Y \mid N^a(dt), N^d(dt), \mathcal{F}_{t-}] \\
&\qquad + \mathbb{E}_{PG^*}[Y \mid N^a(dt), N^d(dt), \mathcal{F}_{t-}] - \mathbb{E}_{PG^*}[Y \mid N^d(dt), \mathcal{F}_{t-}] \\
&\qquad + \mathbb{E}_{PG^*}[Y \mid N^d(dt), \mathcal{F}_{t-}] - \mathbb{E}_{PG^*}[Y \mid \mathcal{F}_{t-}] \\
&\quad = \Big(N^\ell(dt) \Big(\mathbb{E}_{PG^*}[Y \mid N^\ell(dt)L(t), N^\ell(dt), N^a(dt), N^d(dt), \mathcal{F}_{t-}] \\
&\qquad\qquad - \ \mathbb{E}_{PG^*}[Y \mid N^\ell(dt), N^a(dt), N^d(dt), \mathcal{F}_{t-}] \Big) \\
&\qquad + \Big(N^\ell(dt) - \Lambda^\ell(dt) \Big) \Big(\mathbb{E}_{PG^*}[Y \mid N^\ell(dt) = 1, N^a(dt), N^d(dt), \mathcal{F}_{t-}] \\
&\qquad\qquad - \ \mathbb{E}_{PG^*}[Y \mid N^\ell(dt) = 0, N^a(dt), N^d(dt), \mathcal{F}_{t-}] \Big) \\
&\qquad + \Big(N^a(dt) - \Lambda^a(dt) \Big) \Big(\mathbb{E}_{PG^*}[Y \mid N^a(dt) = 1, N^d(dt), \mathcal{F}_{t-}] \\
&\qquad\qquad - \ \mathbb{E}_{PG^*}[Y \mid N^a(dt) = 0, N^d(dt), \mathcal{F}_{t-}] \Big) \Big) \\
&\qquad + \Big(N^d(dt) - \Lambda^d(dt) \Big) \Big(1 - \mathbb{E}_{PG^*}[Y \mid N^d(dt), \mathcal{F}_{t-}] \Big).
\end{aligned}
$$

This gives the integral representation. □

A.5. Representation of the second-order remainder $R_2(P, P_0)$. We use $\bar{\mathcal{O}}_{t-}$ to denote the space where $\bar{O}(t-)$ takes its values and $\underline{\mathcal{O}}_t$ the space where $\underline{O}(t) = \{O(s) : t \le s \le \tau\}$ takes its values. We note that,

$$
\begin{aligned}
\mathbb{E}_{PG^*}\big[Y \mid L(t), N^\ell(dt), N^a(dt), N^d(dt), \mathcal{F}_{t-}\big] &= \int_{\underline{\mathcal{O}}_t} Y \prod_{s \ge t} dG^*_s \prod_{s > t} dQ_s, \\
\mathbb{E}_{PG^*}\big[Y \mid \mathcal{F}_{t-}\big] &= \int_{\underline{\mathcal{O}}_t} Y \prod_{s \ge t} dG^*_s \, dQ_s,
\end{aligned}
$$

so that we can write out $P_0 D^*(P)$ for any $P \in \mathcal{M}$ using the representation of the canonical gradient in Corollary 1 as follows,

$$
\begin{aligned}
&\int_{\mathcal{O}} D^*(P)\, dP_0 \\
&\quad = \int_{\mathcal{O}} \int_0^\tau \frac{d\bar{G}_t^*}{d\bar{G}_t} \Big(\mathbb{E}_{PG^*}\left[Y \mid L(t), N^\ell(dt), N^a(dt), N^d(dt), \mathcal{F}_{t-} \right] - \mathbb{E}_{PG^*}\left[Y \mid \mathcal{F}_{t-} \right] \Big)\, dP_0 \\
&\quad = \int_0^\tau \int_{\mathcal{O}} \frac{d\bar{G}_t^*}{d\bar{G}_t} \Bigg(Y \prod_{s \geq t} dG_s^* \prod_{s>t} dQ_s \prod_{s<t} dG_{0,s} \prod_{s \leq t} dQ_{0,s} - \\
&\qquad\qquad Y \prod_{s \geq t} dG_s^*\, dQ_s \prod_{s<t} dG_{0,s}\, dQ_{0,s} \Bigg) \\
&\quad = \int_0^\tau \int_{\mathcal{O}} Y \frac{d\bar{G}_t^*}{d\bar{G}_t} \prod_{s<t} dG_{0,s} dQ_{0,s} \left(dQ_{0,t} - dQ_t \right) \prod_{s \geq t} dG_s^* \prod_{s>t} dQ_s \\
&\quad = \int_0^\tau \int_{\mathcal{O}} Y \frac{d\bar{G}_{0,t}}{d\bar{G}_t} \prod_{s<t} dG_s^* dQ_{0,s} \left(dQ_{0,t} - dQ_t \right) \prod_{s \geq t} dG_s^* \prod_{s>t} dQ_s \\
&(*1) \\
&\quad = \int_0^\tau \int_{\mathcal{O}} Y \left(\frac{d\bar{G}_{0,t}}{d\bar{G}_t} - 1 \right) \prod_{s \leq \tau} dG_s^* \prod_{s<t} dQ_{0,s} \left(dQ_{0,t} - dQ_t \right) \prod_{s>t} dQ_s \\
&(*2) \\
&\qquad + \int_0^\tau \int_{\mathcal{O}} Y \prod_{s \leq \tau} dG_s^* \prod_{s<t} dQ_{0,s} \left(dQ_{0,t} - dQ_t \right) \prod_{s>t} dQ_s \\
&\quad = \underbrace{R_2(P, P_0)}_{(*1)} + \underbrace{\Psi^{G^*}(P_0) - \Psi^{G^*}(P)}_{(*2)}.
\end{aligned}
$$

The relation $(*2) = \Psi^{G^*}(P_0) - \Psi^{G^*}(P)$ follows from the Duhamel equation (Andersen et al., 1993),

$$\begin{aligned}
\Psi^{G^*}(P_0) - \Psi^{G^*}(P) &= \mathbb{E}_{P_0^{G^*}}\big[Y\big] - \mathbb{E}_{P^{G^*}}\big[Y\big] \\
&= \int_{\mathcal{O}} Y \prod_{s\in[0,\tau]} dG_s^*\, dQ_{0,s} - \int_{\mathcal{O}} Y \prod_{s\in[0,\tau]} dG_s^*\, dQ_s \\
&= \int_{\mathcal{O}} Y \prod_{s\in[0,\tau]} dG_s^* \left(\prod_{s\in[0,\tau]} dQ_{0,s} - \prod_{s\in[0,\tau]} dQ_s \right) \\
&= \int_{\mathcal{O}} Y \prod_{s\in[0,\tau]} dG_s^* \left(\int_0^\tau \prod_{s\in[0,t)} dQ_{0,s} \left(dQ_{0,t} - dQ_t\right) \prod_{s\in(t,\tau]} dQ_s \right) \\
&= \int_0^\tau \int_{\mathcal{O}} Y \prod_{s\in[0,\tau]} dG_s^* \prod_{s\in[0,t)} dQ_{0,s} \left(dQ_{0,t} - dQ_t\right) \prod_{s\in(t,\tau]} dQ_s.
\end{aligned}$$

This now implies that,

$$\Psi^{G^*}(P) - \Psi^{G^*}(P_0) = -P_0 D^*(P) + R_2(P, P_0), \tag{38}$$

with the second-order remainder $R_2(P, P_0)$ given by $(*1)$ above,

$$\begin{aligned}
&R_2(P, P_0) \\
&= \int_0^\tau \int_{\mathcal{O}} Y \left(\frac{d\bar{G}_{0,t}}{d\bar{G}_t} - 1 \right) \prod_{s\le\tau} dG_s^* \prod_{s<t} dQ_{0,s} \left(dQ_{0,t} - dQ_t\right) \prod_{s>t} dQ_s \\
&= \int_0^\tau \int_{\mathcal{O}} Y \frac{1}{\bar{g}_t} (\bar{g}_{0,t} - \bar{g}_t) \prod_{s\le\tau} dG_s^* \prod_{s<t} dQ_{0,s} \left(dQ_{0,t} - dQ_t\right) \prod_{s>t} dQ_s.
\end{aligned}$$

A.6. Proof of Lemma 1.

Proof. Our target parameter is:

$$\Psi^{G^*}(P) = \int Y \prod_{t\le\tau} dQ_t\, dG_t^*,$$

Recall that the product integral is defined as the limit of the product of $d\Lambda$ and densities over small intervals forming a partition of $(0, \tau]$, $0 = t_0 < t_1 < \cdots < t_M = \tau$, $\max_m |t_m - t_{m-1}| \to 0$ (Andersen et al., 1993, Section II.6). We will here use the notation dG_{t_m} and dQ_{t_m} to refer to the measure over $(t_{m-1}, t_m]$. Now we can write:

$$\Psi^{G^*}(P) = \lim_{\max_m |t_m - t_{m-1}| \to 0} \int Y \prod_{0=t_0<t_1<\cdots<t_M} dQ_t\, dG_t^*,$$

By Fubini's theorem, and rearranging the order of integrating, we can rewrite this in terms of iterated integrals. First,

$$\Psi^{G^*}(P) = \lim_{\max_m |t_m - t_{m-1}| \to 0} \int \left(\int Y dQ_{t_M} \right) \prod_{t \leq t_{M-1}} dQ_t \, dG^*_t.$$

The innermost integral is the expectation of Y with respect to conditional distribution dQ_{t_M} of the non-interventional part, conditional on the past $\mathcal{F}_{t_{M-1}}$. Accordingly, we can write:

$$\Psi^{G^*}(P) = \lim_{\max_m |t_m - t_{m-1}| \to 0} \int \mathbb{E}_{P^{G^*}}[Y \mid \mathcal{F}_{t_{M-1}}] \prod_{t \leq t_{M-1}} dQ_t \, dG^*_t.$$

We again rewrite this by use of Fubini's theorem:

$$\begin{aligned} &\Psi^{G^*}(P) \\ &= \lim_{\max_m |t_m - t_{m-1}| \to 0} \int \left(\int \mathbb{E}_{P^{G^*}}[Y \mid \mathcal{F}_{t_{M-1}}] \, dG^*_{t_{M-1}} \right) \prod_{t \leq t_{M-1}} dQ_t \prod_{t \leq t_{M-2}} dG^*_t. \end{aligned}$$

Here we note that the innermost integral is:

$$Z^{G^*}_{t_{M-1}} = \mathbb{E}_{P^{G^*}}[Y \mid L(t_{M-1}), N^\ell(t_{M-1}), N^a(t_{M-1}), N^d(t_{M-1}), \mathcal{F}_{t_{M-2}}].$$

In a similar manner, we next write $\Psi^{G^*}(P)$ as an integral over,

$$Z^{G^*}_{t_{M-1}, L(t_{M-1})} = \mathbb{E}_{P^{G^*}}[Y \mid N^\ell(t_{M-1}), N^a(t_{M-1}), N^d(t_{M-1}), \mathcal{F}_{t_{M-2}}],$$

and next as an integral over $\mathbb{E}_{P^{G^*}}[Y \mid \mathcal{F}_{t_{M-2}}]$, i.e.,

$$\Psi^{G^*}(P) = \lim_{\max_m |t_m - t_{m-1}| \to 0} \int \mathbb{E}_{P^{G^*}}[Y \mid \mathcal{F}_{t_{M-2}}] \prod_{t \leq t_{M-2}} dQ_t \, dG^*_t.$$

Here we have integrated out each non-interventional component over the interval $(t_{M-2}, t_{M-1}]$ according to its conditional distribution, which, for the monitoring process N^x, is uniquely characterized by the intensity measure Λ^x. By backwards induction, applying the process above for all the intervals backwards through time, we last reach the final expectation over the marginal distribution of L_0, corresponding to the desired intervention-specific mean outcome $\Psi(P) = Z^{G^*}_{0, L(0)}$. □

APPENDIX B: HIGHLY ADAPTIVE LASSO (HAL)

In Section 6.1 we presented a particular parametrization of the interventional part, the non-interventional intensities and the conditional expectations. In the following, we first restate our parametrizations of likelihood components (Section B.1), next we review the integral representation of càdlàg functions and the HAL estimator (Section B.2), then we prove Lemma 2 (Section B.3) and finally we present the HAL proof that shows that the HAL estimator has the right rate of convergence (Section B.4).

B.1. Parametrizations and loss functions. We here elaborate on the parametrizations used in Section 6.4. Recall the notation,

$$\bar{O}_k = \big\{\big(L_0, s, dN^a(s), A(s), dN^\ell(s), L(s), dN^d(s), dN^c(s)\big) : s \in \{T_j\}_{j=0}^k\big\}.$$

We can use a hazard rate factorization of the conditional distribution of $A(t)$, and further parametrize it in terms of a function f^A,

$$\lambda^A(a \,|\, A(t) \geq a, \bar{O}(t)) = \operatorname{expit}(f^A(a, t, \bar{O}(t))), \quad a \in \mathcal{A},$$

which we represent in terms of a sum over functions $(a, t, \bar{O}_k) \mapsto f_k^A(a, t, \bar{O}_k)$ with values in $\mathbb{R}$,

$$f^A(a, t, \bar{O}(t)) = \sum_{k=0}^{K} \mathbb{1}\{K_t = k\}\, f_k^A(a, t, \bar{O}_k). \tag{39}$$

To keep the notation simpler here, we assume that $L(t)$ is discrete-valued as well and use an equivalent parametrization:

$$\lambda^L(\ell \,|\, L(t) \geq \ell, \bar{O}(t)) = \operatorname{expit}(f^L(\ell, t, \bar{O}(t))),$$

which we represent in terms of a sum over functions $(\ell, t, \bar{O}_k) \mapsto f_k^L(\ell, t, \bar{O}_k)$ with values in $\mathbb{R}$,

$$f^L(\ell, t, \bar{O}(t)) = \sum_{k=0}^{K} \mathbb{1}\{K_t = k\}\, f_k^L(\ell, t, \bar{O}_k). \tag{40}$$

For the absolutely continuous case, we parametrize the intensity process λ^x for which $\Lambda^x(t \,|\, \mathcal{F}_{t-}) = \int_0^t \lambda^x(s \,|\, \mathcal{F}_{s-})ds$ for each $x = \ell, a, d, c$ as follows,

$$\lambda^x(t \,|\, \bar{O}(t)) = \exp(f^x(t, \bar{O}(t))),$$

where,

$$f^x(t, \bar{O}(t)) = \sum_{k=0}^{K} \mathbb{1}\{K_t = k\}\, f_k^x(t, \bar{O}_k).$$

B.2. Representation of càdlàg functions with finite variation norm. As stated in Section 6.1, we define the variation norm of $f \in \mathbb{D}^k([0,\eta])$,

$$\|f\|_{\mathrm{v}} = |f(0)| + \sum_{s \in \mathcal{P}(\{1,\ldots,k\})} \int_{(0_s,\eta_s]} |f(dx_s, 0_{-s})|,$$

where $x_s = (x_j : j \in s)$, $x_{-s} = (x_j : j \notin s)$ and $x_s \to f(x_s, 0_{-s})$ is the s-specific section of f that sets the coordinates in the complement of the index set s equal to zero. If the variation norm is finite, $\|f\|_{\mathrm{v}} < \infty$, then f admits the representation (Gill, van der Laan and Wellner, 1995),

$$f(x) = f(0) + \sum_{s \in \mathcal{P}(\{1,\ldots,k\})} \int_{(0_s,x_s]} f(du_s, 0_{-s}),$$

where $f(du_s, 0_{-s})$ is the measure generated by the càdlàg section function $u_s \to f(u_s, 0_{-s})$. Now, particularly, Assumption 4 implies that f_k^x as a function of $\boldsymbol{w}_k = (t, \bar{O}_k)$ admits the representation,

$$f_k^x(\boldsymbol{w}_k) = f_k^x(0) + \sum_{s \in \mathcal{P}(\{1,\ldots,kp'+d_0\})} \int_{(0_s,\boldsymbol{w}_{k,s}]} f_k^x(du_s, 0_{-s}).$$

B.2.1. *Finite sum representation over indicator basis functions.* Consider an approximation of f_k^x by $f_{k,h}^x$ with a finite support. The above integral representation of $f_{k,h}^x \in \mathcal{J}_{\mathrm{v},M}$ can be written in terms of a finite linear combination of indicator basis functions,

$$f_{k,h}^x(\boldsymbol{w}_k) = \tilde{\beta}_{k,0}^x + \sum_{s \in \mathcal{P}(\{1,\ldots,kp'+d_0\})} \sum_{j \in \tilde{\mathcal{I}}_{k,h,s}} \tilde{\phi}_{k,s,j}^x(\boldsymbol{w}_k) \tilde{\beta}_{k,s,j}^x. \tag{41}$$

Here we have used $\tilde{\mathcal{I}}_{k,h,s}$ to denote the index set of the support points of the sth section function. Moreover, $\tilde{\beta}_{k,s,j}$ is the measure that $f_{k,h}^x$ assigns to the cube defined by j and $\tilde{\phi}_{k,s,j}(\boldsymbol{w}_k) = \mathbb{1}\{\tilde{m}_{h,s,j} \leq \boldsymbol{w}_{k,s}\}$ is the indicator that the support point $\tilde{m}_{h,s,j}$ is smaller than or equal to $\boldsymbol{w}_{k,s}$. Note that the variation norm of $f_{k,h}^x$ is the sum of the absolute values of its coefficients:

$$\|f_{k,h}^x\|_{\mathrm{v}} = \|\beta\|_1 = |\tilde{\beta}_{k,0}^x| + \sum_{s \in \mathcal{P}(\{1,\ldots,kp'+d_0\})} \sum_{j \in \tilde{\mathcal{I}}_{k,h,s}} |\tilde{\beta}_{k,s,j}^x|. \tag{42}$$

A finite sum representation like (41) can be used to approximate any $f_k^x \in \mathcal{J}_{\mathrm{v},M^x}$, as long as the discretization is chosen fine enough (van der Laan, 2017) and yields the corresponding discrete sum representation of f^x:

$$f_h^x(t, \bar{O}(t)) = \sum_{k=0}^{K} \mathbb{1}\{K_t = k\}\, f_{k,h}^x(t, \bar{O}_k). \tag{43}$$

B.2.2. *The HAL estimator.* The HAL estimator is defined by minimizing the empirical risk over all linear combinations of indicator basis function for a specific set of support points under the constraint that the variation norm is bounded by the constant M^x:

$$\hat{f}_n^x = \underset{f_h^x, \|f_h^x\|_v \leq M^x}{\operatorname{argmin}} \mathbb{P}_n \mathcal{L}_x(f_h^x). \tag{44}$$

By selecting the particular support defined by the actual n observations $\{O_i\}_{i=1}^n$, the minimizer of the finite-dimensional minimization problem (44) equals the minimizer of the empirical risk over all functions f^x with variation norm smaller than M^x.

Separating the representation of $f_{k,h}^x$ in (41) over this support defined by $\{O_i\}_{i=1}^n$ into terms involving the time axis and terms involving $\bar{O}_k$ gives (24) in Section 6.5:

$$f_{k,\beta}^x(t, \bar{O}_k) = \sum_{r=0}^{\bar{K}_n} \mathbb{1}\{t_r \leq t\} \Bigg(\beta_{r,k,0}^x + \sum_{s \in \mathcal{P}(\{1,\ldots,kp'+d_0\})} \sum_{j \in \mathcal{I}_{k,s}} \phi_{k,s,j}^x(\bar{O}_k) \beta_{r,k,s,j}^x \Bigg),$$

where, as in Section 6.5, $\mathcal{I}_{k,s}$ is the index set for the unique observed values of $\bar{O}_{k,s}$, the s-specific coordinates $\bar{O}_k$, $t_0 < \cdots < t_{\bar{K}_n}$ is the ordered sequence of unique times of changes from Display (3), $\beta_{r,k,s,j}$ is the measure that $f_{k,\beta}^x$ assigns to cube j for $t_r \leq t$ and $\phi_{k,s,j}(\bar{O}_k) = \mathbb{1}\{m_{k,s,j} \leq \bar{O}_{k,s}\}$ is the indicator that the support point $m_{k,s,j}$ is smaller than or equal to $\bar{O}_{k,s}$.

According to (42), we replace the constraint $\|f_h^x\|_v \leq M^x$ in (44) by $\|f_h^x\|_1 \leq M^x$, so that we can now define the HAL estimator by,

$$\hat{f}_n^x = \underset{f_\beta^x, \|\beta\|_1 \leq M^x}{\operatorname{argmin}} \mathbb{P}_n \mathcal{L}_x(f_\beta^x),$$

corresponding to (25) in Section 6.5.

B.3. Donsker class conditions. We here provide the proof of Lemma 2. Subsequently, we state and prove a Lemma 9 needed for the HAL proof. To reduce complexity of the presentation as much as possible, we write up the formulas for binary $A(t)$ and $L(t)$.

PROOF. (Lemma 2).
Consider the following representation of the efficient influence curve,

$$D^*(P) = \int_0^\tau \left(\prod_{s<t} \frac{dG_s^*}{dG_s} \right) \left(\int Y \prod_{s>t} dQ_s \prod_{s \geq t} dG_s^* - \int Y \prod_{s \geq t} dQ_s dG_s^* \right).$$

First, we demonstrate that we can represent $d\bar{G}_t = \prod_{s<t} dG_s$ in terms of $f^G = (f^x : x = A, c)$:

$$d\bar{G}_t = \prod_{k=0}^{K_t} \Big(\big(\text{expit}(f_k^A(1, T_k, \bar{O}_k)\big)^{A(T_k)} \big(1 - \text{expit}(f_k^A(1, T_k, \bar{O}_k))\big)^{1-A(T_k)} \Big)^{\Delta N^a(T_k)} \Big(\exp(f_k^c(T_k, \bar{O}_k)\big)^{\Delta N^c(T_k)} \Big)$$
$$\exp\Big(- \sum_{r=1}^{\bar{K}_n} \sum_{k=0}^{K_t} \mathbb{1}\{T_k \in [t_{r-1}, t_r)\} \big(\min(t_r, T_{k+1}, t) - T_k \big) \exp(f_k^c(T_k, \bar{O}_k)) - \sum_{r=1}^{\bar{K}_n} \sum_{k=0}^{K_t} \mathbb{1}\{t_r \in (T_k, T_{k+1}]\} \big(\min(t_{r+1}, T_{k+1}, t) - t_r \big) \exp(f_k^c(t_r, \bar{O}_k)) \Big),$$

and, in the same way, we represent $\prod_{s \geq t} dG_s$ as:

$$\prod_{k=K_t+1}^{K} \Big(\big(\text{expit}(f_k^A(1, T_k, \bar{O}_k)\big)^{A(T_k)} \big(1 - \text{expit}(f_k^A(1, T_k, \bar{O}_k))\big)^{1-A(T_k)} \Big)^{\Delta N^a(T_k)} \Big(\exp(f_k^c(T_k, \bar{O}_k)\big)^{\Delta N^c(T_k)} \Big)$$
$$\exp\Big(- \sum_{r=1}^{\bar{K}_n} \sum_{k=K_t}^{K} \mathbb{1}\{T_k \in [t_{r-1}, t_r)\} \big(\min(t_r, T_{k+1}) - \max(T_k, t) \big) \exp(f_k^c(T_k, \bar{O}_k)) - \sum_{r=1}^{\bar{K}_n} \sum_{k=K_t}^{K} \mathbb{1}\{t_r \in (T_k, T_{k+1}]\} \big(\min(t_{r+1}, T_{k+1}) - \max(t_r, t) \big) \exp(f_k^c(t_r, \bar{O}_k)) \Big).$$

We have equivalent versions of $d\bar{G}_t^* = \prod_{s<t} dG_s^*$ and $\prod_{s \geq t} dG_s^*$.

Similarly, we demonstrate that $\prod_{s>t} dQ_s$ can be represented in terms of $f^Q = (f^x : x = L, \ell, a, d)$:

$$\prod_{k=K_t+1}^{K} \Big(\big(\text{expit}(f_k^L(1, T_k, \bar{O}_k)\big)^{L(T_k)} \big(1 - \text{expit}(f_k^L(1, T_k, \bar{O}_k))\big)^{1-L(T_k)} \Big)^{\Delta N^\ell(T_k)} \big(\exp(f_k^\ell(T_k, \bar{O}_k))\big)^{\Delta N^\ell(T_k)} \big(\exp(f_k^a(T_k, \bar{O}_k))\big)^{\Delta N^a(T_k)} \big(\exp(f_k^d(T_k, \bar{O}_k))\big)^{\Delta N^d(T_k)} \Big)$$
$$\prod_{x \in \{\ell, a, d\}} \exp\Big(- \sum_{r=1}^{\bar{K}_n} \sum_{k=K_t}^{K} \mathbb{1}\{T_k \in [t_{r-1}, t_r)\} \big(\min(t_r, T_{k+1}) - \max(T_k, t) \big) \exp(f_k^x(T_k, \bar{O}_k)) - \sum_{r=1}^{\bar{K}_n} \sum_{k=K_t}^{K} \mathbb{1}\{t_r \in (T_k, T_{k+1}]\} \big(\min(t_{r+1}, T_{k+1}) - \max(t_r, t) \big) \exp(f_k^x(t_r, \bar{O}_k)) \Big),$$

and, likewise, $\prod_{s \geq t} dQ_s$:

$$\prod_{k=K_t}^{K} \Big(\big(\operatorname{expit}(f_k^L(1, T_k, \bar{O}_k))^{L(T_k)} \big(1 - \operatorname{expit}(f_k^L(1, T_k, \bar{O}_k))\big)^{1-L(T_k)}\big)^{\Delta N^\ell(T_k)} \\ \big(\exp(f_k^\ell(T_k, \bar{O}_k))\big)^{\Delta N^\ell(T_k)} \big(\exp(f_k^a(T_k, \bar{O}_k))\big)^{\Delta N^a(T_k)} \big(\exp(f_k^d(T_k, \bar{O}_k))\big)^{\Delta N^d(T_k)} \Big)$$

$$\prod_{x \in \{\ell, a, d\}} \exp\Big(- \sum_{r=1}^{\bar{K}_n} \sum_{k=K_t}^{K} \mathbb{1}\{T_k \in [t_{r-1}, t_r)\} \big(\min(t_r, T_{k+1}) - \max(T_k, t)\big) \exp(f_k^x(T_k, \bar{O}_k)) \\ - \sum_{r=1}^{\bar{K}_n} \sum_{k=K_t}^{K} \mathbb{1}\{t_r \in (T_k, T_{k+1}]\} \big(\min(t_{r+1}, T_{k+1}) - \max(t_r, t)\big) \exp(f_k^x(t_r, \bar{O}_k)) \Big).$$

Notice that there is only a difference between $\prod_{s>t} dQ_s$ and $\prod_{s \geq t} dQ_s$ at the actual jump times.

This shows the first statement of Lemma 2. Since $D^*(P) = D^*(f^x : x = L, A, \ell, a, c, d)$ is a sum of functions that are càdlàg and have finite variation norm and since $d\bar{G}_t^*/d\bar{G}_t$ is uniformly bounded away from zero which preserves the Donsker property of the ratio (van der Vaart and Wellner, 1996), then $\{D^*(f^x : x = L, A, c, a, \ell, d)\}$ is a Donsker class. □

LEMMA 9. *For a set of constants $M^x < \infty$, $x = L, A, a, \ell, d, c$, we have that $\{\mathcal{L}_x(f^x)\}$ is a Donsker class.*

PROOF. The log-likelihood loss $\mathcal{L}_A$ for f^A can be written,

$$\mathcal{L}_A(f^A)(O) = \sum_{k=0}^{K} \Big(- \Delta N^a(T_k) \big(A(T_k) \log\big(1 + \exp(-f_k^A(1, T_k, \bar{O}_k))\big) \\ + \big(1 - A(T_k)\big) \log\big(1 + \exp(f_k^A(0, T_k, \bar{O}_k))\big)\big) \Big),$$

and equivalently for log-likelihood loss $\mathcal{L}_L$ for f^L,

$$\mathcal{L}_L(f^L)(O) = \sum_{k=0}^{K} \Big(- \Delta N^\ell(T_k) \big(L(k) \log\big(1 + \exp(-f_k^L(1, T_k, \bar{O}_k))\big) \\ + \big(1 - L(k)\big) \log\big(1 + \exp(f_k^L(0, T_k, \bar{O}_k))\big)\big) \Big).$$

For $x = c, a, \ell, d$, we write the loss function as:

$$\mathcal{L}_x(f^x)(O) = \sum_{k=0}^{K} \Big(\Delta N^x(T_k) f_k^x(T_k, \bar{O}_k) - \sum_{r=1}^{\bar{K}_n} \mathbb{1}\{T_k \in [t_{r-1}, t_r)\} \big(\min(t_r, T_{k+1}) - T_k \big) \exp(f_k^x(T_k, \bar{O}_k)) - \sum_{r=1}^{\bar{K}_n} \mathbb{1}\{t_r \in (T_k, T_{k+1}]\} \big(\min(t_{r+1}, T_{k+1}) - t_r \big) \exp(f_k^x(t_r, \bar{O}_k)) \Big).$$

By Assumption 4, f_k^x is uniformly bounded. Also note that $\inf_{f_k^x} (1+\exp(f_k^x)) \geq 0$. Accordingly, f^x only ranges over values on which $x \mapsto \exp(x)$ and $x \mapsto \log x$ are Lipschitz. Since the set of càdlàg functions with finite variation norm is a Donsker class and since the Donsker property is preserved under Lipschitz transformations and also under products and sums (van der Vaart and Wellner, 1996) we conclude that $\mathcal{L}_x(f^x)$ for all $x = L, A, c, a, \ell, d$ is a Donsker class. □

B.4. HAL proof (Theorem 3). We proceed on the basis of the general HAL proof (van der Laan, 2017). What we need to show is that,

$$d_x(\hat{f}_n^x, f_0^x) = P_0 \mathcal{L}_x(\hat{f}_n^x) - P_0 \mathcal{L}_x(f_0^x) = o_P(n^{-1/2}), \tag{45}$$

for $x = Q, G$. By the general HAL proof, we have that $d_x(\hat{f}_n^x, f_0^x)$ is bounded by,

$$-(\mathbb{P}_n - P_0)\big(\mathcal{L}_x(\hat{f}_n^x) - \mathcal{L}_x(f_0^x)\big).$$

Since $\mathcal{L}_x(\hat{f}_n^x) - \mathcal{L}_x(f_0^x)$ falls in a Donsker class (Lemma 9) we have $d_x(\hat{f}_n^x, f_0^x) = O_P(n^{-1/2})$ which again implies that $P_0(\mathcal{L}_x(\hat{f}_n^x) - \mathcal{L}_x(f_0^x))^2 = O_P(n^{-1/2})$. The latter implication relies on (22) stated in Assumption 3, and holds for the squared error loss and the log-likelihood loss as long as these are uniformly bounded (c.f., Assumption 3). It then follows by the Donsker theorem that $-(\mathbb{P}_n - P_0)(\mathcal{L}_x(\hat{f}_n^x) - \mathcal{L}_x(f_0^x)) = o_p(n^{-1/2})$ which gives $d_x(\hat{f}_n^x, f_0^x) = o_P(n^{-1/2})$.

APPENDIX C: OVERVIEW

C.1. Overview of notation.

$L_0 \in \mathbb{R}^{d_0}$	baseline covariate vector	
$A(t) \in \mathcal{A}$	treatment decision at time t	
$L(t) \in \mathbb{R}^d$	covariate vector t	
$N^a(t)$	counting process recording changes in treatment	
$N^\ell(t)$	counting process recording changes in covariates	
$N^c(t)$	counting process recording changes in censoring status	
$N^d(t)$	counting process recording changes in death status	
Λ^x	the cumulative intensity characterizing the compensator of N^x	
π_t	the conditional density of $A(t)$; $\pi_{0,t}(a \mid \mathcal{F}_{t-}) = P(A(t) = a \mid \mathcal{F}_{t-})$	
μ_t	the conditional density of $L(t)$; $\mu_{0,t}(\ell \mid \mathcal{F}_{t-})$	
K_t	the subject-specific total number of unique events in $[0, t]$	p. 6
$K = K_\tau$	the subject-specific total number of unique events in $[0, \tau]$	
$T_1^a < \cdots < T^a_{N^a(\tau)}$	the subject-specific jump times of N^a	
$T_1^\ell < \cdots < T^\ell_{N^\ell(\tau)}$	the subject-specific jump times of N^ℓ	
$T_1 < \cdots < T_K$	subject-specific unique event times	
$O = \bar{O}(\tau)$	subject-specific observed data in $[0, \tau]$	p. 7
$P_0 \in \mathcal{M}$	the distribution of O	p. 8
$\mathcal{M}$	the statistical model containing P_0	p. 10
G, g	interventional part of $P \in \mathcal{M}$ and its density	p. 9
Q, q	non-interventional part of $P \in \mathcal{M}$ and its density	
G^*, g^*	intervention and its density	
$P_{Q,G^*} = P^{G^*}$	the post-interventional distribution defined by the g-computation formula	
$\Psi^{G^*} : \mathcal{M} \to \mathbb{R}$	target parameter for fixed intervention G^*	p. 11
$\boldsymbol{Z}$	$\boldsymbol{Z} = \left(Z_t^{G^*}, Z^{G^*}_{t,L(t)}, \Lambda^\ell(t), \Lambda^a(t), \Lambda^d(t) : t \in [0, \tau]\right)$	p. 14
$Z_t^{G^*}$	$Z_t^{G^*} = \mathbb{E}_{P^{G^*}}\left[Y \,\middle|\, L(t), N^\ell(t), N^a(t), N^d(t), \mathcal{F}_{t-}\right]$, $t \in [0, \tau]$	p. 13
$Z^{G^*}_{t,L(t)}$	$Z^{G^*}_{t,L(t)} = \mathbb{E}_{P^{G^*}}\left[Y \,\middle|\, N^\ell(t), N^a(t), N^d(t), \mathcal{F}_{t-}\right]$, $t \in [0, \tau]$	p. 13
$h_t^{G^*}$	clever weights, $t \in (0, \tau]$	p. 17
h_t^ℓ, h_t^a, h_t^d	clever covariates, $t \in (0, \tau]$	p. 17
$O_1, \ldots, O_n \sim P_0$	observed data	p. 7
$\mathbb{P}_n$	the empirical distribution of $\{O_i\}_{i=1}^n$	
$t_0 < t_1 < \cdots < t_{\bar{K}_n}$	the ordered sequence of unique times of changes $\cup_{i=1}^n \{T_{i,k}\}_{k=1}^{K_i}$	p. 7
$\bar{K}_n = \sum_{i=1}^n K_i$	the total number of observation times for the data $\{O_i\}_{i=1}^n$	
$\hat{G}_n$	estimator for G_0 on $[0, \tau]$	
$\hat{\boldsymbol{Z}}_n^k$	estimator for $\boldsymbol{Z}$	
$\hat{P}_n^*$	targeted estimator, characterized by $(\boldsymbol{Z}_n^*, \hat{G}_n)$, where $\boldsymbol{Z}_n^* = \boldsymbol{Z}_n^{k=k^*}$	
$\hat{\psi}_n^{G^*} = \Psi^{G^*}(\hat{P}_n^*)$	TMLE estimator for the target parameter	

C.2. Overview of targeting algorithm. We here provide a more detailed overview of the targeting procedure described in Section 7. As in Section 7, we here use the notation:

$$Z^{G^*}_{t_r} = \mathbb{E}_{P^{G^*}}[Y \mid L(t_r), N^\ell(t_r), N^a(t_r), N^d(t_r), \mathcal{F}_{t_{r-1}}], \tag{46}$$

$$Z^{G^*}_{t_r, L(t_r)} = \mathbb{E}_{P^{G^*}}[Y \mid N^\ell(t_r), N^a(t_r), N^d(t_r), \mathcal{F}_{t_{r-1}}], \tag{47}$$

for $r = 1, \ldots, \bar{K}_n$. Further, we now introduce:

$$Z^{G^*}_{t_r, N^\ell(t_r)} = \mathbb{E}_{P^{G^*}}[Y \mid N^a(t_r), N^d(t_r), \mathcal{F}_{t_{r-1}}], \tag{48}$$

$$Z^{G^*}_{t_r, N^a(t_r)} = \mathbb{E}_{P^{G^*}}[Y \mid N^d(t_r), \mathcal{F}_{t_{r-1}}] \tag{49}$$

$$Z^{G^*}_{t_r, N^d(t_r)} = \mathbb{E}_{P^{G^*}}[Y \mid \mathcal{F}_{t_{r-1}}], \tag{50}$$

where the subscript '$N^x(t_r)$', $x = \ell, a, d$, tells us what was last integrated out over the interval $(t_{r-1}, t_r]$. Given current estimators $\hat{\boldsymbol{Z}}^k_n$ for $\boldsymbol{Z}$, we construct estimators,

$$\hat{Z}^{G^*}_{t_r,k}, \hat{Z}^{G^*}_{t_r,L(t_r),k}, \hat{Z}^{G^*}_{t_r,N^\ell(t_r),k}, \hat{Z}^{G^*}_{t_r,N^a(t_r),k}, \hat{Z}^{G^*}_{t_r,N^d(t_r),k},$$

for (46)–(50). Moreover, given estimators for (46)–(50), we can provide estimators $\hat{h}^\ell_{t_r,k}, \hat{h}^a_{t_r,k}, \hat{h}^d_{t_r,k}$ for the clever covariates $h^\ell_{t_r}, h^a_{t_r}, h^d_{t_r}$ by:

$$\begin{aligned}
\hat{h}^\ell_{t_r,k} &= \sum_{\delta=0,1} (2\delta - 1)\hat{Z}^{G^*}_{t_r,L(t_r),k}(N^\ell(t_r) = N^\ell(t_{r-1}) + \delta), \\
\hat{h}^a_{t_r,k} &= \sum_{\delta=0,1} (2\delta - 1)\hat{Z}^{G^*}_{t_r,N^\ell(t_r),k}(N^a(t_r) = N^a(t_{r-1}) + \delta), \\
\hat{h}^d_{t_r,k} &= 1 - \hat{Z}^{G^*}_{t_r,N^a(t_r),k}(N^d(t_r) = 0).
\end{aligned}$$

Now we can carry out the individual targeting update steps as described in Section 7 (Sections 7.1 and 7.2). This gives:

$$\begin{aligned}
\hat{Z}^{G^*}_{t_r,L(t_r),k} &\mapsto \hat{Z}^{G^*}_{t_r,L(t_r),k+1} \\
\hat{\Lambda}^\ell_k &\mapsto \hat{\Lambda}^\ell_{k+1} \\
\hat{\Lambda}^a_k &\mapsto \hat{\Lambda}^a_{k+1} \\
\hat{\Lambda}^d_k &\mapsto \hat{\Lambda}^d_{k+1}
\end{aligned}$$

Starting from $\hat{Z}^{G^*}_{t_r,L(t_r),k+1}$, we use $\hat{\Lambda}^\ell_{k+1}$ to integrate out $N^\ell(t_r)$ to obtain the updated $\hat{Z}^{G^*}_{t_r,N^\ell(t_r),k+1}$. Similarly, we use $\hat{\Lambda}^a_{k+1}$ to obtain the updated

$\hat{Z}^{G^*}_{t_r,N^a(t_r),k+1}$ from $\hat{Z}^{G^*}_{t_r,N^\ell(t_r),k+1}$ and lastly $\hat{\Lambda}^d_{k+1}$ to obtain $\hat{Z}^{G^*}_{t_r,N^d(t_r),k+1}$ from $\hat{Z}^{G^*}_{t_r,N^a(t_r),k+1}$.

To proceed, recall that $\hat{Z}^{G^*}_{t_r,N^d(t_r),k+1}$ estimates $\mathbb{E}_{P^{G^*}}[Y \mid \mathcal{F}_{t_{r-1}}]$, i.e.,

$$\mathbb{E}_{P^{G^*}}[Y \mid N^c(t_{r-1}), A(t_{r-1}), L(t_{r-1}), N^\ell(t_{r-1}), N^a(t_{r-1}), N^d(t_{r-1}), \mathcal{F}_{t_{r-2}}].$$

The distributions of $N^c(t_{r-1}), A(t_{r-1})$ are specified by our intervention G^*, so that we can now further obtain an updated $\hat{Z}^{G^*}_{t_{r-1},k+1}$ from $\hat{Z}^{G^*}_{t_r,N^d(t_r),k+1}$ by integrating out $N^c(t_{r-1}), A(t_{r-1})$ according to dG^* over $(t_{r-2}, t_{r-1}]$. For example, if our intervention imposes no censoring and sets $A(t)$ to a^* throughout (see Equation (6) in Section 2.2), then we have:

$$\hat{Z}^{G^*}_{t_{r-1},k+1} = \hat{Z}^{G^*}_{t_r,N^d(t_r),k+1}(N^c(t_{r-1}) = 0, A(t_{r-1}) = a^*).$$

This means that we now have updated estimators:

$$\hat{Z}^{G^*}_{t_r,k+1}, \hat{Z}^{G^*}_{t_r,L(t_r),k+1}, \hat{Z}^{G^*}_{t_r,N^\ell(t_r),k+1}, \hat{Z}^{G^*}_{t_r,N^a(t_r),k+1}, \hat{Z}^{G^*}_{t_r,N^d(t_r),k+1},$$

for the entire sequence (46)–(50). We can now proceed with the next update $k+1$ to $k+2$ exactly as outlined above.

C.3. Overview diagrams.

Overview: Super learning.

For each of $x = L, A, \ell, a, c, d$:

Define loss: $(O, f^x) \mapsto \mathcal{L}_x(f^x)(O)$
Define a library of estimators: $O \mapsto \hat{f}^x_{n,m}(O)$ indexed by $m = 1, \ldots, M$

Include in the library a HAL estimator:
$\hat{f}^x_{m,n} = \underset{\beta:\|\beta\|_1<C^x}{\operatorname{argmin}} \mathbb{P}_n \mathcal{L}_x(f^x_h)$

Let $\hat{m}^x_n \in \mathbb{R}^M$ be the loss function based cross-validation selector:
$\hat{m}^x_n := \underset{m}{\operatorname{argmin}} \frac{1}{V} \sum_{v=1}^{V} \mathbb{P}^1_{n,v} \mathcal{L}_x(\hat{f}^x_{m,n}(\mathbb{P}^0_{n,v}))$
($\mathbb{P}^0_{n,v}, \mathbb{P}^1_{n,v}$ denotes the empirical distribution of the training and validation sample, respectively, in a V-fold cross validation scheme)

Define the (discrete) super learner: $\hat{f}^M_n := \hat{f}^x_{n,\hat{m}^x_n}$

Table 3

HAL estimation using `R` software.

`hal9001::fit_hal(X, Y, family=family)`

generates a HAL design matrix consisting of basis functions corresponding to covariates and interactions

makes a call to:
`glmnet::glmnet(x=X, y=Y, family=family)`

automatically selects a CV-optimal value of this regularization parameter:
`glmnet::cv.glmnet(x=X, y=Y, family=family)`

Logistic regression:

`X`	HAL design matrix
`Y`	outcome variable (factor with two levels)
`family`	`'binomial'`

Squared error loss:

`X`	HAL design matrix
`Y`	outcome variable (real-valued)
`family`	`'gaussian'`

Intensity estimation:

`X`	HAL design matrix
`Y`	Number of events observed for each combination of covariates
`R`	The amount of risk time for each combination of covariates `log(R)` is included as an offset in the regression formula
`family`	`'poisson'`

TABLE 4

H. C. Rytgaard, T. A. Gerds
Section of Biostatistics
University of Copenhagen
Øster Farimagsgade 5
1014 København K
Denmark
E-mail: hely@sund.ku.dk
tagteam@sund.ku.dk

M. J. van der Laan
Division of Biostatistics
and
Center for Targeted Machine
Learning and Causal Inference
101 Haviland Hall
Berkeley, California, 94720
USA
E-mail: laan@berkeley.edu

Manuscript II

Random forests for survival analysis

Helene C. Rytgaard and Thomas A. Gerds

Details:

The manuscript is published in Wiley StatsRef: Statistics Reference Online for John Wiley & Sons Ltd. (2018).

Wiley StatsRef:
Statistics Reference Online

Random Forests for Survival Analysis

By *Helene C. Rytgaard and Thomas A. Gerds*

Keywords: *survival analysis, random forests, machine learning, competing risks, survival trees, prediction, variable importance.*

Abstract: A random forest for survival analysis is a machine learning method that combines bootstrap aggregation with randomized survival trees for right-censored time-to-event data. The approach does not assume a model and thus provides a flexible alternative to Cox regression. It can be used to predict the risk of an event for individual subjects in the presence or absence of competing risks, and for the discovery of risk factors in low- and high-dimensional settings.

1 Introduction

A random forest for survival analysis is a random forest[1] (*see* **Random Forests**) with a time-to-event outcome. It is an ensemble learner that uses bootstrap perturbations of the data to grow survival trees, providing a method that can detect marginal and higher-level associations between a potentially high-dimensional covariate space and survival chances[2]. Inside the box, this machine learning technique relies on a data-driven algorithm that does not impose a model for the data-generating mechanism, requiring only that the user selects a number of hyperparameters (Figure 2). A random forest for survival analysis provides a flexible alternative to parametric and semiparametric survival models such as Cox proportional hazard regression (*see* **Proportional Hazards Model, Cox's**). A key feature is that it works in high-dimensional settings[3] where the number of predictors exceeds the sample size, for example in omics data[4] (*see* **Genetics and Genomics, Statistics in**). Applications of the method can be found in multiple fields, such as prediction of credit risk default in economics[5] and risk factor analysis in medical research[6].

Survival data are characterized by right censoring which means that not all subjects are followed until the event of interest occurs[7] (*see* **Censored Data Analysis**). The observation of a subject is called *right-censored*, if the event of interest did not occur within the subject-specific follow-up period. It is then assumed that the event occurs at a later time, later than (to the right of) the censoring time. This changes when there are competing risks[8] (*see* **Competing Risk Analysis**). A competing risk is an event after which the event of interest either cannot occur (e.g., because the subject died) or is for some other reason no longer of interest for the analysis.

University of Copenhagen, Copenhagen, Denmark

DOI: 10.1002/9781118445112.stat08010

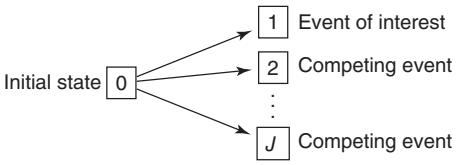


Figure 1. Multistate representation of a situation with competing risks. Each subject is in state 0 at time 0 and is then followed until the first event occurs. Right-censoring means that the subject is still in state 0 at the end of follow-up and it is thus not known which event occurs first and when.

We consider a situation with $J \geq 1$ mutually exclusive types of events (Figure 1) which includes the special case without competing risks ($J = 1$). The dataset, in the following referred to as the learning data, consists of $n \in \mathbb{N}$ samples,

$$\{(\boldsymbol{X}_1, T_1, \delta_1), \ \ldots \ , (\boldsymbol{X}_n, T_n, \delta_n)\}$$

For subject i, the variable $\delta_i \in \{0, 1, \ \ldots \ , J\}$ indicates if the observation is censored ($\delta_i = 0$) or if the event of interest has occurred ($\delta_i = 1$) or, when J > 1, if a competing risk has occurred ($\delta_i > 1$). Instead of the uncensored time-to-event T_i^* of subject i, observed is only $T_i = \min(T_i^*, C_i) \in \mathbb{R}_+$, where C_i is the censoring time. The feature space is spanned by a vector of covariates:

$$\boldsymbol{X}_i = (X_{1,i}, \ \ldots \ , X_{p,i}) \in \mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_p \subseteq \mathbb{R}^p$$

In survival analysis, we operate with parameters such as the event-free survival probability, $S_{\boldsymbol{x}}(t) = P(T^* > t \mid \boldsymbol{X} = \boldsymbol{x})$, and the cause-$j$ specific absolute risk function (cumulative incidence),

$$F_j(t \mid \boldsymbol{x}) = P(T^* \leq t, \delta = j \mid \boldsymbol{X} = \boldsymbol{x})$$

Both quantities can only be estimated for times that satisfy at least $0 \leq t \leq \tau = \max\{T_1, \ \ldots \ , T_n\}$, and both depend on the cause-specific hazard rates for all J competing causes of the event[8]. The cause-j specific hazard rate is the instantaneous probability of a cause-j event at time t given no event until just before time t:

$$\lambda(t \mid \boldsymbol{x}) = \lim_{\mathrm{d}t \to 0} \frac{P(t \leq T^* < t + \mathrm{d}t, \delta = j \mid \boldsymbol{X} = \boldsymbol{x})}{\mathrm{d}t \cdot S_{\boldsymbol{x}}(t)}$$

2 Algorithms and Implementations

A random forest for survival analysis essentially consists of an ensemble of randomized survival trees (Figure 2) (*see* **Regression Trees**). There are at least the following different implementations in the statistical software R[9], in the add-on packages `randomForestSRC`[10], `party`[11] and `ranger`[12]. Furthermore, Mogensen and Gerds[13] pointed out how an implementation of random forest for uncensored data can be combined with pseudo-values obtained with the Kaplan–Meier or the Aalen–Johansen statistic.

The number of trees is a hyperparameter chosen by the user. The randomization of the trees consists of the following parts. First, each tree is grown on a bootstrap sample of the learning data, drawing the bootstrap sample with or without replacement (*see* **Bootstrap with Examples**). Second, in the process of growing the survival trees, the algorithm selects the next binary split using only a random subset

Wiley StatsRef: Statistics Reference Online, © 2014–2018 John Wiley & Sons, Ltd.

DOI: 10.1002/9781118445112.stat08010

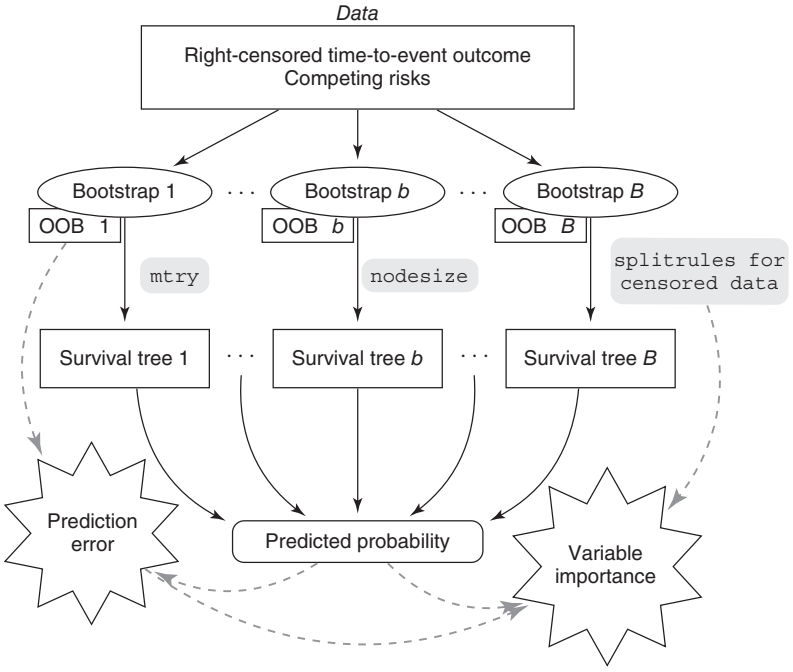


Figure 2. Illustration of the random forest algorithm for survival analysis. The parameter `nodesize` determines the constraint to control when to stop the tree growing process and `mtry` is the size of $\mathcal{P}'$.

$\mathcal{P}' \subseteq \{1, \ldots, p\}$ of the available covariates. The cardinality of the set $\mathcal{P}'$ is a hyperparameter which is usually called `mtry`. When there are continuous covariates, a third layer of randomization can be implemented which chooses a random subset of all the available binary splits that one can form based on the values of a continuous covariate (hyperparameter `msplit` of `randomForestSRC`). In random forests, the trees are typically grown as deep as possible, usually under the constraint that all nodes must contain a minimum number of observations.

The key to applying random forests in survival analysis consists of a revision of the tree-building process. For a comprehensive overview of survival trees, see Ref. 14. In general, a tree amounts to a partitioning algorithm that recursively implements binary splits of the covariate space $\mathcal{X}$ into disjoint subspaces. A survival tree, in particular, uses a specific splitting rule that explicitly takes into account the right-censored nature of the data. In each split, the partitioning is performed only along a single axis (Figure 3), and optimal splits are chosen such that, as the tree grows, the regions of the covariate space defined by the subsequent splits will become more and more similar in terms of survival. The randomization of the process of growing the trees (`mtry`, `msplit`) of a forest increases the variability of the trees. The information obtained from the terminal nodes of the trees of the forest will thus differ accordingly, and is aggregated to form the forest ensemble estimates.

There are different variants of the random forest algorithm for survival analysis and also other tree-based methods applicable to right-censored data, varying with respect to how trees are grown and how aggregation of tree-specific information is performed. The two major algorithms are the random survival forests[2] and the conditional inference forests[11,15,16]. Other alternatives are[17,18].

DOI: 10.1002/9781118445112.stat08010

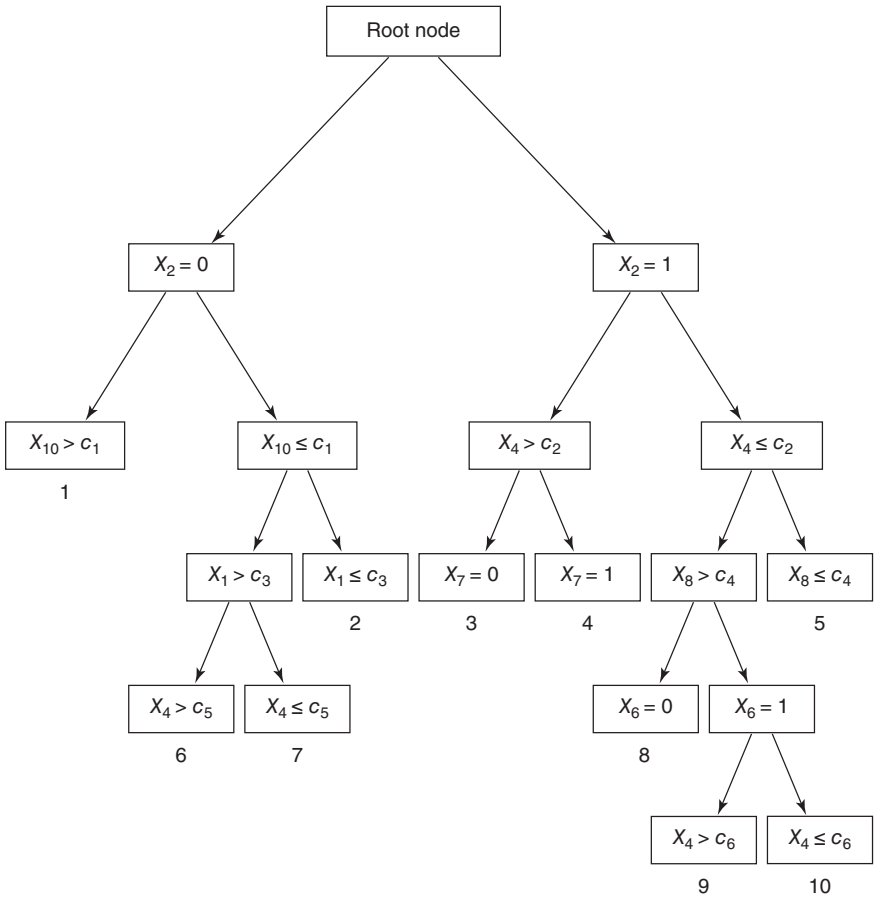


Figure 3. Example of a simple tree with 10 terminal nodes grown on a dataset with covariates $X_1, \ldots, X_{10}$. For instance, terminal node 1 corresponds to $\{X_2 = 0\} \cap \{X_{10} > c_1\}$.

3 Growing Survival Trees

Splitting rules for growing survival trees are usually based on two-sample tests for right-censored data[19,20]. In this way, subjects are discriminated according to their expected survival outcome such that survival within the daughter nodes of the next split is more similar than between daughter nodes. The particular choice of the test statistic defines the measure of similarity. Other examples of splitting rules for survival trees are those based on measures of within-node homogeneity such as the distance between Kaplan–Meier estimates of the survival curves[21,22].

Any nonterminal node of a survival tree corresponds to a subspace $\mathcal{X}' \subseteq \mathcal{X}$. A proposed split of the node leads to a partition, $\mathcal{X}'_l \cup \mathcal{X}'_r = \mathcal{X}'$, into two disjoint daughter nodes $\{i\colon\ X_i \in \mathcal{X}'_l\}$ and $\{i\colon\ X_i \in \mathcal{X}'_r\}$. For each such split, we can perform a statistical two-sample test for survival equivalence or any other criterion. The standard choice is the log-rank test for the null hypothesis of equal survival probabilities

Wiley StatsRef: Statistics Reference Online, © 2014–2018 John Wiley & Sons, Ltd.

DOI: 10.1002/9781118445112.stat08010

(*see* **Logrank Test**). We use counting process notation[7] $N_i(t) = \mathbb{1}\{T_i \leq t,\ \delta_i = 1\}$ and $Y_i(t) = \mathbb{1}\{T_i > t\}$ and define for the left daughter node,

$$\overline{N}_l(t) = \sum_{i=1}^{n} \mathbb{1}\{\boldsymbol{X}_i \in \mathcal{X}_l'\} \cdot N_i(t), \quad \overline{Y}_l(t) = \sum_{i=1}^{n} \mathbb{1}\{\boldsymbol{X}_i \in \mathcal{X}_l'\} \cdot Y_i(t)$$

and correspondingly $\overline{N}_r(t)$, $\overline{Y}_r(t)$ for the right daughter node. The Fleming–Harrington's family of test statistics[23] can be written as,

$$Z^{\mathrm{LR}}(t) = \frac{1}{\hat{\sigma}^{\mathrm{LR}}} \int_0^t W(s)\, \frac{\overline{Y}_l(s)\overline{Y}_r(s)}{\overline{Y}_l(s) + \overline{Y}_r(s)} \left(\frac{d\overline{N}_l(s)}{\overline{Y}_l(s)} - \frac{d\overline{N}_r(s)}{\overline{Y}_r(s)} \right) \tag{1}$$

where $W(t_i) = 1$ corresponds to the standard log-rank test. Varying the weights $W(s)$ in Equation (1) can be used to put more or less emphasis on early and late time-points. The partitioning into daughter nodes $\mathcal{X}_l'$, $\mathcal{X}_r' = \mathcal{X}'/\mathcal{X}_l'$ is chosen such that $|Z^{\mathrm{LR}}(t)|$ is maximized.

The daughter nodes defined by the split $\mathcal{X}_l'$, $\mathcal{X}_r' = \mathcal{X}'/\mathcal{X}_l'$, are limited to differ only along a single axis of $\mathcal{X}$, that is, only according to the values of one of the p covariates. For example, a split performed along the qth axis, when X_q is continuous or ordinal, is of the form,

$$\mathcal{X}_l' = \mathcal{X}' \cap \{X_q < c\}, \quad \mathcal{X}_r' = \mathcal{X}' \cap \{X_q \geq c\}$$

for some value $c \in \mathbb{R}$ that is based on the observed values of $X_q \in \mathcal{X}'$. The choice of splitting criterion can be used to make splits with certain properties more likely. For instance, the log-rank test statistic is optimal when the hazard ratio between the daughter nodes is proportional whereas variations over $W(t_i)$ can be used to target splits with either short-term or long-term effects. Notably, the split point selection is also related to variable selection[18,24].

3.1 Competing Risks

For forests applied to competing risks analysis, different choices of splitting rules have different interpretations in terms of distinguishing variables with direct effect on the hazard of a specific event of interest and variables also having an indirect effect through the competing events. See also Ref. 8. Let $\lambda_{j,l}(\cdot)$, $\lambda_{j,r}(\cdot)$, and $\lambda_{j,0}(\cdot)$ denote the hazard function in the left daughter, the right daughter, and the parent node, respectively, and likewise let $F_{j,l}(\cdot)$, $F_{j,r}(\cdot)$, and $F_{j,0}(\cdot)$ denote the cumulative distribution functions. On the one hand, the log-rank test is based on the test of equal cause-specific hazards,

$$H_0\colon\ \lambda_{j,l}(t) = \lambda_{j,r}(t) = \lambda_{j,0}(t), \quad t \leq \tau$$

and thus takes only cause-specific risk factors into account. On the other hand, effects from variables on a cause-specific event of interest also indirectly through the hazard function of competing events manifest themselves on the cumulative incidence,

$$F_j(t \mid \boldsymbol{x}) = \int_0^t S(u- \mid \boldsymbol{x})\ \lambda_j(u \mid \boldsymbol{x})\ \mathrm{d}u = \int_0^t \exp\left(- \int_0^u \sum_{k=1}^{J} \lambda_k(s \mid \boldsymbol{x})\ \mathrm{d}s \right)\ \lambda_j(u \mid x)\ \mathrm{d}u$$

Variables with such effects can be distinguished by use of Gray's test[25] instead, which is the test of the null hypothesis

$$H_0\colon\ F_{j,l}(t) = F_{j,r}(t) = F_{j,0}(t), \quad t \leq \tau$$

DOI: 10.1002/9781118445112.stat08010

4 Ensemble Prediction

We first note the difference between making predictions based on an isolated (stand-alone) survival tree and making predictions based on a forest. When an isolated survival tree is used for the prediction of a subject with covariate value $\boldsymbol{x}$, the estimator operates on the part of the learning data which shares the terminal node with $\boldsymbol{x}$, and the prediction for any subject sharing this terminal node will be the same. The intuition is that the terminal nodes of a tree will contain subjects that are similar with respect to their expected survival outcome. The prediction of a random forest is based on all the learning data (including possible bootstrap replicates) of all terminal nodes of the B trees in which $\boldsymbol{x}$ falls.

There are different approaches to aggregate the data of the terminal nodes shared with $\boldsymbol{x}$. For example, Ref. 16 collects the learning samples of the terminal nodes of $\boldsymbol{x}$ into a new dataset (with possible bootstrap repetition), and then the ensemble Kaplan–Meier estimate (*see* **Kaplan–Meier Estimator**) is computed from this dataset. This can be written as,

$$\hat{S}_{\boldsymbol{x}}(t) = \prod_{s \le t} \left(1 - \frac{\sum_{i=1}^{n} \sum_{b=1}^{B} n_{i,b} \cdot L_{i,b}(\boldsymbol{x}) \cdot N_i(\mathrm{d}s)}{\sum_{i=1}^{n} \sum_{b=1}^{B} n_{i,b} \cdot L_{i,b}(\boldsymbol{x}) \cdot Y_i(s)} \right)$$

where the function,

$$L_{i,b}(\boldsymbol{x}) = \sum_{k=1}^{K_b} \mathbb{1}\{\boldsymbol{X}_i \in C_k^b\} \mathbb{1}\{\boldsymbol{x} \in C_k^b\}$$

indicates which subjects of the learning data share the terminal node with $\boldsymbol{x}$ in the bth tree, and $n_{i,b} \geq 0$ is the number of times sample i is used in the tree. In Refs 2 and 8, on the other hand, Nelson–Aalen (*see* **Nelson–Aalen Estimator**), Kaplan–Meier, and Aalen–Johansen (*see* **Aalen–Johansen Estimator**) estimates are computed for each survival tree separately based on the learning data that share the terminal node with $\boldsymbol{x}$, and then these estimates are averaged over trees.

5 Predictive Accuracy

Traditionally, the machine learning community uses predictive accuracy to validate their models[26]. In right-censored survival data, standard measures of prediction accuracy are the time-dependent area under the receiver operating characteristic[27–29] and the time-dependent Brier score and time-dependent logarithmic score[30,31]. When the aim is to assess the prediction performance of a random forest, one can use the out-of-bag data[8]. When the aim is to compare the performance of the random forest model with other survival prediction models, one needs an outer cross-validation loop to estimate the prediction accuracy[32].

6 Variable Importance

Random forests can be used to rank covariates in terms of their association with the survival outcome, their variable importance (VIMP)[2,33], and for the investigation of pairwise variable interactions. The most commonly used measures of VIMP are based on predictive accuracy. The idea is to compare a measure of prediction performance when the forest is grown first with and then without each variable separately. To avoid fitting a new forest p times, in implementations this is often accomplished by growing the forest with

Wiley StatsRef: Statistics Reference Online, © 2014–2018 John Wiley & Sons, Ltd.

DOI: 10.1002/9781118445112.stat08010

a "noised-up" version of the variable. When the prediction performance is decreasing more for a noised-up version of one variable than for a noised-up version of another, the former is concluded to be more important than the latter.

A different approach to measuring VIMP is the so-called *minimal depth*[34]. For any variable X_j, this is defined as the shortest distance between the root node of the tree and a node that splits on X_j, a so-called X_j-subtree. In Figure 3, for example, the minimal depth of variable X_2 is 0 and the minimal depth of variable X_4 is 2. The randomness resulting from sampling $\mathcal{P}'$ in each split is averaged out over the trees, and hence the earlier a variable is used for a split, the more important it is for discriminating survival outcome. Thus, the minimal depth is directly linked to the choice of splitting criterion, and, accordingly, variations of the splitting rule can be used to target different interpretations of the minimal depth in order to, for instance, distinguish variables with long- or short-term effects on outcome. Note also that X_j-subtrees for different variables X_j can be used to extract information on interactions, by considering second-order subtrees. In particular, variables interacting with X_j can be identified from other variables' subtrees within X_j-subtrees[34].

7 Asymptotic Properties

Asymptotic properties of random forest estimators are a concern of more recent research. Consistency has been studied for different variants of random forests and is proved for the survival forest ensemble under the assumption of a discrete covariate space $\mathcal{X}$[35]; for quantile regression forests[36] and classification forests[37], consistency is proved under the more general assumption of a continuous $\mathcal{X}$. Asymptotic normality has been looked into more recently in Ref. 38, for a particular variant of regression forests.

References

[1] Breiman, L. (2001) Random forests. *Mach. Learn.* **45** (1), 5–32.
[2] Ishwaran, H., Kogalur, U.B., Blackstone, E.H., and Lauer, M.S. (2008) Random survival forests. *Ann. Appl. Stat.* **2** (3), 841–860.
[3] Ishwaran, H., Kogalur, U.B., Chen, X., and Minn, A.J. (2011) Random survival forests for high-dimensional data. *Stat. Anal Data Min.* **4** (1), 115–132.
[4] Chen, X. and Ishwaran, H. (2012) Random forests for genomic data analysis. *Genomics* **99** (6), 323–329.
[5] Fantazzini, D. and Figini, S. (2009) Random survival forests models for SME credit risk measurement. *Methodol. Comput. Appl. Probab.* **11** (1), 29–45.
[6] Hsich, E., Gorodeski, E.Z., Blackstone, E.H. *et al.* (2011) Identifying important risk factors for survival in patient with systolic heart failure using random survival forests. *Cir. Cardiovasc. Qual. Outcomes* **4** (1), 39–45.
[7] Andersen, P.K., Borgan, O., Gill, R.D., and Keiding, N. (1993) *Statistical Models Based on Counting Processes*, Springer Series in Statistics, Springer, New York.
[8] Ishwaran, H., Gerds, T.A., Kogalur, U.B. *et al.* (2014) Random survival forests for competing risks. *Biostatistics* **15** (4), 757–773.
[9] R Core Team (2017) R: A Language and Environment for Statistical Computing, *R Foundation for Statistical Computing*, Vienna, Austria.
[10] Ishwaran, H. and Kogalur, U.B. (2007) Random survival forests for R. *R News* **7** (2), 25–31.
[11] Hothorn, T., Bühlmann, P., Dudoit, S. *et al.* (2006) Survival ensembles. *Biostatistics* **7** (3), 355–373.
[12] Wright, M.N. and Ziegler, A. (2017) ranger: a fast implementation of random forests for high dimensional data in C++ and R. *J. Stat. Softw.* **77** (1), 1–17.
[13] Mogensen, U.B. and Gerds, T.A. (2013) A random forest approach for competing risks based on pseudo-values. *Stat. Med.* **32** (18), 3102–3114.
[14] Bou-Hamad, I., Larocque, D., and Ben-Ameur, H. (2011) A review of survival trees. *Stat. Surv.* **5**, 44–71.
[15] Hothorn, T., Hornik, K., and Zeileis, A. (2006) Unbiased recursive partitioning: a conditional inference framework. *J. Comput. Graph. Stat.* **15** (3), 651–674.

DOI: 10.1002/9781118445112.stat08010

[16] Hothorn, T., Lausen, B., Benner, A., and Radespiel-Tröger, M. (2004) Bagging survival trees. *Stat. Med.* **23** (1), 77–91.
[17] Zhu, R. and Kosorok, M.R. (2012) Recursively imputed survival trees. *J. Am. Stat. Assoc.* **107** (497), 331–340.
[18] Wright, M.N., Dankowski, T., and Ziegler, A. (2017) Unbiased split variable selection for random survival forests using maximally selected rank statistics. *Stat. Med.* **36** (8), 1272–1284.
[19] Segal, M.R. (1988) Regression trees for censored data. *Biometrics* **44** (1), 35–47.
[20] LeBlanc, M. and Crowley, J. (1993) Survival trees by goodness of split. *J. Am. Stat. Assoc.* **88** (422), 457–467.
[21] Gordon, L. and Olshen, R.A. (1985) Tree-structured survival analysis. *Cancer Treat. Rep.* **69** (10), 1065–1069.
[22] Moradian, H., Larocque, D., and Bellavance, F. (2016) L_1 splitting rules in survival forests. *Lifetime Data Anal.* 1–21. doi: 10.1007/s10985-016-9372-1.
[23] Harrington, D.P. and Fleming, T.R. (1982) A class of rank test procedures for censored survival data. *Biometrika* **69** (3), 553–566.
[24] Nasejje, J.B., Mwambi, H., Dheda, K., and Lesosky, M. (2017) A comparison of the conditional inference survival forest model to random survival forests based on a simulation study as well as on two applications with time-to-event data. *BMC Med. Res. Methodol.* **17** (1), 115.
[25] Gray, R.J. (1988) A class of K-sample tests for comparing the cumulative incidence of a competing risk. *Ann. Stat.* **16** (3), 1141–1154.
[26] Breiman, L. (2001) Statistical modeling: the two cultures. *Stat. Sci.* **16** (3), 199–231.
[27] Heagerty, P.J., Lumley, T., and Pepe, M.S. (2000) Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics* **56**, 337–344.
[28] Chambless, L.E. and Diao, G. (2006) Estimation of time-dependent area under the ROC curve for long-term risk prediction. *Stat. Med.* **25** (20), 3474–3486.
[29] Blanche, P., Dartigues, J.F., and Jacqmin-Gadda, H. (2013) Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. *Stat. Med.* **32** (30), 5381–5397.
[30] Gerds, T.A. and Schumacher, M. (2006) Consistent estimation of the expected Brier score in general survival models with right-censored event times. *Biom. J.* **48** (6), 1029–1040.
[31] van Houwelingen, H.C. and Putter, H. (2012) *Dynamic Prediction in Clinical Survival Analysis*, CRC Press, Boca Raton, FL.
[32] Mogensen, U.B., Ishwaran, H., and Gerds, T.A. (2012) Evaluating random forests for survival analysis using prediction error curves. *J. Stat. Softw.* **50** (11). doi: 10.18637/jss.v050.i11.
[33] Ishwaran, H. (2007) Variable importance in binary regression trees and forests. *Electron. J. Stat.* **1**, 519–537.
[34] Ishwaran, H., Kogalur, U.B., Gorodeski, E.Z. *et al.* (2010) High-dimensional variable selection for survival data. *J. Am. Stat. Assoc.* **105** (489), 205–217.
[35] Ishwaran, H. and Kogalur, U.B. (2010) Consistency of random survival forests. *Stat. Probab. Lett.* **80** (13), 1056–1064.
[36] Meinshausen, N. (2006) Quantile regression forests. *J. Mach. Learn. Res.* **7**, 983–999.
[37] Biau, G., Devroye, L., and Lugosi, G. (2008) Consistency of random forests and other averaging classifiers. *J. Mach. Learn. Res.* **9**, 2015–2033.
[38] Wager, S. and Athey, S. (2017) Estimation and inference of heterogeneous treatment effects using random forests. *J. Am. Stat. Assoc.* doi: 10.1080/01621459.2017.1319839.

Wiley StatsRef: Statistics Reference Online, © 2014–2018 John Wiley & Sons, Ltd.

DOI: 10.1002/9781118445112.stat08010

Manuscript III

Application of generalized random forests for survival analysis

Helene C. Rytgaard

Details:

The manuscript is published in Proceedings of the 21st European Young Statisticians Meeting (2019).

Application of generalized random forests for survival analysis

Helene Charlotte Rytgaard[*1]

[1]*Section of Biostatistics, University of Copenhagen*

Abstract: We are interested in estimating treatment effects on the absolute risk of an event in a survival analysis setting. The particular approach taken in this paper is based on the generalized random forest (GRF) [2] methodology that we adapt to right-censored data. We formulate the estimation problem in terms of counterfactual outcomes where both treatment and censoring act as a coarsening on the underlying survival time, and define our target parameter as the solution to an inverse probability weighted estimating equation. To grow the forest, we use a partitioning scheme (splitting criteria) based on the influence function for our target parameter. The result is a nonparametric estimator for the treatment effect on survival.

Keywords: Survival analysis, random forests, causal inference, treatment effects, censored data.

1 Introduction

Estimation of average treatment effects by means of machine learning methods has applications in fields such as biostatistics and econometrics and is a popular alternative to parametric and semiparametric methods. This article is concerned with the adaption of the generalized random forest (GRF) [2] framework, a recent extension of the original random forest [4] based on subsampling and honesty, to estimation of treatment effects based on right-censored data. The GRF methodology is formulated in terms of estimating equations of the form,

$$\mathbb{E}\big[\psi_{\theta(x),\nu(x)}(O) \,|\, X = x\big] = 0, \tag{1}$$

*Corresponding author: hely@sund.ku.dk

for estimation of a parameter $\theta(x)$ based on data $(X_i, O_i) \in \mathcal{X} \times \mathcal{O}$, $\mathcal{X} \subseteq \mathbb{R}^p$ where $\psi_{\theta(x),\nu(x)}(\cdot)$ is a scoring function and $\nu(x)$ is an optional nuisance parameter. GRFs have been applied to estimation of heterogeneous treatment effects [12, 2] but not for censored time-to-event outcomes.

A random forest consists of trees where each tree recursively splits subsamples of data using a specific partitioning scheme. Central to the GRF algorithm is that the partitioning scheme targets specifically the estimation of $\theta(x)$. The idea is to label subjects with the influence function of a local estimator for the target parameter. Then a split is implemented such as to maximize heterogeneity in the labeled subjects. By averaging over the neighborhoods defined by each tree, the forest outputs a weighting function that can be used to find solutions to the estimating equation (1).

To adapt the GRF methodology to the survival analysis setting we consider a specific estimation equation that involves a Kaplan-Meier integral for which we derive the influence function. The forest weights define a kernel function based on which we construct an estimator that solves the estimating equation of interest. That way, we obtain a nonparametric estimator, allowing for covariate-dependent censoring, that is targeted directly towards the treatment effect on survival.

2 Setting and notation

Suppose we make $n \in \mathbb{N}$ independent and identically distributed observations of,

$$X \in \mathbb{R}^p, \quad O = (A, \tilde{T}, \Delta) \in \{0,1\} \times \mathbb{R}_+ \times \{0,1\},$$

where $\tilde{T}$ is a continuous time-to-event outcome observed under right-censoring, $\Delta \in \{0,1\}$ is an indicator of event, $X \in \mathcal{X} \subseteq \mathbb{R}^p$ is a vector of baseline covariate values, and $A \in \{0,1\}$ is a binary treatment assigned at baseline. We represent the observed data $(X, A, \tilde{T}, \Delta)$ as a many-to-one mapping on the full data structure (X, T^0, T^1) induced by a coarsening by (A, C) [10, 9]. Here, C is the censoring time and T^a is the uncensored counterfactual event time that would result if treatment had been set to $A = a$. The observed survival outcome variables are then given as $\tilde{T} = T^A \wedge C$ and $\Delta = \mathbb{1}\{T^A \leq C\}$.

Our interest is in the counterfactual distributions $F^a(t \,|\, x) = P(T^a \leq t \,|\, X = x)$ for $a = 0, 1$. We further use the notation $F(t, a \,|\, x) = P(T \leq t, A = a \,|\, x)$, $G(t, a \,|\, x) = P(C > t, A = a \,|\, x)$, $H^\delta(t, a \,|\, x) = P(\tilde{T} \leq t, \Delta = \delta, A = a \,|\, X = x)$ for $\delta = 0, 1$ and $H(t, a \,|\, x) = P(\tilde{T} \geq t, A = a \,|\, X = x)$. We assume coarsening at random (CAR) [10, 6] and positivity, $P(C > t_0, A = a \,|\, X) >$

$\eta > 0$, a.s. for $a = 0, 1$ and a fixed timepoint $t_0 > 0$. We note that, under these assumptions, the conditional density of an observation O (with respect to an appropriate dominating measure) can be expressed as,

$$\begin{aligned} &P(\tilde{T} \in dt, \Delta = 1, A = a \,|\, X = x) \\ &\quad = P(T^a \in dt \,|\, X = x) P(C > t \,|\, A = a, X = x) P(A = a \,|\, X = x), \end{aligned}$$

and we have the following relations,

$$F^a(dt \,|\, x) = \frac{H^1(dt, a \,|\, x)}{G(t, a \,|\, x)}, \qquad G(t \,|\, a, x) = \prod_{s \in (0,t]} \left(1 - \frac{H^0(ds, a \,|\, x)}{H(s, a \,|\, x)}\right), \quad (2)$$

where $\prod$ denotes the product integral [1].

3 Kernel estimation

We are concerned with estimation of $\theta(x) = \theta_1(x) - \theta_0(x)$, where,

$$\theta_a(x) = \int_0^\infty \mathbb{1}\{t > t_0\} \, dF^a(t \,|\, x), \quad a = 0, 1. \quad (3)$$

The dependence on the timepoint of interest, t_0, is implicit in the notation for $\theta_a(x)$. We note that $\theta_a(x)$ is defined as a functional of the distribution F^a of the unobservable T^a. By CAR and positivity, we can rewrite (3) using (2) as,

$$\theta_a(x) = \int_0^\infty \mathbb{1}\{t > t_0\} \, \frac{H^1(dt, a \,|\, x)}{G(t, a \,|\, x)}, \quad a = 0, 1. \quad (4)$$

This corresponds to an inverse probability weighted estimating equation of the form,

$$\mathbb{E}\left[\sum_{a \in \{0,1\}} (2a - 1) \left(\int_0^\infty \mathbb{1}\{t > t_0\} \, \frac{\mathbb{1}\{\tilde{T} \le t, \Delta = 1, A = a\}}{G(t, a \,|\, x)} \right) - \theta(x) \,\middle|\, X = x \right] = 0,$$

with nuisance parameters (H^1, G). We consider the following estimators, for a kernel weighting function $K(x, x') \ge 0$,

$$\hat{H}^\delta_K(t, a \,|\, x) = \sum_{i=1}^n K(x, x_i) \, \mathbb{1}\{\tilde{T}_i \le t, \Delta_i = \delta, A_i = a\}, \quad \text{for } \delta = 0, 1,$$

$$\hat{H}_K(t, a \,|\, x) = \sum_{i=1}^n K(x, x_i) \, \mathbb{1}\{\tilde{T}_i > t, A_i = a\}.$$

The kernel function $K(x, x')$ is used to place more weight on observations in the covariate space $\mathcal{X}$ that are close to x. We define the estimators,

$$\hat{\theta}_{K,a}(x) = \int_0^\infty \mathbb{1}\{t > t_0\} \frac{\hat{H}^1_K(dt, a \,|\, x)}{\hat{G}_K(t, a \,|\, x)}, \;\; \hat{G}_K(t \,|\, a, x) = \prod_{s \leq t} \left(1 - \frac{\hat{H}^0_K(ds, a \,|\, x)}{\hat{H}_K(s, a \,|\, x)}\right).$$

In the GRF framework we replace the kernel weighting function $K(x, x')$ by forest-based weights as we will show in the following.

4 GRF for survival analysis

A random forest consists of a set of $B \in \mathbb{N}$ trees that each provides a partitioning of the covariate space. The following outlines the tree building process for the b^{th} tree in the GRF framework.

1. *Subsampling.* An index set $\mathcal{J}_b$ of size $s_n < n$ is sampled randomly from $\{1, \ldots, n\}$ without replacement.

2. *Honesty.* The index set $\mathcal{J}_b$ is divided randomly into $\mathcal{J}^1_b \cup \mathcal{J}^2_b$ of sizes $\lfloor s_n/2 \rfloor$ and $\lceil s_n/2 \rceil$.

3. *Splitting.* The tree is grown by recursively implementing binary axis-aligned splits of the covariates space based on the samples $\{i : i \in \mathcal{J}^1_b\}$. We describe the particular splitting rule used for estimation of our target parameter below.

The randomness induced by subsampling together with randomly selecting a smaller set of variables as candidates for a split ensures diversity of the different trees of the forest.

Splitting is central to the tree building scheme. In the GRF framework, splitting rules are targeted specifically towards the target parameter $\theta(x)$. Particularly, the idea is to implement splits of a mother node $M \subseteq \mathcal{X}$ into daughters $D_1 \cup D_2 = M$ so as to maximize,

$$\mathcal{L}(D_1, D_2) \equiv \sum_{j=1}^{2} P(X \in D_j \,|\, X \in M)\, \mathbb{E}\big[\big(\hat{\theta}_{D_j} - \theta(X)\big)^2 \,|\, X \in D_j\big]. \qquad (5)$$

Here, $\hat{\theta}_{D_j}$ is the estimate of the target parameter in the j^{th} daughter node, corresponding to the kernel weight $K_{D_j}(x, x') = \mathbb{1}\{x' \in D_j\}$. As proposed in [2], we will approximate the splitting criterion in (5) in the following way. Let

$\hat{\theta}_M$ be the estimator for the target parameter, corresponding to the kernel weight $K_M(x, x') = \mathbb{1}\{x' \in M\}$. Define,

$$\Psi_a(H^1, G) = \int_0^\infty \mathbb{1}\{t > t_0\} \frac{H^1(dt, a)}{G(t, a)}.$$

The influence function of the estimator $\hat{\theta}_M = \Psi(\hat{H}^1_M, \hat{G}_M)$ may be derived as the Gâteaux derivative of the functional $\Psi(H^1, G) = \Psi_1(H^1, G) - \Psi_0(H^1, G)$ in direction of δ_{O_i} [5, 11]. This influence function is given as $\text{IF}(H^1, G) = \text{IF}_1(H^1, G) - \text{IF}_0(H^1, G)$, where, for $a = 0, 1$,

$$\begin{aligned}
\text{IF}_a(H^1, G)(O_i) = \Bigg(& \frac{\mathbb{1}\{\tilde{T}_i > t_0, \Delta_i = 1, A_i = a\}}{G(\tilde{T}_i, a \,|\, X_i)} + \mathbb{1}\{A_i = a\} \times \\
& \Bigg(\frac{1 - \Delta_i}{H(\tilde{T}_i, a \,|\, X_i)} \int_{\tilde{T}_i}^\infty \mathbb{1}\{t > t_0\} \frac{H^1(dt, a \,|\, X_i)}{G(t, a \,|\, X_i)} \\
& - \int_0^\infty \mathbb{1}\{t > t_0\} \frac{H^1(dt, a \,|\, X_i)}{G(t, a \,|\, X_i)} \Bigg(\int_0^{t \wedge \tilde{T}_i} \frac{H^0(ds, a \,|\, X_i)}{\big(H(s, a \,|\, X_i)\big)^2} \Bigg) \Bigg) \Bigg) \\
& - \Psi_a(H^1, G).
\end{aligned}$$

Now, we can approximate the splitting criterion defined in (5) by,

$$\tilde{\mathcal{L}}(D_1, D_2) \equiv \sum_{j=1,2} \frac{1}{\sum_{i=1}^n \mathbb{1}\{X_i \in D_j\}} \Bigg(\sum_{\{i : X_i \in D_j\}} \text{IF}(\hat{H}^1_M, \hat{G}_M)(O_i) \Bigg)^2. \quad (6)$$

The estimated influence function $\text{IF}(\hat{H}^1_M, \hat{G}_M)(O_i)$ represents the rate of change in $\hat{\theta}_M$ in direction of $O_i \in D_j$, and the criterion defined by (6) seeks to separate samples in a way such that the estimates in the daughter nodes, $\hat{\theta}_{D_j}$, $j = 1, 2$, differ as much as possible from the estimate in the mother node, $\hat{\theta}_M$.

Node M specific estimation and the approximation by (6) is defined locally for $X \in M$. When the splitting process is repeated iteratively, we move through smaller and smaller neighborhoods defined by each current mother node. We let $L_b(x) \subseteq \mathcal{X}$ denote the terminal node of the b^{th} tree that contains $x \in \mathcal{X}$. Forest weights are obtained by averaging over the neighborhoods $L_b(x)$, $b = 1, \ldots, B$,

$$\alpha_i(x) = \frac{1}{B} \sum_{b=1}^B \alpha_{b,i}(x), \text{ where, } \alpha_{b,i}(x) = \frac{\mathbb{1}\{X_i \in L_b(x), i \in \mathcal{J}^2_b\}}{\sum_{k=1}^n \mathbb{1}\{X_k \in L_b(x), k \in \mathcal{J}^2_b\}}. \quad (7)$$

Our forest estimator for $\theta(x)$ is defined as,

$$\hat{\theta}_\alpha(x) = \sum_{a\in\{0,1\}} (2a-1) \int_0^\infty \mathbb{1}\{t > t_0\} \frac{d\hat{H}^1_\alpha(t,a)}{\hat{G}_\alpha(t,a)},$$

using the kernel function defined by the forest $K(x, x_i) = \alpha_i(x)$. The terminal nodes shrinking around x for $n \to \infty$ implies that $K(x, x_i) = \alpha_i(x) \to \delta_x$ for $n \to \infty$.

5 Discussion

In this paper we have demonstrated how the GRF methodology can be adapted to right-censored data. We have proposed a forest-based kernel weighted estimator of the treatment effect on the absolute risk and derived the influence curve to be used for the recursive splitting scheme. That way, estimation is targeted directly towards the treatment effect and optimized for the timepoint of interest.

We note that this stands in contrast to the existing random forest algorithms for survival analysis, see for instance [7, 8]. For these, splitting rules are typically based on two-sample tests for right-censored data focusing on survival estimation over the whole time range. Our approach will be useful in the application of average treatment effects as a variable importance measure. Another extension of interest deals with competing risks analysis. Here our methods could be used to rank a list of treatments in terms of their effect on hospitalization with depression or bipolar disorder in presence of the competing risk of death.

References

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

[2] Athey, S., Tibshirani, J. and Wager, S. (2019). Generalized random forests. *The Annals of Statistics.* 47.2, 1148–1178.

[3] Bickel, P. J. and Klaassen, C. A. J. and Ritov, Y. and Wellner, J. A. (1993). *Efficient and adaptive inference in semiparametric models*, Johns Hopkins University Press, Baltimore.

[4] Breiman, L. (2001). Random forests. *Machine learning* 45.1, 5–32.

[5] Gill, R. D. (1994). *Lectures on survival analysis.* Lectures on survival analysis. Springer, Berlin, Heidelberg.

[6] Gill, R. D. and van der Laan, M. J. and Robins, J. M. (1997). Coarsening at random: Characterizations, conjectures, counter-examples. *Proceedings of the First Seattle Symposium in Biostatistics*, 255–294.

[7] Ishwaran, H., Kogalur, U. B. and Blackstone, E. H. and Lauer, M. S. and others. (2008). Random survival forests. *The Annals of Applied Statistics*, 2.3, 841–860.

[8] Rytgaard, H. C. and Gerds, T. A. (2018). Random Forests for Survival Analysis. *Wiley StatsRef: Statistics Reference Online* 1–8.

[9] Tsiatis, A. (2007). *Semiparametric theory and missing data*, Springer Science & Business Media.

[10] van der Laan, M. J. & Robins, J. M. (2003). *Unified methods for censored longitudinal data and causality*, Springer Science & Business Media.

[11] van der Vaart, A. W. (2000). *Asymptotic statistics*, Cambridge university press.

[12] Wager, S., and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association.* 113.523, 1228-1242.

Manuscript IV

Average treatment effects with generalized random forests for survival and competing risks analysis

Helene C. Rytgaard, Claus T. Ekstrøm, Lars V. Kessing and Thomas A. Gerds

Details:

In preparation.

Abstract

In this paper we present a data-adaptive estimation procedure for estimation of average treatment effects in a time-to-event setting based on generalized random forests. In these kinds of settings, the definition of causal effect parameters are complicated by competing risks; here we distinguish between treatment effects on the crude and the net probabilities, respectively. To handle right-censoring, and to switch between crude and net probabilities, we propose a two-step procedure for estimation, applying inverse probability weighting to construct time-point specific weighted outcomes as input for the forest. The forest adaptively handles confounding of the treatment assigned by applying a splitting rule that targets a causal parameter. We demonstrate that our method is effective for a causal search through a list of treatments to be ranked according to the magnitude of their effect. We further apply our method to a dataset from the Danish health registries where it is of interest to discover drugs with an unexpected protective effect against relapse of severe depression.

Ranking of average treatment effects with generalized random forests for time-to-event outcomes

Helene C. W. Rytgaard[1,*], Claus T. Ekstrøm[1], Lars V. Kessing[2], and Thomas A. Gerds[1]

[1]Section of Biostatistics, University of Copenhagen, Oester Farimagsgade 5, Copenhagen Denmark

[2]Copenhagen Affective Disorder research Center (CADIC),

Psychiatric Center Copenhagen, Rigshospitalet, University of Copenhagen

January 28, 2021

1 Introduction

Drug repurposing is an important low-cost method for drug discovery which is typically based on a data-driven experimental approach. In this paper, our general aim is the ability to rank a list of treatment variables according to their effect on a time-to-event outcome. We consider average treatment effect estimation based on generalized random forests in a time-to-event setting with competing risks. We find two aspects particularly important when having to search through a potentially large list of treatments. First, our methods should be as flexible as possible; if we have many treatments, it will be impossible to correctly specify parametric models for the outcome distribution and for all treatment propensities with main effects and interactions. Second, we need a real-valued measure to be used for ranking that should have a sensible interpretation. Compared to other methods for average treatment effect estimation in right-censored and competing risks settings (see, e.g., Ozenne et al., 2019), the methods presented in this paper do not require specification of models for treatment propensity and outcome distribution. Further, we discuss the choice between different causal parameters in the competing risks setting when the aim is to identify new active substances.

Our motivation comes specifically from a large-scale observational registry study on drug purchases and development of psychiatic disorders. Here the goal is to discover if drugs that are already in clinical use may have a protective effect against depression. Psychiatric disorders is a field where the pharmaceutical industry has substantially withdrawn from developing new drugs; thus, in the absence of new randomized clinical trials, and to supplement the expensive and time-consuming generation of data from clinical trials, a systematic search through all drug purchases in the registry data is a cost-efficient way to identify new treatments as well as to discover adverse side-effects. Specific findings can then subsequently be further investigated in randomized trials.

A random forest (Breiman, 2001) is a popular data-driven algorithm that can be used for variable importance analysis, i.e., to rank variables according to their association with the outcome of interest (Ishwaran et al., 2007; Strobl et al., 2008). Mostly, these variable importance measures are based on prediction performance and target the difference in prediction error for a forest that uses the observed version of a specific variable (for which we measure the importance) compared to a forest that uses a randomized version of that variable (Breiman, 2001; Ishwaran et al., 2007). Another measure is the minimal depth (Ishwaran et al., 2010, 2011) which utilizes the distance from the root node of a tree to the first node where there is a split on the variable of interest. The smaller the minimal depth for a given variable, the more important that variable is considered to be.

Our approach in this paper is different in that we consider the use of causal treatment effect parameters as a variable importance measure. Similar approaches have also been considered in the context of high-dimensional biomarker discovery, see, for example, Tuglus and van der Laan (2008); Bembom et al. (2009); Wang and van der Laan (2011). In a counterfactual framework (Neyman, 1923; Rubin, 1974), treatment effect parameters are formally defined as a difference between expected counterfactual outcomes. Under a set of structural and distributional assumptions the parameters are linked to the observed data. We formulate causal parameters in terms of

average differences of event probabilities at pre-specified time horizons of interest, allowing us to report a time-point specific measure of the effect of a particular treatment.

Generalized random forests (GRFs) (Wager and Athey, 2018; Athey et al., 2019) are a recent extension of Breiman's random forests that have been applied to provide nonparametric inference for heterogeneous treatment effects in settings with real-valued and uncensored outcomes of interest. The GRF algorithm is implemented to optimize estimation of the causal treatment effect specifically. Here we implement GRFs for time-to-event outcomes by using inverse probability weighting to make the GRF implementation directly applicable to our setting with right-censoring and competing risks. In the competing risks setting, we further discuss the distinction between treatment effects on crude and net probabilities. These considerations are closely related to the work of Young et al. (2018).

For proof of concept and illustration, we analyze Danish registry data on all Danish citizens who have a first time diagnosis with depression registered. We follow these patients until depression relapse, onset of other mental disorders, death without relapse, or right-censoring. We consider all drugs purchased by any patient in the eight weeks between the depression diagnosis and the start of follow-up. We then apply our two-step procedure separately to each drug and rank the drugs according to the magnitude of their treatment effects.

The article is organized as follows. In Section 2 we introduce the setting and notation for survival and competing risks data. In Section 2.1 we define our target parameters in terms of counterfactual outcomes, and we discuss the distributional assumptions under which we can identify the parameters from the observed data. In Section 3 we review the generalized random forest methodology and present our weighting approach for making the methodology applicable to time-to-event data. In Section 4 we introduce and discuss the use of average treatment effects specifically for the purpose of variable importance analysis. In Section 5 we study the performance using simulated data. In Section 6 we analyze Danish registry data. We close with a discussion in Section 7.

2 Setting and notation

In time-to-event settings subjects are observed from study entry to the occurrence of an event of interest or a competing event. If no event of any kind is observed within the subject-specific follow-up time, the subject is right-censored. Specifically, we consider a competing risks situation with $J \geq 2$ mutually exclusive types of events. For sake of presentation, we assume throughout that $J = 2$. We denote by T_i the uncensored event time, by $\Delta_i \in \{1, 2\}$ the event type and by C_i the censoring time, such that the observed data are $\tilde{T}_i = \min(T_i, C_i)$ and $\tilde{\Delta}_i = \mathbb{1}\{T_i \leq C_i\}\Delta_i$. Moreover, $\boldsymbol{X}_i \in \mathcal{X} \subseteq \mathbb{R}^p$ is a vector of baseline covariates and $\boldsymbol{A}_i = (A_{1,i}, \ldots, A_{K,i}) \in \{0, 1\}^K$ is a vector of $K \in \mathbb{N}$ binary treatment variables. The data consist of $n \in \mathbb{N}$ independent samples, $\big\{(\boldsymbol{X}_1, \boldsymbol{A}_1, \tilde{T}_1, \tilde{\Delta}_1), \ldots, (\boldsymbol{X}_n, \boldsymbol{A}_n, \tilde{T}_n, \tilde{\Delta}_n)\big\}$. We are interested in estimating the effect of the treatment variable $A_k \in \{0, 1\}$ on the probability of events of type $j = 1$. We refer to the other type of events ($j = 2$) as competing events, or competing risks.

We define our target parameter in terms of counterfactuals, using a notation with superscripts to define interventions. In particular, we define T^a as the uncensored counterfactual event time and Δ^a as the corresponding event indicator that would result from setting treatment A_k to a. Further, for $j = 1, 2$, we use $T^{j,a}$ to denote the uncensored counterfactual event time of type j that would result if treatment A_k had been set to a in a hypothetical world where cause j is the only cause. Lastly, we denote by $\pi_k(\boldsymbol{x}) = P(A_k = 1 \,|\, \boldsymbol{X} = \boldsymbol{x})$ the propensity score of treatment A_k conditional on $\boldsymbol{X} = \boldsymbol{x}$, $\boldsymbol{x} \in \mathcal{X}$. Note that we distinguish between the counterfactual event time variable T^a with a single superscript and the counterfactual event time variable $T^{j,a}$ with double superscript. Note also that when studying the treatment A_k, the other treatments can enter the vector of baseline covariates.

2.1 Treatment effects in presence of competing risks

2.1.1 The competing risks problem revisited

As a motivation for our later discussions on causal parameters for treatment effect ranking, we here briefly revisit the problems with causal inference in competing risks settings. In particular, in presence of competing risks, the one-to-one correspondence between the cause-specific hazard and the absolute risk is lost (Andersen et al., 2012), and the effect of variables on the cause-specific hazard may be quite different from their effect on the absolute risk

(Gray, 1988). Specifically, variables may have an indirect effect on the absolute risk only through its effect on the cause-specific hazard of the competing event. Consider the following example.

Example 2.1 *Suppose that it is of interest to rank two treatments, A_1 and A_2, according to their effect on the event of interest. Assume that the cause-specific hazard rates are given by:*

$$\lambda_1(t \mid A_1, A_2) = e^{-0.2A_1 - 0.2A_2},$$
$$\lambda_2(t \mid A_1, A_2) = e^{-0.2A_1},$$

for the event of interest (λ_1) and the competing event (λ_2). Clearly, A_1 and A_2 have the same effect on the hazard of the event of interest. Nonetheless, the cause-specific cumulative incidence of the event of interest also depends on the hazard rate of the competing event:

$$\begin{aligned} F_1(t \mid A_1, A_2) &= \int_0^t \lambda_1(s \mid A_1, A_2) e^{-\int_0^s \sum_{j=1,2} \lambda_j(u|A_1,A_2)du} ds \\ &= \frac{e^{-0.2A_1-0.2A_2}(1 - e^{-t(e^{-0.2A_1-0.2A_2}+e^{-0.2A_1})})}{e^{-0.2A_1-0.2A_2} + e^{-0.2A_1}}. \end{aligned}$$

Now, assume that A_1 and A_2 are both Bernoulli variables with $P(A_k = 1) = 0.5$, $k = 1, 2$. This implies that:

$$\mathbb{E}[F_1(1 \mid 1, A_2) - F_1(1 \mid 0, A_2)] = -0.0137,$$
$$\mathbb{E}[F_1(1 \mid A_1, 1) - F_1(1 \mid A_1, 0)] = -0.061,$$

i.e., the average treatment effects on the cause-specific risk beyond $t = 1$ are very different. Likewise, we see at $t = 2$ that:

$$\mathbb{E}[F_1(2 \mid 1, A_2) - F_1(2 \mid 0, A_2)] = -0.0049$$
$$\mathbb{E}[F_1(2 \mid A_1, 1) - F_1(2 \mid A_1, 0)] = -0.0586,$$

i.e., there is almost no effect of treatment A_1 but a considerable effect of A_2. Thus, in this example, due solely to the effect that A_1 has on the competing cause-specific hazard rate, we would conclude very different effects of the two treatments on the cumulative risk of cause 1.

Recall that our goal is variable importance and the ability to rank a list of treatment variables according to their effect on a specific time-to-event outcome. Example 2.1 illustrates the interpretational issues with absolute risks in the presence of competing risks, and the question is if we would like to conclude different effects for the two treatments A_1 and A_2. This problem is not solved by analyzing the cause-specific hazard rates alone. Specifically, these are defined conditional on post-treatment mechanisms and therefore cannot be ascribed an interpretation as a measure of a causal treatment effect (Hernán, 2010; Martinussen et al., 2018).

In the following we distinguish between effects on *crude* and *net* probabilities, respectively, to characterize the effect of a treatment variable A_k on the occurrence of events of type $j = 1$. We emphasize that the choice between crude and net effects corresponds to the choice between different causal parameters, and altogether depends upon the goal of the analysis. In summary, we argue that:

1. Causal effects on **crude probabilities** are used for describing the real world; crude probabilities allow us to infer on treatment effects that would actually occur in a given population.

2. Causal effects on **net probabilities** are defined in hypothetical worlds without competing risks, and reflect effects of etiological nature. They allow us to infer treatment effects directly on the event type of interest without interference from indirect effects on the competing event time.

Different assumptions on the underlying data-generating mechanisms are necessary when focus is on crude or on net probabilities as we describe in Sections 2.2.1 and 2.3.1, respectively. Importantly, the assumptions needed to identify net probabilities are considerably more ambitious. In Section 2.2, we start by discussing treatment effects on crude probabilities. In Section 2.3, we present treatment effects on net probabilities. In Section 4 we consider a variable importance analysis where the subject matter interest is not in the treatment effects on the crude probability scale; rather, we want to assess treatment effects only directly on the occurrence of type $j = 1$ events.

2.2 Effects on crude probabilities

Recall that the random variables T^0 and T^1 denote the uncensored counterfactual event times that would result if treatment had been set to $A_k = 0$ or $A_k = 1$, respectively. The conditional treatment effect of A_k on the crude risk of events of type 1 before a fixed time horizon $t_0 > 0$ is defined as

$$\begin{aligned}\theta_{\text{crude}}(\boldsymbol{x}) = P(T^1 \leq t_0, \Delta^1 = 1 \,|\, \boldsymbol{X} = \boldsymbol{x}) \\ - \; P(T^0 \leq t_0, \Delta^0 = 1 \,|\, \boldsymbol{X} = \boldsymbol{x}),\end{aligned} \tag{1}$$

for $\boldsymbol{x} \in \mathcal{X}$. The parameter in (1) has a corresponding average,

$$\begin{aligned}\bar{\theta}_{\text{crude}} &= \mathbb{E}[\theta_{\text{crude}}(\boldsymbol{X})] \\ &= P(T^1 \leq t_0, \Delta^1 = 1) - P(T^0 \leq t_0, \Delta^0 = 1),\end{aligned} \tag{2}$$

the average treatment effect (ATE) on the crude risk at time t_0. The quantities $P(T^a \leq t_0, \Delta^a = 1)$, $a = 0, 1$, in (2), referred to as the crude probabilities, are the cumulative incidence functions (Gray, 1988) of the event of interest for a hypothetical treated and a hypothetical untreated population, respectively. These crude probabilities also depend on the hazard rate of the competing event, since, at any time, the event of interest can only occur for subjects who have survived all risks so far. A treatment which reduces the hazard rate of the competing risk increases the event-free survival probability and thereby indirectly increases the crude risk of the event of interest, and vice versa (see also Ishwaran et al., 2014). Particularly, as also illustrated in Example 2.1 of Section 2.1.1, a treatment effect reflected in a non-zero value of $\bar{\theta}_{\text{crude}}$ will occur also if there is only an indirect effect of the treatment on the outcome of interest via the hazard rate of the competing event.

2.2.1 Identifiability of treatment effects on crude probabilities

The average treatment effect on crude probabilities $\bar{\theta}_{\text{crude}}$ is defined in terms of counterfactual random variables such that identifying $\bar{\theta}_{\text{crude}}$ from the observed data requires some distributional assumptions (Hernan and Robins, 2020). First, an assumption of *consistency* entails that the event time under treatment $A_k = a$, T^a, corresponds to the event time we would observe for a subject who was actually observed to be given treatment $A_k = a$. This requires that the treatment A_k can only be administered in one way. Second, an assumption of *no unmeasured confounding* relates both to the treatment and the censoring mechanism. We assume that, conditional on covariates $\boldsymbol{X}$, the counterfactual outcome (T^a, Δ^a) is independent of the observed treatment A_k. Further, we assume that, conditional on covariates $\boldsymbol{X}$ and observed treatment A_k, the outcome (T, Δ) is independent of the censoring time C. Lastly, we assume *positivity*, that $P(C \geq t_0 \,|\, A_k = a, \boldsymbol{X})\,(\pi_k(\boldsymbol{X}))^a(1 - \pi_k(\boldsymbol{X}))^{1-a} > \eta$ for $a = 0, 1$ and some $\eta > 0$.

2.3 Effects on net probabilities

Recall that the counterfactual random variables $T^{1,0}$ and $T^{1,1}$ are the (uncensored) counterfactual event times that would have been observed in a hypothetical world in which cause $j = 1$ is the only cause and where treatment had been set to $A_k = 0$ and $A_k = 1$, respectively. Particularly, $T^{1,0}$ and $T^{1,1}$ are *latent* times that are not always observed in the real world due to cause $j = 2$ events and due to right-censoring. We emphasize that, opposed to the crude risks $P(T^a \leq t_0, \Delta^a = 1)$, $a = 0, 1$, in Equation (2), the net risks $P(T^{1,a} \leq t_0)$, $a = 0, 1$, are not affected by the (indirect) effect that a treatment may have on the hazard rate of the competing risk. They are interpreted as net probabilities for the event of interest in a hypothetical world where the competing event cannot happen. The conditional treatment effect of A_k on the net risk of events of type 1 is defined as follows,

$$\theta_{\text{net}}(\boldsymbol{x}) = P(T^{1,1} \leq t_0 \,|\, \boldsymbol{X} = \boldsymbol{x}) - P(T^{1,0} \leq t_0 \,|\, \boldsymbol{X} = \boldsymbol{x}), \tag{3}$$

for $\boldsymbol{x} \in \mathcal{X}$, with the corresponding average,

$$\bar{\theta}_{\text{net}} = \mathbb{E}[\theta_{\text{net}}(\boldsymbol{X})] = P(T^{1,1} \leq t_0) - P(T^{1,0} \leq t_0), \tag{4}$$

which is the average treatment effect (ATE) on the net probability. Notably, a treatment effect reflected in a non-zero value $\bar{\theta}_{\text{net}}$ will only occur if the studied treatment has a direct effect on the event of interest, whereas a treatment effect reflected in a non-zero value $\bar{\theta}_{\text{crude}}$ will occur also if there is an indirect effect of the treatment on the outcome of interest via the hazard rate of the competing event.

2.3.1 Identifiability of treatment effects on net probabilities

Identification of $\bar{\theta}_{\text{net}}$ from observed data requires additional assumptions to the ones stated in Section 2.2.1. Notably, $\bar{\theta}_{\text{net}}$ are formulated in terms of an extra layer of counterfactual reasoning, as it is defined in terms of counterfactual event times $T^{1,a}$ we would had seen has there been no occurrences of competing events. *Consistency* requires that the counterfactual time to event $j = 1$, $T^{1,a}$, corresponds to the actual event time for a subject who remained uncensored, free of competing events and was observed to be given treatment $A_k = a$. *Positivity* additionally includes the probability of competing events being bounded away from zero: $P(T^{2,A_k} \geq t_0 \,|A_k = a, \boldsymbol{X}) > \eta'$, for $a = 0, 1$ and for some $\eta' > 0$. Most critical is the additional assumption of *no unmeasured confounding* for the competing event time T^{2,A_k}; it is needed that T^{1,A_k} and T^{2,A_k} are conditionally independent given covariates $\boldsymbol{X}$ and treatment A_k. As previously mentioned, we stress that this is a very strong assumption: Whether A_k and $\boldsymbol{X}$ together include all factors that we believe to be predictive of both event types depends very much on the nature of the competing events and how rich the measured set of covariates is.

3 Generalized random forests with inverse probability weighted outcomes

Generalized random forests (GRFs) (Athey et al., 2019) are a recent generalization of the original random forest algorithm (Breiman, 2001), a machine learning tool that adaptively searches the covariate space by recursive sample splitting.

Generally, a forest consists of $B \in \mathbb{N}$ randomized trees, where the bth tree of the forest is grown by recursively splitting the covariate space according to some split criterion. GRFs provide a data-adaptive approach to estimation of average treatment effects for uncensored data, particularly, for a generic outcome variable $Y \in \mathbb{R}$,

$$\theta(\boldsymbol{x}) = \mathbb{E}[Y \mid A_k = 1, \boldsymbol{X} = \boldsymbol{x}] - \mathbb{E}[Y \mid A_k = 0, \boldsymbol{X} = \boldsymbol{x}].$$

A key part of the generalized random forest algorithm is the splitting rule that targets specifically the estimation of the quantity of interest $\theta(\boldsymbol{x})$. Each tree applies a splitting rule that adaptively makes binary partitions of the covariate space such as to maximize heterogeneity in $\theta(\boldsymbol{x})$. By averaging over neighborhoods defined by the trees, the forest produces a neighborhood function that is used as a kernel for the estimation of $\theta(\boldsymbol{x})$. In the Supplementary Material (Appendix C) we describe the local gradient-based criterion for making splits and the kernel-based estimator for $\theta(\boldsymbol{x})$ as proposed by Athey et al. (2019), and we further review the structural model formulation of treatment effects of Athey et al. (2019, Section 6) and its relation to our setting with the counterfactual formulation.

The problem in our setting is that we do not observe the actual outcomes of interest. For the parameter $\bar{\theta}_{\text{crude}}$, for example, we do not observe $Y := \mathbb{1}\{T \leq t_0, \Delta = 1\}$ due to right-censoring. In this section we assume that we are given a conditional distribution function G such that $G(t \,|\, A_k, \boldsymbol{X}) = P(C > t \,|\, A_k, \boldsymbol{X})$. Based on G, we define the inverse probability weighted outcome,

$$\tilde{Y} := \frac{\mathbb{1}\{\tilde{T} \leq t_0, \tilde{\Delta} = 1\}}{G(\tilde{T}- \,|\, A_k, \boldsymbol{X})}. \tag{5}$$

For this outcome, we show in Section 3.1 below that

$$\begin{aligned}
&\theta_{\text{crude}}(\boldsymbol{x}) \\
&\quad = P(T^1 \leq t_0, \Delta^1 = 1 \,|\, \boldsymbol{X} = \boldsymbol{x}) - P(T^0 \leq t_0, \Delta^0 = 1 \,|\, \boldsymbol{X} = \boldsymbol{x}) \\
&\quad = \mathbb{E}\big[\tilde{Y} \,\big|\, \boldsymbol{X} = \boldsymbol{x}, A_k = 1\big] - \mathbb{E}\big[\tilde{Y} \,\big|\, \boldsymbol{X} = \boldsymbol{x}, A_k = 0\big].
\end{aligned}$$

The idea is that we can apply GRFs directly to our weighted outcome $\tilde{Y}$. This provides an estimator $\hat{\theta}_{\text{crude}}(\boldsymbol{x})$ for $\theta_{\text{crude}}(\boldsymbol{x})$ and thereby an estimator for the corresponding average effect

$$\hat{\bar{\theta}}_{\text{crude}} = \frac{1}{n} \sum_{i=1}^{n} \hat{\theta}_{\text{crude}}(\boldsymbol{X}_i). \tag{6}$$

This leads to the following two-step approach:

Step 1. The conditional distribution function G is estimated based on the full dataset and is used to construct the weighted outcome $\tilde{Y}$ as defined by Equation (5).

Step 2. A generalized random forest is applied with $\tilde{Y}$ as outcome, yielding estimates $\hat{\theta}_{\text{crude}}(\boldsymbol{x})$, $\boldsymbol{x} \in \mathcal{X}$, and the ATE is then estimated simply by averaging as in Equation (6).

An equivalent two-step approach is utilized to estimate the effect on net probabilities. We note that this requires, in addition to an estimator for the conditional distribution G, an estimator for the conditional distribution function G_2 such that $G_2(t \mid A_k, \boldsymbol{X}) = P(T^{2,a} > t \mid A_k, \boldsymbol{X})$, and construction of the inverse probability weighted outcome

$$\tilde{Y}' := \frac{\mathbb{1}\{\tilde{T} \leq t_0, \tilde{\Delta} = 1\}}{G(\tilde{T}- \mid A_k, \boldsymbol{X}) G_2(\tilde{T}- \mid A_k, \boldsymbol{X})}. \tag{7}$$

Thus, to construct the weights, we need to model the survival functions of both the latent time to a competing risk event and the censoring time.

3.1 Identifiability by inverse probability weighting

The assumptions stated in Section 2.2.1 allow us to link the distribution of the counterfactual variables to the observed data distribution. Since,

$$\begin{aligned}\mathbb{E}\big[\tilde{Y} \,\big|\, \boldsymbol{X}, A_k\big] &= \mathbb{E}\left[\frac{\mathbb{1}\{\tilde{T} \leq t_0, \tilde{\Delta} = 1\}}{G(\tilde{T}- \mid A_k, \boldsymbol{X})} \,\middle|\, \boldsymbol{X}, A_k\right] \\ &= \mathbb{E}\big[\mathbb{1}\{T \leq t_0, \Delta = 1\} \mid A_k, \boldsymbol{X}\big],\end{aligned}$$

it follows that,

$$\begin{aligned}&\bar{\theta}_{\text{crude}}(\boldsymbol{x}) \\ &= P(T^1 \leq t_0, \Delta^1 = 1 \mid \boldsymbol{X} = \boldsymbol{x}) - P(T^0 \leq t_0, \Delta^0 = 1 \mid \boldsymbol{X} = \boldsymbol{x}) \\ &= P(T \leq t_0, \Delta = 1 \mid \boldsymbol{X} = \boldsymbol{x}, A_k = 1) \\ &\qquad\qquad - P(T \leq t_0, \Delta = 1 \mid \boldsymbol{X} = \boldsymbol{x}, A_k = 0) \\ &= \mathbb{E}[\tilde{Y} \mid \boldsymbol{X} = \boldsymbol{x}, A_k = 1] - \mathbb{E}[\tilde{Y} \mid \boldsymbol{X} = \boldsymbol{x}, A_k = 0].\end{aligned}$$

Similarly, we identify $\theta_{\text{net}}(\boldsymbol{x})$. More details can be found in the Supplementary Material (Appendix B).

3.2 Estimation of inverse probability weights

To implement our two-step approach, we need consistent estimators for the nuisance parameters G and G_2 on $[0, t_0]$. We here describe an approach based on the reverse Kaplan-Meier estimator stratified on a subset of categorical covariates $\boldsymbol{Z} \subset \{A_k, \boldsymbol{X}\}$. This approach is appropriate in our illustrative data example (Section 6), whereas other settings may require more sophisticated approaches as we discuss in Section 7. Based on the Kaplan-Meier approach, we estimate the censoring survival distribution function G, conditional on $\boldsymbol{Z}$, as follows,

$$\hat{G}(t \mid \boldsymbol{z}) = \prod_{t_k \leq t} \left(1 - \frac{\sum_{i=1}^n \mathbb{1}\{T_i = t_k, \Delta_i = 0, \boldsymbol{Z}_i = \boldsymbol{z}\}}{\sum_{i=1}^n \big(\mathbb{1}\{T_i \geq t_k\} - \mathbb{1}\{T_i = t_k, \Delta_i > 0, \boldsymbol{Z}_i = \boldsymbol{z}\}\big)}\right).$$

We handle ties in the event times with the usual convention that the event of interest happens before competing events and censoring events. Similarly, we may estimate G_2 with Kaplan-Meier estimator for the competing event time conditional on $\boldsymbol{Z}$,

$$\hat{G}_2(t \mid \boldsymbol{z}) = \prod_{t_k \leq t} \left(1 - \frac{\sum_{i=1}^n \mathbb{1}\{T_i = t_k, \Delta_i = 2, \boldsymbol{Z}_i = \boldsymbol{z}\}}{\sum_{i=1}^n (\mathbb{1}\{T_i \geq t_k\} - \mathbb{1}\{T_i = t_k, \Delta_i \neq 2, \boldsymbol{Z}_i = \boldsymbol{z}\})}\right).$$

Under the working assumption that $G(t \mid A_k, \boldsymbol{X}) = P(C > t \mid A_k, \boldsymbol{X}) = P(C > t \mid \boldsymbol{Z})$, standard arguments (e.g. Andersen et al., 1993) lead to $\hat{G}(t|\boldsymbol{Z}) \to G(t|\boldsymbol{Z})$ a.s. as $n \to \infty$ for all $t \leq t_0$, and likewise for $\hat{G}_2(t|\boldsymbol{Z})$. However, violation of the working assumption may lead to asymptotic bias in Step 1 of our two-step approach which may also lead to bias in the ranking of the treatment variables. In Section 7, we discuss the bias-variance trade-off and how one may relax the working assumptions.

4 Variable importance

Suppose we have a list of treatments, $A_1, A_2, \ldots, A_K$, $K \in \mathbb{N}$, that we would like to rank according to either their crude or their net effect. We assume that the treatments are binary variables with $A_k = 1$ indicating treatment and $A_k = 0$ no treatment, $k = 1, \ldots, K$.

We will consider the use of our data-adaptive generalized random forest estimation procedure to rank the list of treatment variables. Specifically, we continue our discussion from Section 2.1.1 to distinguish between crude and net probabilities for the purpose of ranking. The problem with crude probabilities is that they reflect a mixture of effects on the hazard rate of the event of interest and effects on the hazard rate of the competing risks. Net probabilities, on the other hand, describe a hypothetical world without competing events and have thereby been criticized (Andersen and Keiding, 2012). Nevertheless, they allow us to study the effect of a particular drug in a way that is independent of the effect that this drug may have on the hazard rate of the competing events. We argue that for the purpose of drug discovery, it may be desirable to restrict the search to drugs that would have an effect in the hypothetical world where all competing causes are eliminated.

To obtain a ranking of the treatments, we apply the two-step approach of Section 3 which yields estimates $\hat{\bar{\theta}}_{\text{net},k}$ for the treatment effects on the net probability scale for all drugs A_k, $k = 1, \ldots, K$. For comparison and illustration, we also compute estimates $\hat{\bar{\theta}}_{\text{crude},k}$ for the treatment effect on the crude probability scale. A standard delta method argument using the standard errors $\hat{\sigma}_n(\boldsymbol{x})$ for the conditional estimates (as provided by Athey et al., 2019, Theorem 5 and Section 6) yields asymptotic normality of the forest estimators $\hat{\bar{\theta}}_{\text{net},k}$, $\hat{\bar{\theta}}_{\text{crude},k}$ for the average treatment effects, based on which we construct confidence intervals. Generally, we say that a treatment A_k has a protective effect if the upper confidence limit is below zero, a harmful effect if the lower confidence limit is above zero, and a neutral effect if zero is contained in the confidence interval.

Clearly, the asymptotic standard error also contains a contribution from the uncertainty of the weights constructed in Step 1 of our procedure. However, in our experience, these contributions are often very small in real data applications. In our simulations and illustrative data analysis, we only show confidence intervals which ignore the statistical uncertainty due to Step 1 and thus rely solely on the theory that applies for construction of confidence intervals for forest estimation without this step. Despite these shortcomings, we note that in our simulation studies (Section 5), the coverage of the confidence intervals lies nicely around 95%.

5 Simulation study

To evaluate the performance of our proposed methodology, and as a proof of concept, we test our algorithm on simulated data. Our simulations further illustrate the difference between treatment effects on the crude and net probability scales. We here explain the design of the simulations. Further details in the form of `R`-code can be found on github, see Section 8. We start by simulating covariates, $\boldsymbol{X} = (X_1, \ldots, X_6)$. We let X_1, X_4, X_5, X_6 be uniformly distributed on the unit interval $(0, 1)$, X_2 be categorical with three ordered categories, and X_4 be categorical with four ordered categories. We consider a setting where we compare $K = 10$ treatment variables drawn from Bernoulli distributions that are all dependent on one of the covariates, $\mathbb{E}[A_k \mid \boldsymbol{X}] = \text{expit}(\beta_0^k + \beta_1^k X_{l_k})$, with $l_k \in \{1, \ldots, 6\}$.

Given treatments and covariates, three latent event times T^1, T^2, C are simulated according to Weibull distributions. The Weibull distribution of the latent censoring time is specified independently of covariate and treatment variables. The Weibull distribution of the latent time to the event of interest is specified with a shape parameter dependent on X_1, X_3 and A_1. The Weibull distribution of the latent competing event time is specified with a shape

parameter dependent on X_1, X_2 and A_2. Our simulation design can thus be summarized as follows:

$$\begin{aligned}
&\text{Event of interest:} && T^1 \sim A_1 + X_1 + X_3 \\
&\text{Competing event:} && T^2 \sim A_2 + X_1 + X_2 \\
&\text{Censoring:} && C \sim 1
\end{aligned}$$

We simulate counterfactuals such that we know the true value of the average treatment effect parameters, $\bar{\theta}_{\text{net}}, \bar{\theta}_{\text{crude}}$: That is, we draw from the distributions in the hypothetical scenarios where we control treatment assignment and occurrence of censoring and, for the net effects, competing risk events.

Throughout this section, we focus on three of the treatment variables: A_1 that has a direct effect on the event of interest, A_2 that has an effect only on the competing event, and A_3 that has no effect at all. The true values of $\bar{\theta}_{\text{net},A_k}, \bar{\theta}_{\text{crude},A_k}$, $k = 1, 2, 3$, are as follows:

$$\begin{aligned}
&\text{Effects on net probabilities:} && \bar{\theta}_{\text{net},A_1} = -0.113, \\
& && \bar{\theta}_{\text{net},A_2} = 0, \\
& && \bar{\theta}_{\text{net},A_3} = 0. \\
&\text{Effects on crude probabilities:} && \bar{\theta}_{\text{crude},A_1} = -0.083, \\
& && \bar{\theta}_{\text{crude},A_2} = -0.047, \\
& && \bar{\theta}_{\text{crude},A_3} = 0.
\end{aligned}$$

Our aim is to show that weighting yields unbiased estimation of the ATEs and further to explore the effect of confounding and sample size. Our simulations consist of the following two parts:

1. *Effect estimation and coverage.* We simulate $M = 1000$ datasets with sample size $n = 1000$ from the data-generating distribution. We look at effect estimates and coverage of the confidence intervals based on the standard error estimates provided by the forest.

2. *Ranking effectiveness.* For sample sizes $n \in \{100, 200, 500, 1000, 1500, 2000\}$, we simulate $M = 500$ datasets from the data-generating distribution. For each dataset, we use our algorithm to estimate the variable importance of the treatments $A_1, \ldots, A_{10}$, in form of estimates $\hat{\bar{\theta}}^m_{\text{net},A_k}$ and $\hat{\bar{\theta}}^m_{\text{crude},A_k}$ for $k = 1, \ldots, 10$ and $m = 1, \ldots, M$. For A_k, $k = 1, \ldots, 10$, we define,

$$\mathcal{R}^M_{\text{net}}(A_k) := \frac{1}{M} \sum_{m=1}^{M} \prod_{k' \neq k} \mathbb{1}\{\hat{\bar{\theta}}^m_{\text{net},A_k} \leq \hat{\bar{\theta}}^m_{\text{net},A_{k'}}\}, \tag{8}$$

$$\mathcal{R}^M_{\text{crude}}(A_k) := \frac{1}{M} \sum_{m=1}^{M} \prod_{k' \neq k} \mathbb{1}\{\hat{\bar{\theta}}^m_{\text{crude},A_k} \leq \hat{\bar{\theta}}^m_{\text{crude},A_{k'}}\}, \tag{9}$$

as the fraction of simulation repetitions (out of $M = 500$) where the treatment variable A_k is ranked "most important" among $A_1, \ldots, A_{10}$ in terms of the effect on net and crude probabilities, respectively. We report the ability of our method to, for instance, detect treatment A_1 as the "most important" variable among $A_1, \ldots, A_{10}$.

We consider three different adjustment schemes for the inverse probability weight estimation:

(a) Weight estimators $\hat{G}, \hat{G}_2$ that are adjusted for A_2, X_1 and X_2, i.e., $\boldsymbol{Z} = \{A_2, X_1, X_2\}$.

(b) Weight estimators $\hat{G}, \hat{G}_2$ that are adjusted only for A_2, i.e., $\boldsymbol{Z} = \{A_2\}$.

(c) Weight estimators $\hat{G}, \hat{G}_2$ that are unadjusted, i.e., $\boldsymbol{Z} = \{1\}$.

The weight estimators are constructed outside the forest in Step 1 of our two-step procedure as described in Section 4. Based on the weights, a separate (GRF) forest is applied for each treatment variable A_k to estimate $\bar{\theta}_{\text{net},A_k}$ and $\bar{\theta}_{\text{crude},A_k}$, for $k = 1, \ldots, 10$. We use $B = 1000$ trees to construct each GRF estimate. The parameters $\bar{\theta}_{\text{net}}$ and $\bar{\theta}_{\text{crude}}$ are defined with time horizon $t_0 = 0.5$. For $n = 1000$ we had on average (across $M = 1000$ datasets) 394 competing events, 145 censoring events and 129 events of interest observed before time t_0.

5.1 Simulation results

5.1.1 Effect estimation and coverage

Figure 2 shows mean estimates across $M = 1000$ simulated datasets using adjustment schemes (a)–(c) for estimation of inverse probability weights. Using adjustment scheme (a), treatment A_1 is correctly shown to have a protective effect, both on the scale of net probabilities and on the scale of crude probabilities. On the other hand, whether we conclude a protective effect of A_2 depends on whether we focus on the net probabilities or on the crude probabilities. Confidence intervals all have a coverage around 95% despite the fact that the standard errors do not take the uncertainty of the weight estimation into account. This result was further challenged using varying sample size and for a varying amount of censored observations, but we found no systematic relationship (results not shown).

In adjustment scheme (b) the weights are only adjusted for A_2, but Figure 2 shows that we still achieve 95% coverage with our confidence intervals. Comparing the results for adjustment schemes (b) and (c) in Figure 2 shows that it is crucial to include treatment A_2 in $\boldsymbol{Z}$ in the weight estimation for estimating $\bar{\theta}_{\text{net}}$: Adjustment scheme (c) uses unadjusted estimators for the inverse probability weights and incorrectly estimate a protective effect of treatment A_2 on the net probabilities.

Estimation of $\bar{\theta}_{\text{crude}}$ is hardly affected across the weighting schemes (a)–(c) since the censoring times were generated independent of all treatment and covariate variables. Of course, we can produce biased results for $\bar{\theta}_{\text{crude}}$ with the unadjusted weighting scheme if we let the censoring mechanism depend on treatment variables and covariates.

Across all adjustment schemes (a)–(c), the treatment A_3 is correctly shown to have no effect both in terms of crude and net probabilities. Furthermore, the treatment A_1 is correctly shown to have a protective effect. The coverage of the confidence intervals for these two treatment variables are hardly affected by varying the adjustment set $\boldsymbol{Z}$ for constructing weight estimators.

5.1.2 Ranking effectiveness

Figure 2 shows the fractions $\mathcal{R}^M_{\text{net}}(A_1)$ and $\mathcal{R}^M_{\text{crude}}(A_1)$ as defined in Equations (8) and (9) across different sample sizes. We show only the results from using adjustment scheme (b) and adjustment scheme (c) for estimating the inverse probability weights, as the results for weighting scheme (a) and (b) are similar.

Recall that $\mathcal{R}^M_{\text{net}}(A_k)$ is the fraction of simulation repetitions (out of $M = 500$ total repetitions) where A_k is ranked most important among $A_1, \ldots, A_{10}$ in terms of their effect on the difference in net probabilities. Likewise, $\mathcal{R}^M_{\text{crude}}(A_k)$ is the fraction of simulation repetitions where A_k is ranked most important among $A_1, \ldots, A_{10}$ in terms of their effect on the difference in crude probabilities. Figure 2 shows $\mathcal{R}^M_{\text{net}}(A_k)$ and $\mathcal{R}^M_{\text{crude}}(A_k)$ for each of the three treatment variables A_1, A_2 and A_3. We would like $\mathcal{R}^M_{\text{net}}(A_1), \mathcal{R}^M_{\text{crude}}(A_1)$ to be close to one, and $\mathcal{R}^M_{\text{net}}(A_2), \mathcal{R}^M_{\text{crude}}(A_2), \mathcal{R}^M_{\text{net}}(A_3), \mathcal{R}^M_{\text{crude}}(A_3)$ to be close to zero. We further expect $\mathcal{R}^M_{\text{crude}}(A_2)$ to be larger than $\mathcal{R}^M_{\text{net}}(A_2)$, due to the effect of A_2 on the competing risk event.

Figure 2 shows that both $\mathcal{R}^M_{\text{net}}(A_1)$ and $\mathcal{R}^M_{\text{crude}}(A_1)$ approach one as the sample size n increases: The larger the sample size, the more certain we are to detect the important variable A_1. On the other hand, it also shows that $\mathcal{R}^M_{\text{net}}(A_1)$ and $\mathcal{R}^M_{\text{crude}}(A_1)$ are both rather small for $n = 100$ and $n = 200$. Evidently, we need a certain sample size to be able to detect important variables with high probability.

Across all sample sizes we have that $\mathcal{R}^M_{\text{net}}(A_3), \mathcal{R}^M_{\text{crude}}(A_3)$ are both very small, consistent with the fact that A_3 has no effect at all ($\bar{\theta}_{\text{net},A_3} = \bar{\theta}_{\text{crude},A_3} = 0$). The same is seen for $\mathcal{R}^M_{\text{net}}(A_2)$, except in the scenario where we fail to adjust for A_2 in the estimation of inverse probability weights. At last we note that $\mathcal{R}^M_{\text{crude}}(A_2)$ is overall larger than $\mathcal{R}^M_{\text{net}}(A_2)$, as we would expect.

6 Registry study

We apply our method to our motivating example in which it is of interest to study whether the use of any particular drug decreases the absolute risk of relapse of depression resulting in psychiatric hospitalization. Our aim is to discover new active substances; here net probabilities will allow us to rank drugs according to their direct effect

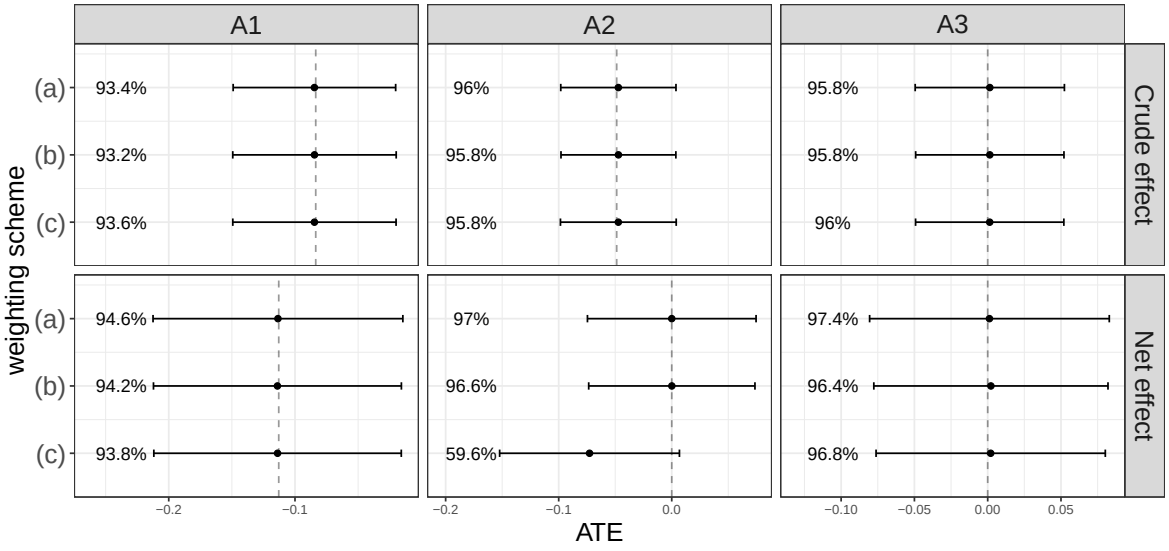


Figure 1: Results of simulation studies. Shown are the results from estimation of $\bar{\theta}_{\text{net},A_k}$ and $\bar{\theta}_{\text{crude},A_k}$, $k = 1, 2, 3$, across $M = 500$ repetitions (all with sample size $n = 500$). The true values are marked by the dashed gray lines. Note that A_2 has an effect on the difference in crude probabilities (true effect $\bar{\theta}_{\text{crude},A_2} = -0.0487$ through the effect on the competing risk event whereas it has no effect on the difference in net probabilities (true effect $\bar{\theta}_{\text{net},A_2} = 0$). The right column shows the coverage, i.e, the fraction of simulations where the confidence interval constructed based on the forest estimate of the standard error contains the true value. In weighting scheme (a) and (b) we used weights that were adjusted for $\boldsymbol{Z} = \{A_2, X_1, X_2\}$ or $\boldsymbol{Z} = \{A_2\}$, both resulting in unbiased estimators. In weighting scheme (c) we used unadjusted weights ($\boldsymbol{Z} = \{1\}$), inducing severe bias in the estimate of $\bar{\theta}_{\text{net},A_2}$. We do not show the results for $A_4, \ldots, A_{10}$ as these are similar to those for A_3.

on depression, isolating this effect from what effect that drug may have on competing events. We compare our estimates of effects on the net probabilities to those on the crude probabilities to investigate their difference.

6.1 Description of data

Data are obtained by linking Danish population-based registers that contain data on all prescribed medical purchases at pharmacies since 1995 and data on all patients treated at hospitals since 1977. A total of 78,700 patients were included who all had a first-time admission with depression after 2005.

Figure 3 illustrates our design. The date of first contact with depression is defined as the index date. Patients with a psychiatric hospitalization in the eight weeks window following the index date are excluded. We group ATC drug codes after their first three digits and define binary exposure variables with the value 1 if there was at least one prescribed purchase within the ATC group in the eight weeks window. Information on conmorbidity is collected during a ten year period before the index date and included as covariates in the analysis, along with sex and age at the index date. Subjects are followed for five years from the end of the exposure window until depression relapse ($\Delta = 1$), a competing event ($\Delta = 2$), or loss to follow-up ($\Delta = 0$). Summary statistics on comorbidities, exposure and number of events can be found in the Supplementary Material (Appendix D).

6.2 Analysis of data

To estimate the treatment effect of each considered treatment A_k on the net and crude probabilities, $\bar{\theta}_{\text{net},A_k}$ and $\bar{\theta}_{\text{crude},A_k}$, we adjust the inverse probability weights for sex, age group and the treatment A_k itself. In the forest we use $B = 200$ trees, and we include sex, age group and all comorbidities as covariates.

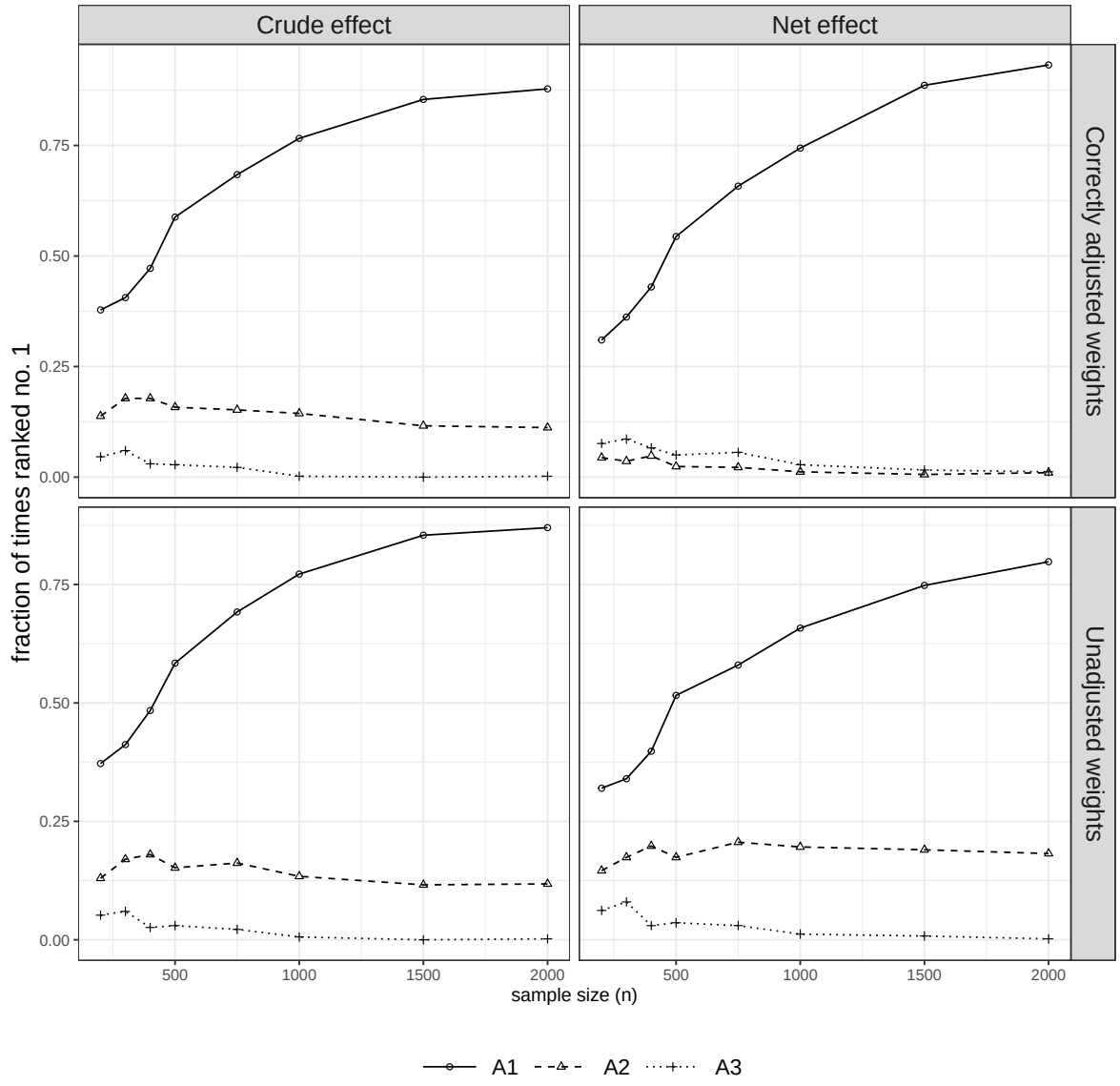


Figure 2: Results of simulation studies. Shown are the fraction of times that each the three treatment variables A_1, A_2, A_3 was ranked most important (across $M = 500$ simulation repetitions) in terms of either the effect on the difference in net probabilities ($\bar{\theta}_{\text{net}}$) or the effect on the difference in crude probabilities ($\bar{\theta}_{\text{crude}}$). In the left plot, estimation of the inverse probability weights were adjusted for A_2 (weighting scheme (b)). In the right plot, we used unadjusted estimators (weighting scheme (c)) for the inverse probability weights.

6.3 Results

Figure 4 shows the causal forest estimates of the effect on net probabilities, $\bar{\theta}_{\text{net}}$, and of the effect on crude probabilities, $\bar{\theta}_{\text{crude}}$, for each drug group. We distinguish between a protective effect (if the upper confidence limit is below zero), a harmful effect (if the lower confidence limit is above zero), and a neutral effect (if zero is contained in the confidence interval). The size of the estimates allows us to rank the treatment groups according to their effect on relapse with depression.

Specifically, recall that $\bar{\theta}_{\text{net}}$ is the treatment effect in the hypothetical world without competing events and

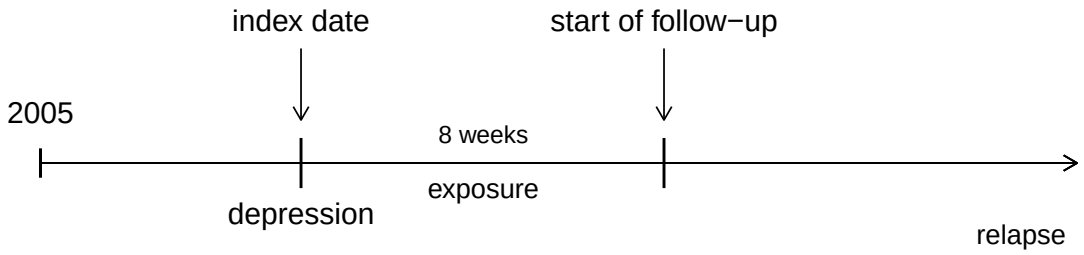


Figure 3: Illustration of our study design. The date of first contact with depression is defined as the index date. Patients with a psychiatric hospitalization in the eight weeks window following the index date are excluded. ATC drug codes are grouped after their first three digits to define binary exposure variables with the value 1 if there was at least one prescribed purchase within the ATC group in the eight weeks window. Information on comorbidity is collected during a ten year period before the index date and included as covariates in the analysis, along with sex and age at the index date.

$\bar{\theta}_{\text{crude}}$ the treatment effect in the real world. When looking for new active substances, we do not wish to report a large treatment effect of a particular drug if this drug effect only came through an effect on a competing risk event. Rather, we would like to rank the drugs according to the direct effect that the drugs may have as represented by the parameter $\bar{\theta}_{\text{net}}$. As we saw in the simulation study, as well as in Example 2.1, there can be a substantial difference between $\bar{\theta}_{\text{net}}$ and $\bar{\theta}_{\text{crude}}$.

Here we see in Figure 4, as well, that the estimates of the two parameters lead to slightly differing conclusions. Consider, for example, the drug group 'A12' (mineral supplements). This drug group is ranked higher in terms of net probabilities than in terms of crude probabilities (although the effect remains insignificant in both cases). On the other end of the spectrum, some drug groups are deemed harmful in terms of their effect on crude probabilities and neutral in terms of their effect on net probabilities: 'A10' (antidiabetics) and 'C10' (lipid modifying agents).

7 Discussion

In this paper we have considered average treatment effect estimation in a time-to-event setting for the purpose of ranking treatments according to their effect on a specific outcome of interest. Particularly, we have discussed the use of two different parameters in the presence of competing risks, defined in terms of net and crude probabilities, respectively, with different interpretations.

In the present paper, we have implemented a data-adaptive estimation method based on generalized random forest, where inverse probability weights are constructed to move from a crude to a net interpretation and to make the forest implementation directly applicable to the time-to-event setting.

Our method makes no parametric model restrictions and benefits from the flexibility of the generalized random forest which adaptively adjusts the propensity of treatment for covariates. However, a weakness of our presented analysis is the use of the Kaplan-Meier method for constructing the inverse probability weights. This may work in large scale registry data where most variables are categorical and a large amount of data are available to estimate the weights separately in all strata defined by the covariates. However, in other applications it may be necessary to

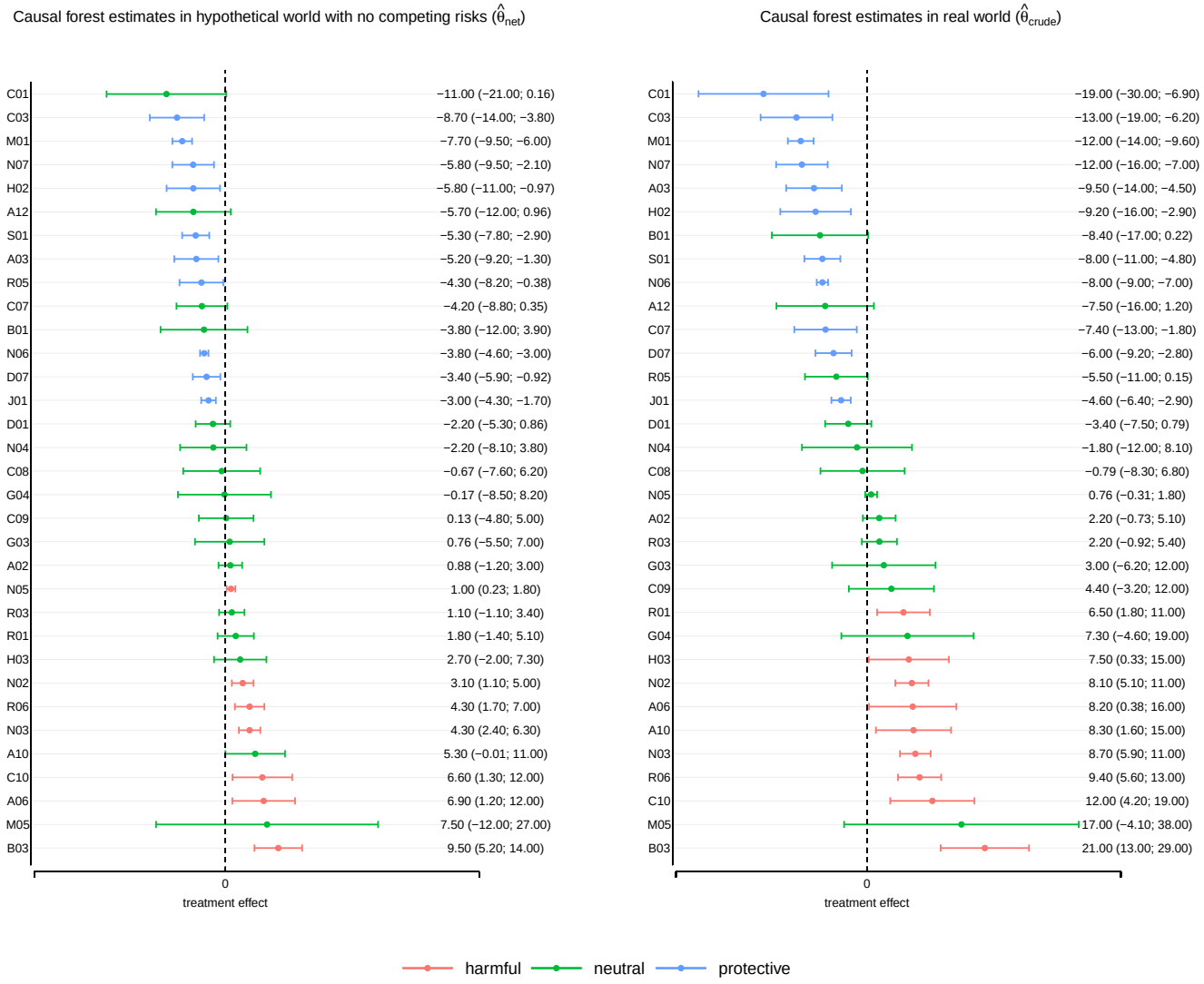


Figure 4: *Left:* Causal forest estimates of $\bar{\theta}_{\text{net}}$ (using adjusted weights to construct weighted outcomes). *Right:* Causal forest estimates of $\bar{\theta}_{\text{crude}}$ (using adjusted weights to construct weighted outcomes). For each ATC group (marked on the x-axis) the plot shows the estimates and the estimated confidence intervals (numbers written on the right). The colors indicate the direction of the effect.

allow that several continuous covariates affect the distributions G and G_2. Semiparametric theory tells us to use a flexible model and to include all covariates that affect the event time to improve robustness and efficiency (van der Laan and Robins, 2003). However, to achieve proper bias-variance trade-off for the target parameter in the second step, the random survival forest used for the weights must be undersmoothed.Another idea for improvement is to handle the weight estimation inside the forest in a one-step approach. Indeed, we may improve upon the current setting by implementing the splitting rule based on the efficient influence function (Robins and Rotnitzky, 1992; van der Laan and Robins, 2003), extending the methods of Rytgaard (2019) to the competing risks setting. In future work we follow this route and revise the implementation of GRFs to adapt it to the event history analysis setting as proposed by Rytgaard (2019).

As an illustration, we have considered a particular application where it is of interest to rank a list of treatments according to their effect on depression. Here we argue for the use of treatment effects on net probabilities: When discussing variable importance and discovering new drugs to treat a disease, it is not of interest to conclude that a variable has a causal effect on the event of interest if that effect is only due to an increased risk of competing events. On the other hand, we emphasize that treatment effects on crude probabilities should be considered if interest is in the real world and the aim is to predict for a given patient. Importantly, net probabilities are meant for ranking drugs when looking for new active substances as part of a drug discovery study, but they are not sensible interpreting the size of the effect, e.g., when counseling a patient.

Another potential future avenue for defining causal effects in the competing risks setting was proposed recently by Stensrud et al. (2019). In their paper, they discuss a different parametrization with a similar aim to isolate the direct effect on the event of interest. Particularly, they assume that the treatment mechanism can be split into two parts, of which only one affects the event of interest and the other only the competing event, such that they can consider a hypothetical scenario in which only the former part of the treatment is changed. However, in our application where the event of interest is depression and the competing risk events include a diagnosis of bipolar disorder, it may very well be the case that the treatment mechanism affecting the one is in fact the same as the one affecting the other.

8 Supplementary Material

`R` code is available on github (`https://github.com/helenecharlotte/grfCausalSearch`). The supplementary material consists of Appendices A–D.

References

Aalen, O. O. and Johansen, S. (1978). An empirical transition matrix for non-homogeneous markov chains based on censored observations. *Scandinavian Journal of Statistics* pages 141–150.

Andersen, P. K., Borgan, Ø., Gill, R. D., and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer Series in Statistics. Springer, New York.

Andersen, P. K., Geskus, R. B., de Witte, T., and Putter, H. (2012). Competing risks in epidemiology: possibilities and pitfalls. *International journal of epidemiology* **41,** 861–870.

Andersen, P. K. and Keiding, N. (2012). Interpretability and importance of functionals in competing risks and multistate models. *Statistics in medicine* **31,** 1074–1088.

Athey, S., Tibshirani, J., and Wager, S. (2019). Generalized random forests. *The Annals of Statistics* **47,** 1148–1178.

Bembom, O., Petersen, M. L., Rhee, S., Fessel, W. J., Sinisi, S. E., Shafer, R. W., and van der Laan, M. J. (2009). Biomarker discovery using targeted maximum-likelihood estimation: Application to the treatment of antiretroviral-resistant hiv infection. *Statistics in medicine* **28,** 152–172.

Breiman, L. (2001). Random forests. *Machine learning* **45,** 5–32.

Gill, R. D., van der Laan, M. J., and Robins, J. M. (1997). Coarsening at random: Characterizations, conjectures, counter-examples. In *Proceedings of the First Seattle Symposium in Biostatistics*, pages 255–294. Springer.

Gray, R. J. (1988). A class of k-sample tests for comparing the cumulative incidence of a competing risk. *The Annals of statistics* **16,** 1141–1154.

Hernán, M. A. (2010). The hazards of hazard ratios. *Epidemiology (Cambridge, Mass.)* **21,** 13.

Hernan, M. A. and Robins, J. M. (2020). *Causal Inference*. Chapman & Hall/CRC, Boca Raton, Fl.

Ishwaran, H. et al. (2007). Variable importance in binary regression trees and forests. *Electronic Journal of Statistics* **1,** 519–537.

Ishwaran, H., Gerds, T. A., Kogalur, U. B., Moore, R. D., Gange, S. J., and Lau, B. M. (2014). Random survival forests for competing risks. *Biostatistics* **15,** 757–773.

Ishwaran, H., Kogalur, U. B., Chen, X., and Minn, A. J. (2011). Random survival forests for high-dimensional data. *Statistical analysis and data mining* **4,** 115–132.

Ishwaran, H., Kogalur, U. B., Gorodeski, E. Z., Minn, A. J., and Lauer, M. S. (2010). High-dimensional variable selection for survival data. *Journal of the American Statistical Association* **105,** 205–217.

Martinussen, T., Vansteelandt, S., and Andersen, P. K. (2018). Subtleties in the interpretation of hazard ratios. *arXiv preprint arXiv:1810.09192* .

Neyman, J. (1923). Sur les applications de la theorie des probabilites aux experiences agricoles: Essai des principes (in polish). english translation by dm dabrowska and tp speed (1990). *Statistical Science* **5,** 465–480.

Ozenne, B. M. H., Scheike, T. H., Stærk, L., and Gerds, T. A. (2019). On the estimation of average treatment effects with right-censored time to event outcome and competing risks. *arXiv preprint arXiv:1907.12912* .

Robins, J. M. and Rotnitzky, A. (1992). Recovery of information and adjustment for dependent censoring using surrogate markers. In *AIDS epidemiology*, pages 297–331. Springer.

Rosenblum, M. and van der Laan, M. J. (2011). Simple examples of estimating causal effects using targeted maximum likelihood estimation.

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

Rytgaard, H. C. (2019). Application of generalized random forests for survival analysis. In *European Young Statisticians Meeting*, page 102.

Scharfstein, D. O., Rotnitzky, A., and Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association* **94,** 1096–1120.

Stensrud, M. J., Young, J. G., Didelez, V., Robins, J. M., and Hernán, M. A. (2019). Separable effects for causal inference in the presence of competing risks. *arXiv preprint arXiv:1901.09472* .

Strobl, C., Boulesteix, A., Kneib, T., Augustin, T., and Zeileis, A. (2008). Conditional variable importance for random forests. *BMC bioinformatics* **9,** 307.

Tsiatis, A. (2007). *Semiparametric theory and missing data*. Springer Science & Business Media.

Tuglus, C. and van der Laan, M. J. (2008). Targeted methods for biomarker discovery, the search for a standard.

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

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

Wang, H. and van der Laan, M. J. (2011). Dimension reduction with gene expression data using targeted variable importance measurement. *BMC bioinformatics* **12,** 312.

Young, J. G., Stensrud, M. J., Tchetgen Tchetgen, E. J., and Hernán, M. A. (2018). A causal framework for classical statistical estimands in failure time settings with competing events. *arXiv preprint arXiv:1806.06136* .

Appendix A

We here detail the identifiability assumptions for the effect on net probabilities and the effect on crude probabilities, respectively.

A.1 Identifiability assumptions for the effect on net probabilities, $\theta_{\text{net}}(\boldsymbol{x})$

Identification of $\theta_{\text{net}}(\boldsymbol{x})$ in terms of the observed data distribution depends on three untestable causal assumptions: Consistency, coarsening at random and positivity.

First, the assumption of consistency entails that the counterfactual event time $T^{1,a}$ corresponds to the observed event time for those subjects who were actually uncensored, free of event type $j = 2$ and were exposed to the treatment level $A_k = a$. Particularly, consistency provides the counterfactual variables as follows:

$$T = \min(T^{1,A_k}, T^{2,A_k}), \text{ and that, } T^{1,a} = T^{1,A_k} \text{ on the event that } A_k = a \text{ for } a = 0, 1. \tag{1a}$$

Here T^{2,A_k} is the uncensored counterfactual event time of type $j = 2$ under the observed treatment.

The second assumption of coarsening at random is characterized as follows. The full data we would have liked to observe are $(\boldsymbol{X}, T^{1,0}, T^{1,1})$. These are not fully observed due to censoring, the competing event and the treatment decision A_k, and we observe only the coarsened data $(\boldsymbol{X}, A_k, \tilde{T}, \tilde{\Delta})$ (Gill et al., 1997; van der Laan and Robins, 2003; Tsiatis, 2007). To identify $(\boldsymbol{X}, T^{1,0}, T^{1,1})$ from the data, we need coarsening at random (CAR) (Gill et al., 1997; van der Laan and Robins, 2003, Section 1.2.3), i.e., that the coarsening mechanism only depends on the full data structure $(\boldsymbol{X}, T^{1,0}, T^{1,1})$ through the observed data structure $(\boldsymbol{X}, A_k, \tilde{T}, \tilde{\Delta})$. Coarsening at random is implied by the following conditional independence conditions:

$$\begin{aligned} &T^{1,a} \perp\!\!\!\perp A_k \,|\, \boldsymbol{X}, \\ &T^{1,A_k} \perp\!\!\!\perp (C, T^{2,A_k}) \,|\, A_k, \boldsymbol{X}, \end{aligned} \tag{1b}$$

for $a = 0, 1$, also refer to as "no unmeasured confounding".

The last assumption of positivity requires for the coarsening mechanism that

$$P(\min(C, T^{2,A_k}) \geq t_0 \,|\, A_k, \boldsymbol{X}) \, (\pi_k(\boldsymbol{X}))^{A_k} (1 - \pi_k(\boldsymbol{X}))^{1-A_k} > \eta > 0, \tag{1c}$$

almost surely.

Under Assumptions 1a, 1b and 1c, we can link the distribution of the counterfactual variables to the observed data distribution as follows:

$$\begin{aligned} &P(\tilde{T} \in dt, \tilde{\Delta} = 1, A_k = a, \boldsymbol{X} \in d\boldsymbol{x}) \\ &\quad = P(\tilde{\Delta} = 1 \,|\, T^{1,A_k} = t, A_k = a, \boldsymbol{X} = \boldsymbol{x}) P(T^{1,A_k} \in dt, A_k = a, \boldsymbol{X} \in d\boldsymbol{x}) \\ &\quad = P(\min(C, T^{2,A_k}) \geq t \,|\, T^{1,A_k} = t, A_k = a, \boldsymbol{X} = \boldsymbol{x}) P(T^{1,A_k} \in dt \,|\, A_k = a, \boldsymbol{X} = \boldsymbol{x}) \\ &\qquad\qquad P(A_k = a \,|\, \boldsymbol{X} = \boldsymbol{x}) P(\boldsymbol{X} \in d\boldsymbol{x}) \\ &\quad = P(\min(C, T^{2,A_k}) \geq t \,|\, A_k = a, \boldsymbol{X} = \boldsymbol{x}) P(T^{1,a} \in dt \,|\, \boldsymbol{X} = \boldsymbol{x}) P(A_k = a \,|\, \boldsymbol{X} = \boldsymbol{x}) P(\boldsymbol{X} \in d\boldsymbol{x}). \end{aligned} \tag{10}$$

Particularly, the first line of Assumption 1b together with the Assumption 1a of consistency implies that

$$\begin{aligned} P(T^{1,A_k} \in dt \,|\, A_k = a, \boldsymbol{X} \in d\boldsymbol{x}) &\overset{1b}{=} P(T^{1,a} \in dt \,|\, A_k = a, \boldsymbol{X} = \boldsymbol{x}) \\ &\overset{1a}{=} P(T^{1,a} \in dt \,|\, \boldsymbol{X} = \boldsymbol{x}), \end{aligned}$$

whereas the second line of Assumption 2a yields that

$$P(\min(C, T^{2,A_k}) \geq t \,|\, T = t, \Delta = 1, A_k = a, \boldsymbol{X} = \boldsymbol{x}) = P(\min(C, T^{2,A_k}) \geq t \,|\, A_k = a, \boldsymbol{X} = \boldsymbol{x}).$$

Assumption 1c ensures that the right hand side of (11) is non-zero and well-defined.

A.2 Identifiability assumptions for the effect on crude probabilities, $\theta_{\text{crude}}(\boldsymbol{x})$

The assumptions needed to identify $\theta_{\text{crude}}(\boldsymbol{x})$ are less restrictive than those needed for $\theta_{\text{net}}(\boldsymbol{x})$ and correspond to the standard setting for right-censored survival times. The consistency assumption for $\theta_{\text{crude}}(\boldsymbol{x})$ can be expressed as

$$T = T^a \text{ and } \Delta = \Delta^a \text{ on the event that } A = a, \text{ for } a = 0, 1. \tag{2a}$$

The full data we would have liked to observe are $(\boldsymbol{X}, T^0, T^1, \Delta^0, \Delta^1)$, but we observe only the coarsened data $(\boldsymbol{X}, A_k, \tilde{T}, \tilde{\Delta})$ due to censoring C and treatment decision A_k. The equivalent of Assumption 1b,

$$\begin{aligned} &(T^a, \Delta^a) \perp\!\!\!\perp A_k \,|\, \boldsymbol{X}, \quad \text{for } a = 0, 1, \\ &(T, \Delta) \perp\!\!\!\perp C \,|\, A_k, \boldsymbol{X}, \end{aligned} \tag{2b}$$

yields coarsening at random. We further make the positivity assumption that,

$$P(C \geq t_0 \,|\, A_k = a, \boldsymbol{X})\,(\pi_k(\boldsymbol{X}))^a (1 - \pi_k(\boldsymbol{X}))^{1-a} > \eta > 0, \tag{2c}$$

almost surely, for $a = 0, 1$.

We can now express the observed data distribution as,

$$\begin{aligned} &P(\tilde{T} \in dt, \tilde{\Delta} = 1, A_k = a, \boldsymbol{X} \in d\boldsymbol{x}) \\ &\quad = P(\tilde{\Delta} \geq 1 \,|\, T = t, \Delta = 1, A_k = a, \boldsymbol{X} = \boldsymbol{x}) P(T \in dt, \Delta = 1, A_k = a, \boldsymbol{X} \in d\boldsymbol{x}) \\ &\quad = P(C \geq t \,|\, T = t, \Delta = 1, A_k = a, \boldsymbol{X} = \boldsymbol{x}) P(T \in dt, \Delta = 1 \,|\, A_k = a, \boldsymbol{X} \in d\boldsymbol{x}) \\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad P(A_k = a \,|\, \boldsymbol{X} = \boldsymbol{x}) P(\boldsymbol{X} \in d\boldsymbol{x}) \\ &\quad = P(C \geq t \,|\, A_k = a, \boldsymbol{X} = \boldsymbol{x}) P(T^a \in dt, \Delta^a = 1 \,|\, \boldsymbol{X} = \boldsymbol{x}) P(A_k = a \,|\, \boldsymbol{X} = \boldsymbol{x}) P(\boldsymbol{X} \in d\boldsymbol{x}), \end{aligned} \tag{11}$$

relying on Assumptions 2a, 2b and 2c. Particularly, the first line of Assumption 2b together with the Assumption 2a of consistency implies that

$$\begin{aligned} P(T \in dt, \Delta = 1 \,|\, A_k = a, \boldsymbol{X} \in d\boldsymbol{x}) &\overset{2b}{=} P(T^a \in dt, \Delta^a = 1 \,|\, A_k = a, \boldsymbol{X} = \boldsymbol{x}) \\ &\overset{2a}{=} P(T^a \in dt, \Delta^a = 1 \,|\, \boldsymbol{X} = \boldsymbol{x}), \end{aligned}$$

whereas the second line of Assumption 2a yields that

$$P(C \geq t \,|\, T = t, \Delta = 1, A_k = a, \boldsymbol{X} = \boldsymbol{x}) = P(C \geq t \,|\, A_k = a, \boldsymbol{X} = \boldsymbol{x}).$$

Assumption 2c ensures that the right hand side of (11) is non-zero and well-defined.

Appendix B

B.3 Weighted outcome for net probabilities

Define the weighted outcome:

$$\tilde{Y}' = \frac{\mathbb{1}\{\tilde{T} \leq t_0, \tilde{\Delta} = 1\}}{G'(\tilde{T}- \,|\, A_k, \boldsymbol{X})},$$

with weights given by

$$G'(\tilde{T}- \,|\, A_k, \boldsymbol{X}) = P(\min(T^{2,A_k}, C) \geq t \,|\, A_k, \boldsymbol{X}).$$

For this weighted outcome we have that,

$$\begin{aligned} \theta_{\text{net}}(\boldsymbol{x}) &= P(T^{1,1} \leq t_0 \,|\, \boldsymbol{X} = \boldsymbol{x}) - P(T^{1,0} \leq t_0 \,|\, \boldsymbol{X} = \boldsymbol{x}) \\ &= \mathbb{E}\big[\tilde{Y}' \,\big|\, \boldsymbol{X} = \boldsymbol{x}, A_k = 1\big] - \mathbb{E}\big[\tilde{Y}' \,\big|\, \boldsymbol{X} = \boldsymbol{x}, A_k = 0\big]. \end{aligned} \tag{12}$$

This follows straightforwardly by the identification in, and just after, Equation (10); indeed, we note that

$$\begin{aligned}
\mathbb{E}[\tilde{Y}' \,|\, \boldsymbol{X}, A_k] &= \mathbb{E}\left[\frac{\mathbb{1}\{\tilde{T} \le t_0, \tilde{\Delta} = 1\}}{G'(\tilde{T}- \,|\, \boldsymbol{X}, A_k)} \,\middle|\, \boldsymbol{X}, A_k \right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\frac{\mathbb{1}\{T^{1,A_k} \le t_0\}\mathbb{1}\{\tilde{\Delta} = 1\}}{G'(\tilde{T}- \,|\, \boldsymbol{X}, A_k)} \,\middle|\, T^{1,A_k}, \boldsymbol{X}, A_k \right] \,\middle|\, \boldsymbol{X}, A_k \right] \\
&= \mathbb{E}\big[\mathbb{1}\{T^{1,A_k} \le t_0\} \,|\, \boldsymbol{X}, A_k\big]\, \mathbb{E}\left[\frac{\mathbb{E}[\mathbb{1}\{\tilde{\Delta} = 1\} \,|\, T^{1,A_k}, \boldsymbol{X}, A_k]}{G'(T^{1,A_k}- \,|\, \boldsymbol{X}, A_k)} \right] \\
&= \mathbb{E}\big[\mathbb{1}\{T^{1,A_k} \le t_0\} \,|\, \boldsymbol{X}, A_k\big]\, \mathbb{E}\left[\frac{G'(T^{1,A_k}- \,|\, \boldsymbol{X}, A_k)}{G'(T^{1,A_k}- \,|\, \boldsymbol{X}, A_k)} \right] \\
&= \mathbb{E}[\mathbb{1}\{T^{1,A_k} \le t_0 \,|\, \boldsymbol{X}, A_k],
\end{aligned}$$

and

$$\mathbb{E}[\mathbb{1}\{T^{1,A_k} \le t_0 \,|\, \boldsymbol{X}, A_k = a] = \mathbb{E}[\mathbb{1}\{T^{1,a} \le t_0\} \,|\, \boldsymbol{X}, A_k = a] = \mathbb{E}[\mathbb{1}\{T^{1,a} \le t_0\} \,|\, \boldsymbol{X}],$$

for $a = 0, 1$, which yields (12).

B.4 Weighted outcome for crude probabilities

For the weighted outcome,

$$\tilde{Y} = \frac{\mathbb{1}\{\tilde{T} \le t_0, \tilde{\Delta} = 1\}}{G(\tilde{T}- \,|\, A_k, \boldsymbol{X})},$$

we have that,

$$\begin{aligned}
\theta_{\text{crude}}(\boldsymbol{x}) &= P(T^1 \le t_0, \Delta^1 = 1 \,|\, \boldsymbol{X} = \boldsymbol{x}) - P(T^0 \le t_0, \Delta^0 = 1 \,|\, \boldsymbol{X} = \boldsymbol{x}) \\
&= \mathbb{E}\big[\tilde{Y} \,\big|\, \boldsymbol{X} = \boldsymbol{x}, A_k = 1\big] - \mathbb{E}\big[\tilde{Y} \,\big|\, \boldsymbol{X} = \boldsymbol{x}, A_k = 0\big].
\end{aligned} \tag{13}$$

This follows straightforwardly by the identification in, and just after, Equation (11); indeed, we note that

$$\begin{aligned}
\mathbb{E}[\tilde{Y} \,|\, \boldsymbol{X}, A_k] &= \mathbb{E}\left[\frac{\mathbb{1}\{\tilde{T} \le t_0, \tilde{\Delta} = 1\}}{G(\tilde{T}- \,|\, \boldsymbol{X}, A_k)} \,\middle|\, \boldsymbol{X}, A_k \right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\frac{\mathbb{1}\{T \le t_0, \Delta = 1\}\mathbb{1}\{\tilde{\Delta} \ge 1\}}{G(T- \,|\, \boldsymbol{X}, A_k)} \,\middle|\, T, \Delta, \boldsymbol{X}, A_k \right] \,\middle|\, \boldsymbol{X}, A_k \right] \\
&= \mathbb{E}\left[\mathbb{1}\{T \le t_0, \Delta = 1\}\, \frac{\mathbb{E}\big[\mathbb{1}\{\tilde{\Delta} \ge 1\} \,|\, T, \Delta, \boldsymbol{X}, A_k\big]}{G(T- \,|\, \boldsymbol{X}, A_k)} \,\middle|\, \boldsymbol{X}, A_k \right] \\
&= \mathbb{E}\left[\mathbb{1}\{T \le t_0, \Delta = 1\}\, \frac{G(T- \,|\, \boldsymbol{X}, A_k)}{G(T- \,|\, \boldsymbol{X}, A_k)} \,\middle|\, \boldsymbol{X}, A_k \right] \\
&= \mathbb{E}\big[\mathbb{1}\{T \le t_0, \Delta = 1\} \,|\, \boldsymbol{X}, A_k\big]
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}[\mathbb{1}\{T \le t_0, \Delta = 1\} \,|\, \boldsymbol{X}, A_k = a] &= \mathbb{E}[\mathbb{1}\{T^a \le t_0, \Delta^a = 1\} \,|\, \boldsymbol{X}, A_k = a] \\
&= \mathbb{E}[\mathbb{1}\{T^a \le t_0, \Delta^a = 1\} \,|\, \boldsymbol{X}],
\end{aligned}$$

for $a = 0, 1$, which yields (13).

Appendix C

To explain the general idea of GRFs, we use a generic (uncensored) random variable $Y \in \mathbb{R}$ and a corresponding generic parameter of interest,

$$\theta(\boldsymbol{x}) = \mathbb{E}[Y \mid A_k = 1, \boldsymbol{X} = \boldsymbol{x}] - \mathbb{E}[Y \mid A_k = 0, \boldsymbol{X} = \boldsymbol{x}],$$

representing the treatment effect of A_k on Y conditional on $\boldsymbol{X} = \boldsymbol{x}$. Athey et al. (2019, Section 6) consider a conditional average partial effect estimation problem which they formulate in terms of a structural model. Below we demonstrate the equivalence of their setting with the counterfactual formulation and show that the conditional average treatment effect estimation problem considered here is a special case. In particular, we show that the parameter $\theta(\boldsymbol{x})$ can be identified in terms of

$$\theta(\boldsymbol{x}) = \frac{\operatorname{cov}(A_k, Y \mid \boldsymbol{X} = \boldsymbol{x})}{\operatorname{Var}(A_k \mid \boldsymbol{X} = \boldsymbol{x})}. \tag{14}$$

This means that $\theta(\boldsymbol{x})$ can be estimated by providing estimators for $\operatorname{cov}(A_k, Y \mid \boldsymbol{X} = \boldsymbol{x})$ and $\operatorname{Var}(A_k \mid \boldsymbol{X} = \boldsymbol{x})$, respectively. The forest outputs weights that can be used to define such estimators as follows. First, forest weights are obtained by averaging over the neighborhoods $L_b(\boldsymbol{x})$ defined by the trees, $b = 1, \ldots, B$,

$$\alpha_i(\boldsymbol{x}) = \frac{1}{B} \sum_{b=1}^{B} \alpha_{b,i}(\boldsymbol{x}), \qquad \text{where,} \quad \alpha_{b,i}(\boldsymbol{x}) = \frac{\mathbb{1}\{\boldsymbol{X}_i \in L_b(\boldsymbol{x})\}}{\sum_{k=1}^{n} \mathbb{1}\{\boldsymbol{X}_k \in L_b(\boldsymbol{x})\}}. \tag{15}$$

Then, the forest estimator $\hat{\theta}_\alpha(\boldsymbol{x})$ is given by,

$$\hat{\theta}_\alpha(\boldsymbol{x}) = \left(\sum_{i=1}^{n} \alpha_i(\boldsymbol{x}) \left(A_i - \bar{A}_{k,\alpha}\right)^2 \right)^{-1} \left(\sum_{i=1}^{n} \alpha_i(\boldsymbol{x}) \left(A_i - \bar{A}_{k,\alpha}\right) \left(Y_i - \bar{Y}_\alpha\right) \right). \tag{16}$$

Here, $\bar{A}_{k,\alpha} = \sum_{i=1}^{n} \alpha_i(\boldsymbol{x}) A_i$ and $\bar{Y}_\alpha = \sum_{i=1}^{n} \alpha_i(\boldsymbol{x}) Y_i$ are estimators for the propensity score $\pi_k(\boldsymbol{x}) = \mathbb{E}[A_k \mid \boldsymbol{X} = \boldsymbol{x}]$ and for $\mathbb{E}[Y \mid \boldsymbol{X} = \boldsymbol{x}]$, respectively. Athey et al. (2019, Theorem 5 and Section 6) provide conditions under which $\hat{\theta}_\alpha$ converges in distribution to a normal distribution centered around the true $\theta(\boldsymbol{x})$. They further propose an estimator $\hat{\sigma}_n(\boldsymbol{x})$ for the standard deviation of the asymptotic distribution.

A key part of the generalized random forest algorithm is the splitting rule that targets specifically the estimation of the quantity of interest $\theta(\boldsymbol{x})$. Each split starts with a mother node $M \subset \mathcal{X}$, corresponding to a subset of $\mathcal{X}$, that is to be split into two daughter nodes $D_1 \cup D_2 = M$. For $l = 1, 2$, let $\hat{\theta}_{D_l}$ be the daughter node local estimate of $\theta(\boldsymbol{x})$ given by (16) with $\alpha_i(\boldsymbol{x}) = \mathbb{1}\{\boldsymbol{X}_i \in D_l\}$ that simply gives weight one to all samples falling in the respective daughter node. To derive their approximate criterion for picking good splits, Athey et al. (2019) use a gradient-based approximation of the mother node estimator $\hat{\theta}_M$. In the setting without censoring and competing risks, as we demonstrate below, it can be seen that the "pseudo-outcomes" used in the "labeling step" of the splitting rule correspond to mother node specific estimates of the efficient influence function for the target parameter. Specifically, the split criterion is based on,

$$\rho_i = W_M^{-1}(A_{k,i} - \bar{A}_M)\Big(Y_i - \bar{Y}_M - \big(A_{k,i} - \bar{A}_M\big)\hat{\theta}_M\Big), \tag{17}$$

where,

$$W_M = \frac{1}{\#\{i \,:\, \boldsymbol{X}_i \in M\}} \sum_{\{i \,:\, X_i \in M\}} (A_{k,i} - \bar{A}_M)^2,$$

and $\bar{A}_M, \bar{Y}_M$ are mother node averages. Each split of a mother node M into daughter nodes D_1, D_2 is carried out such as to maximize,

$$\tilde{\mathcal{L}}(D_1, D_2) = \sum_{l=1}^{2} \frac{1}{\#\{i \,:\, \boldsymbol{X}_i \in D_l\}} \left(\sum_{i \in \{i \,:\, X_i \in D_l\}} \rho_i \right)^2,$$

with ρ_i as defined in (17).

C.5 Equivalence between counterfactual formulation and structural model formulation

We demonstrate the equivalence of the setting of Athey et al. (2019, Section 6) with the counterfactual formulation and show that the conditional average treatment effect estimation problem considered in the main paper (Section 4) is a special case hereof.

Accordingly, we here consider observed data $O = (\boldsymbol{X}, A_k, Y)$, $\boldsymbol{X} \in \mathcal{X}$, $A_k \in \{0, 1\}$ and $Y \in \mathbb{R}$ (uncensored). Further, let Y^1 be the counterfactual outcome that would have been observed under $A_k = 1$, and Y^0 be the counterfactual outcome that would have been observed under $A_k = 0$. The consistency assumption states that

$$Y = A_k Y^1 + (1 - A_k) Y^0, \tag{18}$$

and the exogeneity assumption (no unmeasured confounding) that $(Y^1, Y^0) \perp\!\!\!\perp A_k \mid \boldsymbol{X}$. The conditional treatment effect is defined as,

$$\theta(\boldsymbol{x}) = \mathbb{E}[Y^1 - Y^0 \mid \boldsymbol{X} = \boldsymbol{x}] = \mathbb{E}[Y \mid A_k = 1, \boldsymbol{X} = \boldsymbol{x}] - \mathbb{E}[Y \mid A_k = 0, \boldsymbol{X} = \boldsymbol{x}].$$

The second equality follows under the exogeneity assumption together with the consistency assumption.

Assume on the other hand that,

$$Y_i = a_i + b_i A_k + \varepsilon_i, \tag{19}$$

equivalent to (Athey et al., 2019, Section 6) with our $a_i + \varepsilon_i$ collapsed into just ε_i.

We show that (19) imposes no restriction when A_k is binary. Under consistency, we can express Y as,

$$\begin{aligned}
Y &= A_k Y^1 + (1 - A_k) Y^0 \\
&= A_k Y^1 + (1 - A_k) Y^0 + A_k\big(\mathbb{E}[Y^1 \mid \boldsymbol{X}] - \mathbb{E}[Y^0 \mid \boldsymbol{X}]\big) - A_k \mathbb{E}[Y^1 \mid \boldsymbol{X}] - (1 - A_k)\mathbb{E}[Y^0 \mid \boldsymbol{X}] \\
&\qquad\qquad + \mathbb{E}[Y^0 \mid \boldsymbol{X}] \\
&= \mathbb{E}[Y^0 \mid \boldsymbol{X}] + A_k\big(\mathbb{E}[Y^1 \mid \boldsymbol{X}] - \mathbb{E}[Y^0 \mid \boldsymbol{X}]\big) + A_k\big(Y^1 - \mathbb{E}[Y^1 \mid \boldsymbol{X}] + (1 - A_k)\big(Y^0 - \mathbb{E}[Y^0 \mid \boldsymbol{X}]\big).
\end{aligned}$$

So if we let,

$$\begin{aligned}
a_i &:= \mathbb{E}[Y^0 \mid \boldsymbol{X}_i], \\
b_i &:= \mathbb{E}[Y^1 \mid \boldsymbol{X}_i] - \mathbb{E}[Y^0 \mid \boldsymbol{X}_i], \quad \text{and,} \\
\varepsilon_i &:= (1 - A_{k,i})\big(Y^0 - \mathbb{E}[Y^0 \mid \boldsymbol{X}_i]\big) + A_{k,i}\big(Y^1 - \mathbb{E}[Y^1 \mid \boldsymbol{X}_i]\big),
\end{aligned}$$

we are back on the form in (19).

Further note that,

$$\mathbb{E}[\varepsilon_i \mid A_k, \boldsymbol{X}] = (1 - A_{k,i})\big(\mathbb{E}[Y^0 \mid \boldsymbol{X}_i] - \mathbb{E}[Y^0 \mid \boldsymbol{X}_i]\big) + A_{k,i}\big(\mathbb{E}[Y^1 \mid \boldsymbol{X}_i] - \mathbb{E}[Y^1 \mid \boldsymbol{X}_i]\big) = 0,$$

so that,

$$\mathbb{E}[Y_i \mid A_k, \boldsymbol{X}] = \mathbb{E}[Y^0 \mid \boldsymbol{X}_i] + \theta(\boldsymbol{x}) A_k.$$

C.6 Identification of the target parameter

We demonstrate that,

$$\theta(\boldsymbol{x}) = \mathbb{E}[Y \mid A_k = 1, \boldsymbol{X} = \boldsymbol{x}] - \mathbb{E}[Y \mid A_k = 0, \boldsymbol{X} = \boldsymbol{x}] = \frac{\operatorname{cov}(A_k, Y \mid \boldsymbol{X} = \boldsymbol{x})}{\operatorname{Var}(A_k \mid \boldsymbol{X} = \boldsymbol{x})}. \tag{20}$$

This follows since $A_k \in \{0, 1\}$, so that we have:

$$\begin{aligned}
\operatorname{cov}(A_k, Y \mid \boldsymbol{X} = \boldsymbol{x}) &= \mathbb{E}[A_k Y \mid \boldsymbol{X} = \boldsymbol{x}] - \mathbb{E}[A_k \mid \boldsymbol{X} = \boldsymbol{x}]\, \mathbb{E}[Y \mid \boldsymbol{X} = \boldsymbol{x}] \\
&= \mathbb{E}[A_k Y \mid A_k = 1, \boldsymbol{X} = \boldsymbol{x}]\, \pi_k(\boldsymbol{x}) \\
&\qquad + \mathbb{E}[A_k Y \mid A_k = 0, \boldsymbol{X} = \boldsymbol{x}]\, (1 - \pi_k(\boldsymbol{x})) - \pi_k(\boldsymbol{x})\, \mathbb{E}[Y \mid \boldsymbol{X} = \boldsymbol{x}] \\
&= \mathbb{E}[Y \mid A_k = 1, \boldsymbol{X} = \boldsymbol{x}]\, \pi_k(\boldsymbol{x}) - \pi_k(\boldsymbol{x}) \big(\mathbb{E}[Y \mid A_k = 1, \boldsymbol{X} = \boldsymbol{x}]\, \pi_k(\boldsymbol{x}) \\
&\qquad + \mathbb{E}[Y \mid A_k = 0, \boldsymbol{X} = \boldsymbol{x}]\, (1 - \pi_k(\boldsymbol{x})) \big) \\
&= \mathbb{E}[Y \mid A_k = 1, \boldsymbol{X} = \boldsymbol{x}]\, \pi_k(\boldsymbol{x})(1 - \pi_k(\boldsymbol{x})) \\
&\qquad - \mathbb{E}[Y \mid A_k = 0, \boldsymbol{X} = \boldsymbol{x}]\, \pi_k(\boldsymbol{x})(1 - \pi_k(\boldsymbol{x})) \\
&= \big(\mathbb{E}[Y \mid A_k = 1, \boldsymbol{X} = \boldsymbol{x}] - \mathbb{E}[Y \mid A_k = 0, \boldsymbol{X} = \boldsymbol{x}] \big)\, \pi_k(\boldsymbol{x})(1 - \pi_k(\boldsymbol{x})), \\
&= \big(\mathbb{E}[Y \mid A_k = 1, \boldsymbol{X} = \boldsymbol{x}] - \mathbb{E}[Y \mid A_k = 0, \boldsymbol{X} = \boldsymbol{x}] \big)\, \operatorname{Var}(A_k \mid \boldsymbol{X} = \boldsymbol{x}),
\end{aligned}$$

which yields (20).

C.7 Influence function used for splitting

The influence function used to split in the GRF algorithm for estimation of treatment effects (Athey et al., 2019, Section 6) is,

$$\rho_i = W_M^{-1}(A_{k,i} - \bar{A}_M)\Big(Y_i - \bar{Y}_M - \big(A_{k,i} - \bar{A}_M\big)\hat{\theta}_M\Big), \tag{21}$$

where,

$$W_M = \frac{1}{\#\{i \,:\, \boldsymbol{X}_i \in M\}} \sum_{\{i \,:\, X_i \in M\}} (A_{k,i} - \bar{A}_M)^2,$$

and $\bar{A}_M$, $\bar{Y}_M$ are mother node averages. Note that ρ_i in (21) is a mother node specific estimator for,

$$\phi(Y, A_k) = \big(\operatorname{Var}(A_k \mid \boldsymbol{x})\big)^{-1}\big(A_k - \pi_k(\boldsymbol{x})\big)\big(Y - \mathbb{E}[Y \mid \boldsymbol{X} = \boldsymbol{x}] - \big(A_k - \pi_k(\boldsymbol{x})\big)\theta(\boldsymbol{x})\big). \tag{22}$$

We here demonstrate that $\phi(Y, A_k)$ in (22) can also be written,

$$\phi(Y, A_k) = \left(\frac{A_k}{\pi_k(\boldsymbol{x})} - \frac{1 - A_k}{1 - \pi_k(\boldsymbol{x})}\right)\Big(Y - \mathbb{E}[Y \mid A_k, \boldsymbol{X} = \boldsymbol{x}]\Big), \tag{23}$$

which we recognize as the efficient influence function for estimation of the parameter $\theta(\boldsymbol{x}) = \mathbb{E}[Y \mid A_k = 1, \boldsymbol{X} = \boldsymbol{x}] - \mathbb{E}[Y \mid A_k = 0, \boldsymbol{X} = \boldsymbol{x}]$ (Scharfstein et al., 1999; Rosenblum and van der Laan, 2011).

First note that $\operatorname{Var}(A_k \mid \boldsymbol{x}) = (1 - \pi_k(\boldsymbol{x}))\pi_k(\boldsymbol{x})$ since A_k is binary. Next, by iterated expectations, we have that,

$$\mathbb{E}[Y \mid \boldsymbol{X} = \boldsymbol{x}] = \mathbb{E}[Y \mid A_k = 1, \boldsymbol{X} = \boldsymbol{x}]\, \pi_k(\boldsymbol{x}) + \mathbb{E}[Y \mid A_k = 0, \boldsymbol{X} = \boldsymbol{x}]\, (1 - \pi_k(\boldsymbol{x})).$$

Moreover, we can write $(A_k - \pi_k) = A_k\,(1 - \pi_k) + (1 - A_k)\,\pi_k$. Also recall that $\theta_{\text{net}}(\boldsymbol{x}) = \mathbb{E}[Y \mid A_k = 1, \boldsymbol{X} = \boldsymbol{x}] - \mathbb{E}[Y \mid A_k = 0, \boldsymbol{X} = \boldsymbol{x}]$.

Now rewrite,

$$\begin{aligned}
\frac{A_k}{\pi_k(\boldsymbol{x})} - \frac{(1 - A_k)}{1 - \pi_k(\boldsymbol{x})} &= \frac{A_k(1 - \pi_k(\boldsymbol{x}))}{\pi_k(\boldsymbol{x})(1 - \pi_k(\boldsymbol{x}))} - \frac{(1 - A_k)\pi_k(\boldsymbol{x})}{(1 - \pi_k(\boldsymbol{x}))\pi_k(\boldsymbol{x})} \\
&= \big(A_k - \pi_k(\boldsymbol{x})\big)\, \big(\operatorname{Var}(A_k \mid \boldsymbol{x})\big)^{-1},
\end{aligned}$$

and,

$$\begin{aligned}
&\mathbb{E}[Y \mid A_k = 1, \boldsymbol{X} = \boldsymbol{x}] \\
&\quad = \pi_k(\boldsymbol{x})\, \mathbb{E}[Y \mid A_k = 1, \boldsymbol{X} = \boldsymbol{x}] + (1 - \pi_k(\boldsymbol{x}))\, \mathbb{E}[Y \mid A_k = 1, \boldsymbol{X} = \boldsymbol{x}] \\
&\quad = \mathbb{E}[Y \mid \boldsymbol{X} = \boldsymbol{x}] - \mathbb{E}[Y \mid A_k = 0, \boldsymbol{X} = \boldsymbol{x}](1 - \pi_k(\boldsymbol{x})) + (1 - \pi_k(\boldsymbol{x}))\, \mathbb{E}[Y \mid A_k = 1, \boldsymbol{X} = \boldsymbol{x}] \\
&\quad = \mathbb{E}[Y \mid \boldsymbol{X} = \boldsymbol{x}] - (1 - \pi_k(\boldsymbol{x}))\big(\mathbb{E}[Y \mid A_k = 0, \boldsymbol{X} = \boldsymbol{x}] - \mathbb{E}[Y \mid A_k = 1, \boldsymbol{X} = \boldsymbol{x}]\big) \\
&\quad = \mathbb{E}[Y \mid \boldsymbol{X} = \boldsymbol{x}] + (1 - \pi_k(\boldsymbol{x}))\big(\mathbb{E}[Y \mid A_k = 1, \boldsymbol{X} = \boldsymbol{x}] - \mathbb{E}[Y \mid A_k = 0, \boldsymbol{X} = \boldsymbol{x}]\big),
\end{aligned}$$

and likewise,

$$\begin{aligned}
&\mathbb{E}[Y \mid A_k = 0, \boldsymbol{X} = \boldsymbol{x}] \\
&\quad = \pi_k(\boldsymbol{x})\, \mathbb{E}[Y \mid A_k = 0, \boldsymbol{X} = \boldsymbol{x}] + (1 - \pi_k(\boldsymbol{x}))\, \mathbb{E}[Y \mid A_k = 0, \boldsymbol{X} = \boldsymbol{x}] \\
&\quad = \mathbb{E}[Y \mid \boldsymbol{X} = \boldsymbol{x}] + \mathbb{E}[Y \mid A_k = 0, \boldsymbol{X} = \boldsymbol{x}]\, \pi_k(\boldsymbol{x}) - \pi_k(\boldsymbol{x})\, \mathbb{E}[Y \mid A_k = 1, \boldsymbol{X} = \boldsymbol{x}] \\
&\quad = \mathbb{E}[Y \mid \boldsymbol{X} = \boldsymbol{x}] - \pi_k(\boldsymbol{x})\big(\mathbb{E}[Y \mid A_k = 0, \boldsymbol{X} = \boldsymbol{x}] - \mathbb{E}[Y \mid A_k = 1, \boldsymbol{X} = \boldsymbol{x}]\big) \\
&\quad = \mathbb{E}[Y \mid \boldsymbol{X} = \boldsymbol{x}] + (0 - \pi_k(\boldsymbol{x}))\big(\mathbb{E}[Y \mid A_k = 1, \boldsymbol{X} = \boldsymbol{x}] - \mathbb{E}[Y \mid A_k = 0, \boldsymbol{X} = \boldsymbol{x}]\big).
\end{aligned}$$

Collecting the above, we rewrite (23) as,

$$\begin{aligned}
\phi(Y, A_k) &= \left(\frac{A_k}{\pi_k(\boldsymbol{x})} - \frac{1 - A_k}{1 - \pi_k(\boldsymbol{x})}\right)\Big(Y - \mathbb{E}[Y \mid A_k, \boldsymbol{X} = \boldsymbol{x}]\Big) \\
&= \big(A_k - \pi_k(\boldsymbol{x})\big)\, \big(\mathrm{Var}(A_k \mid \boldsymbol{x})\big)^{-1}\Big(Y - \mathbb{E}[Y \mid \boldsymbol{X} = \boldsymbol{x}] - (A_k - \pi_k(\boldsymbol{x}))\, \theta(\boldsymbol{x})\Big),
\end{aligned}$$

which yields (22).

Appendix D

We here collect descriptive statistics for our data analysis.

Figure 5 shows unadjusted Aalen-Johansen estimators (Aalen and Johansen, 1978) for the risk of readmission with depression and risk of death without relapse, respectively.

Table 1 shows the number of subjects in each age group and in each comorbidity group. Table 2 shows the number of subjects exposed to the different drug groups in the exposure window. Table 3 shows the number of relapse with depression within five years, along with number of subjects who die without depression.

To illustrate the effects of covariates for the estimation of our target parameters, we compare our forest estimates of $\bar{\theta}_{\mathrm{crude}}$ to the naive Aalen-Johansen estimates of crude probabilities stratified on each treatment variable (that leaves out all covariate information), i.e., the nonparametric and unadjusted estimator of,

$$P(T \leq t_0, \Delta = 1 \mid A_k = 1) - P(T \leq t_0, \Delta = 1 \mid A_k = 0).$$

The naive Aalen-Johansen estimates along with confidence intervals and the corresponding causal forest estimates of the treatment effect on the crude probabilities, $\bar{\theta}_{\mathrm{crude}}$, are shown in Figure 6. We see that the treatment effect estimates for some drug groups differ quite a lot for the two methods. Considering these differences, we deduce that there is a covariate effect to be taken into account.

References

Aalen, O. O. and Johansen, S. (1978). An empirical transition matrix for non-homogeneous markov chains based on censored observations. *Scandinavian Journal of Statistics* pages 141–150.

Andersen, P. K., Borgan, Ø., Gill, R. D., and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer Series in Statistics. Springer, New York.

	Male (n=28748)	Female (n=49952)	Total (n=78700)
Infections	6429 (22.4)	14183 (28.4)	20612 (26.2)
Neoplasms	3775 (13.1)	9419 (18.9)	13194 (16.8)
Diseases of blood	674 (2.3)	1403 (2.8)	2077 (2.6)
Diseases of the nervous system	6560 (22.8)	11836 (23.7)	18396 (23.4)
Diseases of the circulatory or respiratory system	9446 (32.9)	16408 (32.8)	25854 (32.9)
Nutritional and metabolic diseases	6863 (23.9)	11752 (23.5)	18615 (23.7)
Diseases of the skin and subcutaneous tissue	2040 (7.1)	3752 (7.5)	5792 (7.4)
Diseases of the musculoskeletal system	7856 (27.3)	15212 (30.5)	23068 (29.3)
Diseases of the genitourinary system and pregnancy, childbirth and the puerperium	4002 (13.9)	21003 (42.0)	25005 (31.8)
age in (0,18]	1900 (6.6)	4595 (9.2)	6495 (8.3)
age in (18,25]	3437 (12.0)	7466 (14.9)	10903 (13.9)
age in (25,30]	2213 (7.7)	4347 (8.7)	6560 (8.3)
age in (30,40]	4668 (16.2)	8457 (16.9)	13125 (16.7)
age in (40,50]	5216 (18.1)	7360 (14.7)	12576 (16.0)
age in (50,60]	4349 (15.1)	5321 (10.7)	9670 (12.3)
age in (60,70]	2689 (9.4)	3486 (7.0)	6175 (7.8)
age in (70,80]	2308 (8.0)	4054 (8.1)	6362 (8.1)
age > 80	1968 (6.8)	4866 (9.7)	6834 (8.7)

Table 1: Comorbidities and demographics of the Danish population-based registry study. Shown are counts (%).

Andersen, P. K., Geskus, R. B., de Witte, T., and Putter, H. (2012). Competing risks in epidemiology: possibilities and pitfalls. *International journal of epidemiology* **41,** 861–870.

Andersen, P. K. and Keiding, N. (2012). Interpretability and importance of functionals in competing risks and multistate models. *Statistics in medicine* **31,** 1074–1088.

Athey, S., Tibshirani, J., and Wager, S. (2019). Generalized random forests. *The Annals of Statistics* **47,** 1148–1178.

Bembom, O., Petersen, M. L., Rhee, S., Fessel, W. J., Sinisi, S. E., Shafer, R. W., and van der Laan, M. J. (2009). Biomarker discovery using targeted maximum-likelihood estimation: Application to the treatment of antiretroviral-resistant hiv infection. *Statistics in medicine* **28,** 152–172.

Breiman, L. (2001). Random forests. *Machine learning* **45,** 5–32.

Gill, R. D., van der Laan, M. J., and Robins, J. M. (1997). Coarsening at random: Characterizations, conjectures, counter-examples. In *Proceedings of the First Seattle Symposium in Biostatistics*, pages 255–294. Springer.

Gray, R. J. (1988). A class of k-sample tests for comparing the cumulative incidence of a competing risk. *The Annals of statistics* **16,** 1141–1154.

Hernán, M. A. (2010). The hazards of hazard ratios. *Epidemiology (Cambridge, Mass.)* **21,** 13.

Hernan, M. A. and Robins, J. M. (2020). *Causal Inference*. Chapman & Hall/CRC, Boca Raton, Fl.

Ishwaran, H. et al. (2007). Variable importance in binary regression trees and forests. *Electronic Journal of Statistics* **1,** 519–537.

Ishwaran, H., Gerds, T. A., Kogalur, U. B., Moore, R. D., Gange, S. J., and Lau, B. M. (2014). Random survival forests for competing risks. *Biostatistics* **15,** 757–773.

Ishwaran, H., Kogalur, U. B., Chen, X., and Minn, A. J. (2011). Random survival forests for high-dimensional data. *Statistical analysis and data mining* **4,** 115–132.

	Male (n=28748)	Female (n=49952)	Total (n=78700)
N06	18740 (65.2)	33327 (66.7)	52067 (66.2)
N05	10049 (35.0)	16784 (33.6)	26833 (34.1)
N02	3384 (11.8)	7467 (14.9)	10851 (13.8)
A02	2509 (8.7)	4486 (9.0)	6995 (8.9)
J01	2118 (7.4)	5739 (11.5)	7857 (10.0)
B01	2687 (9.3)	3484 (7.0)	6171 (7.8)
N03	1713 (6.0)	2965 (5.9)	4678 (5.9)
C03	1610 (5.6)	3450 (6.9)	5060 (6.4)
G03	42 (0.1)	7352 (14.7)	7394 (9.4)
R03	1254 (4.4)	2381 (4.8)	3635 (4.6)
C09	2219 (7.7)	3079 (6.2)	5298 (6.7)
M01	1503 (5.2)	3082 (6.2)	4585 (5.8)
C10	1822 (6.3)	2356 (4.7)	4178 (5.3)
A10	1236 (4.3)	1339 (2.7)	2575 (3.3)
C07	1401 (4.9)	2075 (4.2)	3476 (4.4)
S01	866 (3.0)	2149 (4.3)	3015 (3.8)
C08	1170 (4.1)	1897 (3.8)	3067 (3.9)
A12	815 (2.8)	1907 (3.8)	2722 (3.5)
A06	756 (2.6)	1503 (3.0)	2259 (2.9)
C01	695 (2.4)	1082 (2.2)	1777 (2.3)
G04	1107 (3.9)	328 (0.7)	1435 (1.8)
H03	235 (0.8)	1390 (2.8)	1625 (2.1)
D07	633 (2.2)	1234 (2.5)	1867 (2.4)
N07	897 (3.1)	666 (1.3)	1563 (2.0)
B03	485 (1.7)	1076 (2.2)	1561 (2.0)
R05	370 (1.3)	951 (1.9)	1321 (1.7)
R06	442 (1.5)	1143 (2.3)	1585 (2.0)
A03	334 (1.2)	1010 (2.0)	1344 (1.7)
M05	158 (0.5)	1011 (2.0)	1169 (1.5)
H02	374 (1.3)	755 (1.5)	1129 (1.4)
N04	261 (0.9)	431 (0.9)	692 (0.9)
D01	458 (1.6)	708 (1.4)	1166 (1.5)
R01	357 (1.2)	672 (1.3)	1029 (1.3)

Table 2: Number (%) of subjects purchasing treatments in the Danish population-based registry study.

Ishwaran, H., Kogalur, U. B., Gorodeski, E. Z., Minn, A. J., and Lauer, M. S. (2010). High-dimensional variable selection for survival data. *Journal of the American Statistical Association* **105,** 205–217.

Martinussen, T., Vansteelandt, S., and Andersen, P. K. (2018). Subtleties in the interpretation of hazard ratios. *arXiv preprint arXiv:1810.09192* .

Neyman, J. (1923). Sur les applications de la theorie des probabilites aux experiences agricoles: Essai des principes (in polish). english translation by dm dabrowska and tp speed (1990). *Statistical Science* **5,** 465–480.

Ozenne, B. M. H., Scheike, T. H., Stærk, L., and Gerds, T. A. (2019). On the estimation of average treatment effects with right-censored time to event outcome and competing risks. *arXiv preprint arXiv:1907.12912* .

Robins, J. M. and Rotnitzky, A. (1992). Recovery of information and adjustment for dependent censoring using surrogate markers. In *AIDS epidemiology*, pages 297–331. Springer.

Rosenblum, M. and van der Laan, M. J. (2011). Simple examples of estimating causal effects using targeted maximum likelihood estimation.

Δ	event type	number of subjects	percent of total
0	censoring	67794	86.14 %
1	depression relapse	4613	5.861 %
2	competing event	6293	7.996 %

Table 3: Number of relapse events, censoring and competing events after five years.

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

Rytgaard, H. C. (2019). Application of generalized random forests for survival analysis. In *European Young Statisticians Meeting*, page 102.

Scharfstein, D. O., Rotnitzky, A., and Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association* **94,** 1096–1120.

Stensrud, M. J., Young, J. G., Didelez, V., Robins, J. M., and Hernán, M. A. (2019). Separable effects for causal inference in the presence of competing risks. *arXiv preprint arXiv:1901.09472* .

Strobl, C., Boulesteix, A., Kneib, T., Augustin, T., and Zeileis, A. (2008). Conditional variable importance for random forests. *BMC bioinformatics* **9,** 307.

Tsiatis, A. (2007). *Semiparametric theory and missing data*. Springer Science & Business Media.

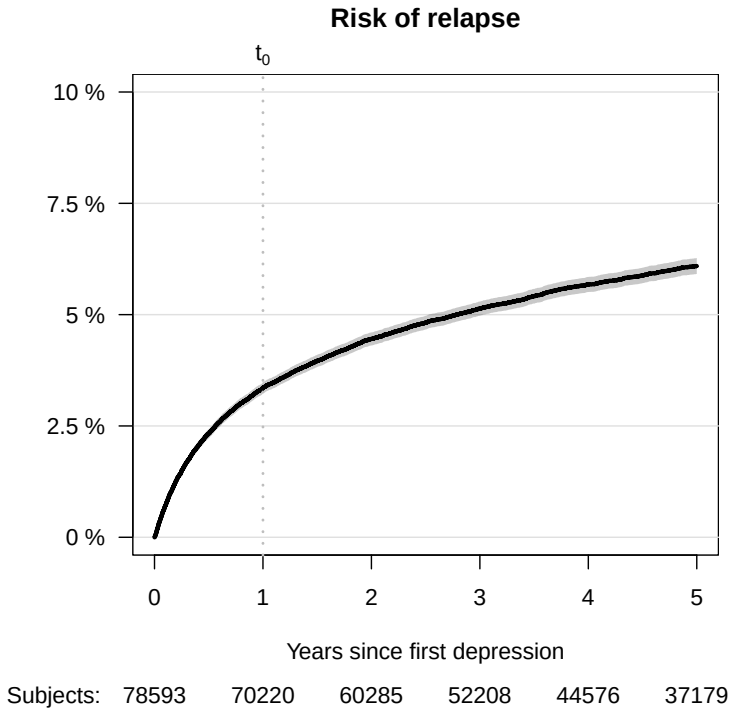
Tuglus, C. and van der Laan, M. J. (2008). Targeted methods for biomarker discovery, the search for a standard.

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

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

Wang, H. and van der Laan, M. J. (2011). Dimension reduction with gene expression data using targeted variable importance measurement. *BMC bioinformatics* **12,** 312.

Young, J. G., Stensrud, M. J., Tchetgen Tchetgen, E. J., and Hernán, M. A. (2018). A causal framework for classical statistical estimands in failure time settings with competing events. *arXiv preprint arXiv:1806.06136* .



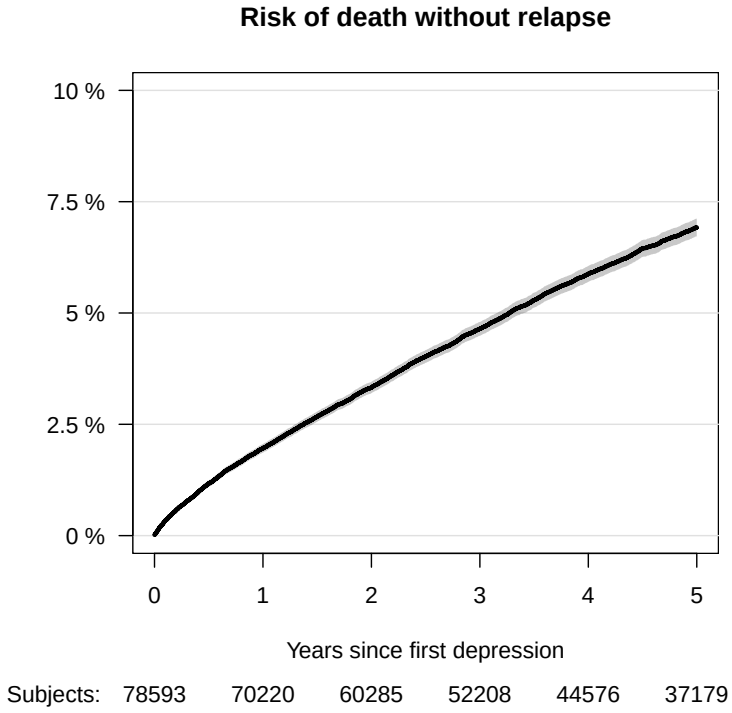


Figure 5: Aalen-Johansen estimators for the risk of readmission with depression (top) and risk of death without relapse (bottom), respectively. We are interested in readmissions within 1 year which is marked on the upper plot. The number of subjects at risk is shown below the plots.

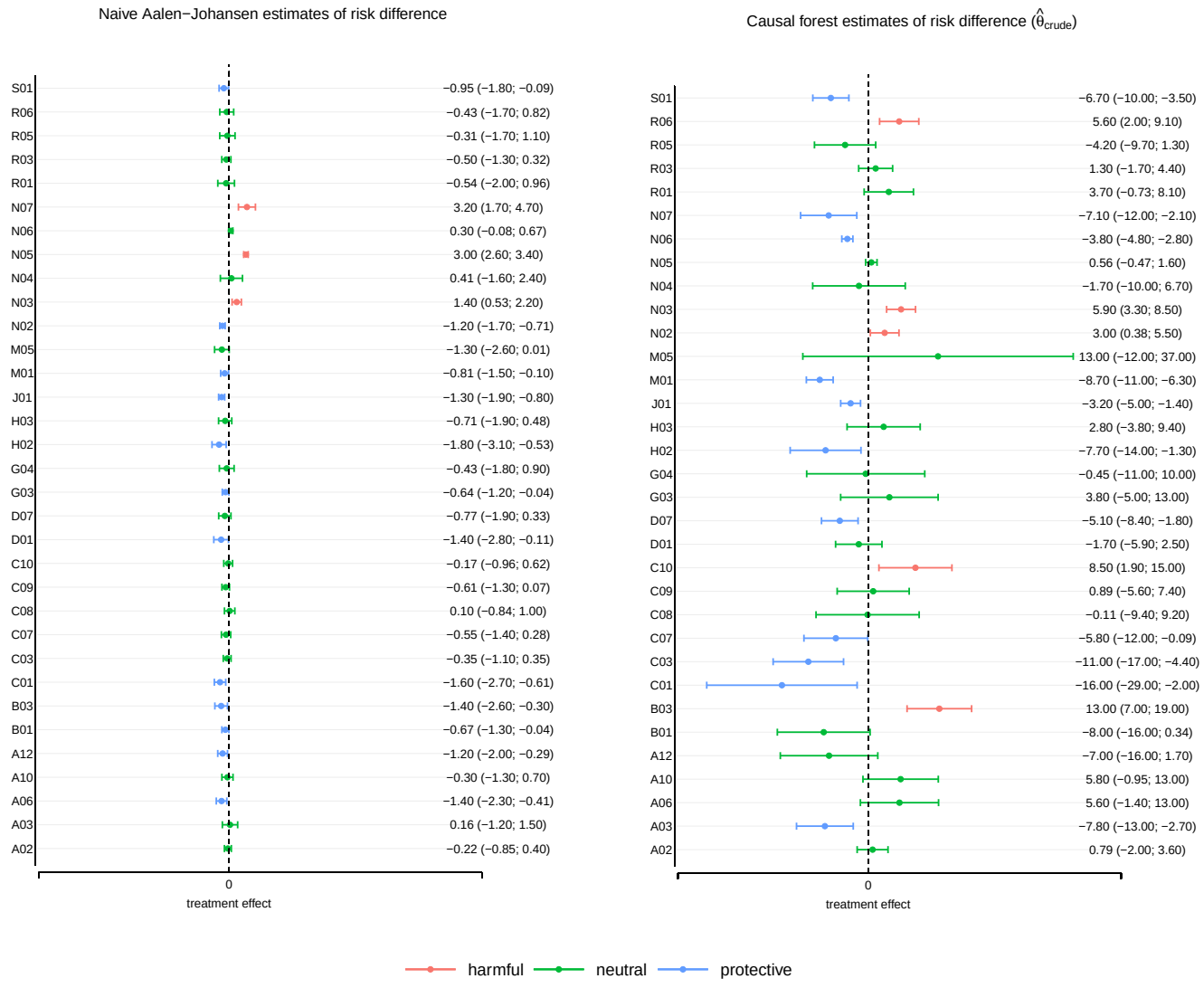


Figure 6: *Left:* Naive Aalen-Johansen estimates of the risk difference, i.e., the difference in crude probabilities stratified on the respective treatment. *Right:* Causal forest estimates of the effect on crude probabilities, $\bar{\theta}_{\text{crude}}$ (using unadjusted weights to construct weighted outcomes). For each ATC group (marked on the y-axis) the plot shows the estimates and the estimated confidence intervals (numbers written on the right). The colors indicate the direction of the effect.