Understanding the Impact of Competing Events on Heterogeneous Treatment Effect Estimation from Time-to-Event Data

Alicia Curth
University of Cambridge

Mihaela van der Schaar
University of Cambridge, The Alan Turing Institute

Abstract

We study the problem of inferring heterogeneous treatment effects (HTEs) from time-to-event data in the presence of *competing* events. Albeit its great practical relevance, this problem has received little attention compared to its counterparts studying HTE estimation without time-to-event data or competing events. We take an outcome modeling approach to estimating HTEs, and consider how and when existing *prediction* models for time-to-event data can be used as plug-in estimators for potential outcomes. We then investigate whether competing events present new challenges for HTE estimation – in addition to the standard confounding problem –, and find that, because there are multiple *definitions* of causal effects in this setting – namely total, direct and separable effects –, competing events *can* act as an additional source of covariate shift depending on the desired treatment effect interpretation and associated estimand. We theoretically analyze and empirically illustrate when and how these challenges play a role when using generic machine learning prediction models for the estimation of HTEs.

1 INTRODUCTION

Competing events are ubiquitous in medical applications where the focus is on the time until occurrence of an adverse event due to a specific cause (Lim et al., 2010; Lambert et al., 2010). Especially when patients have comorbidities, the effect of a treatment on an event of interest can only be assessed when taking into account the presence of risk due to a competing event. For example, when assessing the effectiveness of different cancer treatments for individual cancer patients one may have to consider how to take into account how an individual's risk for cardiovascular events changes due to treatment. This question, however, is far from straightforward: competing events – which act as *mediators* of the treatment on the outcome of interest – give rise to multiple and different definitions of counterfactual risk that could be used depending on the policy or research question of interest, as recently formalized in Young et al. (2020). To see this, note that a treatment which causes a high number of cardiovascular events will automatically result in fewer events due to cancer – which will appear as a protective (total) effect of treatment on the risk of events due to cancer, but may not be the desired interpretation of *what makes a treatment effective* against an adverse outcome of interest. Instead, one could be interested in the direct effect of treatment on outcome (under elimination of competing events) or in the effect of the component in the treatment on outcome that acts only on the primary outcome (Young et al., 2020; Stensrud et al., 2020).

Related work. Possibly because of this conceptual difficulty, heterogeneous treatment effect (HTE) estimation from time-to-event (TTE) data with competing events has received no attention from the machine learning (ML) literature yet. This stands in stark contrast with the ML literature on closely related problems – (a) TTE prediction with competing events (sometimes also referred to as 'competing risks') and (b) HTE estimation with other outcomes – which has flourished in recent years. The literature on the former has adapted a variety of ML methods for risk prediction in the presence of competing events – e.g. using Bayesian nonparametric methods in continuous time (Alaa and van der Schaar, 2017b; Zhang and Zhou, 2018) and neural networks in discrete time (Lee et al., 2018; Wang and Sun, 2022). The literature on the latter has focused on HTE estimation for binary or continuous outcomes, and has either provided model-agnostic strategies to estimate HTEs using *any* ML method (Künzel et al., 2019; Nie and Wager, 2017; Kennedy, 2020; Curth and van der Schaar, 2021a) or adapted specific ML methods to correct for specific challenges of HTE estimation (Shalit et al., 2017; Curth and van der Schaar, 2021b). The largest stream of this literature has focused on analyzing and correcting confounding-induced *covariate shift* (Shalit et al., 2017; Johansson et al., 2018; Hassanpour and Greiner, 2019; Assaad et al., 2021).

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206. Copyright 2023 by the author(s).

Closest to our setting are two recent papers that have investigated covariate shift challenges inherent to HTE estimation for TTE data *without* competing events: Chapfuwa et al. (2021) used generative models for counterfactual TTE analysis in continuous time and Curth et al. (2021a) used neural networks for discrete time analyses. An extended discussion of related work can be found in Appendix A.

Outlook. In this paper, we study heterogeneous treatment effect estimation in the presence of competing events through the causal framework recently established in Young et al. (2020) and used therein to derive estimators for different *average* treatment effects. We begin by considering how the ML toolbox developed for individualized risk prediction in the TTE setting with and without competing events could be used for estimation of different HTEs through a potential outcome modeling approach, or conversely, consider what type of effects are implicitly targeted when different types of ML risk prediction algorithms are used as a basis for treatment decisions. As the ML literature on HTE estimation has focused on the presence of covariate shifts due to confounders, we then move to investigate which forms of covariate shift arise as we target different HTEs. We finally investigate and illustrate their effects empirically across simulation studies. Note that – possibly unconventionally for this literature – our focus here is not on designing or proposing a new method, but rather on *understanding the unique challenges in a new and practically relevant problem* which is why we rely on simple existing methods to allow for clear insights.

Our contributions are thus threefold: Conceptually, we study a new problem in the ML literature on HTE estimation and investigate how one could make use of the strong TTE estimation ML toolbox for solving it. Theoretically, we analyze covariate shift problems that arise therein. Empirically, we obtain insights into how estimation is affected in practice. Overall, we focus on *understanding* the challenges underlying the problem, hoping that the insights that we provide will pave the way for future methodological work on this practically relevant problem.

2 PROBLEM SETUP

We adopt the setup of (Young et al., 2020; Stensrud et al., 2020) in which patients are characterized by pre-treatment characteristics $X{\in}\mathcal{X}$, a binary treatment $A{\in}\{0,1\}$ assigned at baseline, and $\{Y_k\}_{k\in\{1,\dots,K\}}$ and $\{D_k\}_{k\in\{1,\dots,K\}}$, binary indicators for whether the main event and competing event, respectively, have occurred *by time period* $k \leq K$, where K is the maximum time of follow-up. By convention, we assume that D_k precedes Y_k, and that occurrence of either event precludes the other. Further, $\bar{V}_\kappa$ denotes the history $(V_0, \dots, V_\kappa)$ of variable V_k through interval κ. Fig. 1 depicts the assumed underlying causal graph.

This data structure, which is in so-called long format, can equivalently be represented in short format of tuples (X, A, T, E) where T indicates the (discrete) time at which the event occurred (i.e. $T{=}\min k : Y_k{=}1 \vee D_k{=}1$) and $E \in \{Y, D\}$ indicates its type (i.e. $E{=}Y$ if $Y_T{=}1$ else D).

We will generally use long format as it uniquely allows to capture the sequential nature of the problem, but will sometimes use the short format when it simplifies notation. We also assume *no loss to follow-up* due to censoring (i.e. no patient drop-out) for simplicity, but discuss later in Sec. 4.2 how censoring would play a role.

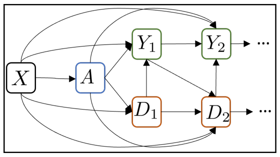


Figure 1: Assumed Causal Graph

Based on the assumed causal structure in Fig. 1, it is easy to see that we can model all risk functions of interest in the competing events literature – e.g. the cause-specific cumulative incidence functions $\mathbb{P}(T \leq k, E = Y|X = x, A = a)$ – by relying on *conditional hazard functions*. These are the hazard (probability) of the main event occurring given that there has been no event yet, i.e.

$$h_Y(k,x,a) = \mathbb{P}(Y_k{=}1|\bar{D}_k{=}\bar{Y}_{k-1}{=}0, X{=}x, A{=}a)$$

and, analogously, the hazard of the competing event occurring given event-free history

$$h_D(k,x,a) = \mathbb{P}(D_k{=}1|\bar{D}_{k-1}{=}\bar{Y}_{k-1}{=}0, X{=}x, A{=}a)$$

These can be used to model e.g. the cause-specific cumulative incidence function, or risk, of an event occurring by time k, as $\mathbb{P}(T \leq k, E{=}Y|X{=}x, A{=}a){=}$

$$\begin{aligned}\mathbb{P}(Y_k{=}1|X{=}x, A{=}a){=}&\textstyle\sum_{l=1}^{k} h_Y(l,x,a)\\ &\times\textstyle\prod_{q=1}^{l-1}(1{-}h_Y(q,x,a))(1{-}h_D(q,x,a))\end{aligned} \tag{1}$$

3 DEFINING AND ESTIMATING HTE GIVEN COMPETING EVENTS

3.1 Preliminaries: Estimating HTEs from TTE data without competing events using prediction models

To introduce the HTE estimation problem, counterfactuals and existing strategies for estimation, we begin with the simpler setting in which there are not competing events. Define the counterfactuals[1] (or potential outcomes) Y_k^a for $a \in \{0,1\}$ and times $k \in \{1,\dots,K\}$ as the event indicator for the scenario in which a patient – possibly countrary to fact – has been *assigned to* treatment $A{=}a$ at baseline, i.e. has been intervened on. Then, we can define heterogeneous

[1] Here, we use the term counterfactual following e.g. Young et al. (2020); Stensrud et al. (2020) exchangeably with the term potential outcome, which is different from Pearl (2009)'s usage of the term; within Pearl's framework we only consider interventional quantities. Throughout, our counterfactuals/potential outcomes Y^a or $Y^{a,\bar{d}}$ correspond to do-operations $do(A = a)$ or $do(A = a, \bar{D}_K = \bar{d})$, respectively.

treatment effects $\tau(x)$ as contrasts of functions of these potential outcomes: For example, if we are interested in *differences in risk* of the event occuring by the end of study K, we have that for a patient with characteristics $X{=}x$, the treatment effect is

$$\tau(x) = \mathbb{P}(Y_K^1 = 1|X = x) - \mathbb{P}(Y_K^0 = 1|X = x) \quad (2)$$

Under the standard ignorability assumptions (Rosenbaum and Rubin, 1983) – which ensure that $\bar{Y}_k^a \perp\!\!\!\perp A|X$ for all k and randomness in treatment assignment for all x – we have that $\mathbb{P}(Y_k^a = 1|X = x) = \mathbb{P}(Y_k = 1|X = x, A = a)$. Therefore a simple and popular way to *estimate* HTEs is through outcome modeling: fitting a standard supervised learning model to predict Y_K as

$$\hat{\mu}^a(x) = \hat{\mathbb{P}}(Y_K = 1|X = x, A = a)$$

e.g. by appending A to X like a standard covariate, or by fitting two separate prediction models, one on each treatment group (these two strategies are often referred to as S- and T-learner, respectively (Künzel et al., 2019)). Then, the HTE can be estimated as $\hat{\tau}(x) = \hat{\mu}^1(x) - \hat{\mu}^0(x)$.

3.2 Total effects: *Using competing event prediction models for estimating HTEs in the presence of competing events allows to estimate total effects*

In a competing events setting, we can similarly define counterfactuals under intervention on treatment Y_k^a (and analogously for the competing event resulting in D_k^a). At first glance, applying the same treatment effect estimation strategy discussed above to the setting with competing events seems appealing: one could use one of the recently proposed predictive ML models for the cause-specific cumulative incidence function (e.g. Lee et al. (2018)'s DeepHit or any other model that allows to estimate eq. (1)) and estimate treatment effect on risk as the difference

$$\hat{\tau}(x){=}\hat{\mathbb{P}}(Y_K{=}1|X{=}x, A{=}1){-}\hat{\mathbb{P}}(Y_K{=}1|X{=}x, A{=}0)$$

which, under similar ignorability assumptions, formalized below, *is* a valid approach.

Assumption 1 (Ignorability w.r.t. treatment) *For each $k \in \{1, \dots, K\}$, we have: (i) Exchangeability w.r.t. A: $Y_k^a, D_k^a \perp\!\!\!\perp A|X$, (ii) Positivity w.r.t. A: $\mathbb{P}(A = a|X = x) > 0$ for $\forall x : \mathbb{P}(X = x) > 0$ and $a \in \{0, 1\}$ and (iii) Consistency w.r.t. A: We observe the counterfactuals associated with the given treatment A, i.e. $Y_k = AY_k^1 + (1 - A)Y_k^0$ and $D_k = AD_k^1 + (1 - A)D_k^0$*

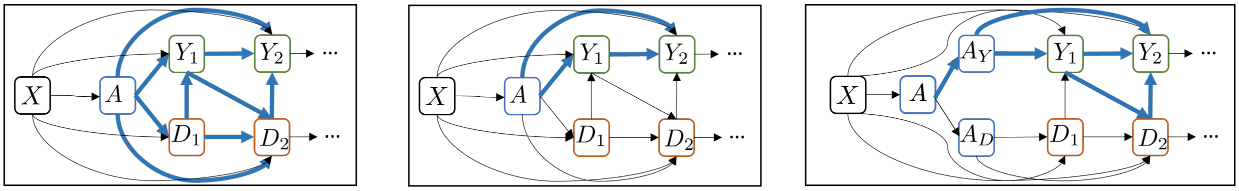
However, using competing event prediction models, which output cause-specific risks, in this way estimates a specific *type* of treatment effect – a total effect (Young et al., 2020) – which may not always be the effect of most natural interest to an investigator. This is because probabilities outputted by cause-specific models depend not only on the occurrence of the primary event, but also on the competing event: Even when the treatment does not affect the primary event at all, it is possible that $\mathbb{P}(Y_K^1 = 1|X = x) \neq \mathbb{P}(Y_K^0 = 1|X = x)$ if treatment affects the competing event because this affects how many individuals are *available* to experience the event (in other words, competing events act as *mediators* in this context (Young et al., 2020)). Formally, this is because, as can be seen in eq. 1, $\mathbb{P}(Y_K^a = 1|X = x)$ depends on the conditional hazard of both the primary and the competing event – thus even if $h_Y(k, x, a)$ is independent of a, the cause-specific risk may not be if $h_D(k, x, a)$ changes with treatment. Fig. 2(A) illustrates the treatment effect's path associated with a total effect. To see why this may be undesirable, consider a cancer treatment that causes all patients in some subgroup in the treatment arm to experience heart failure but has no actual effect on their cancer-related events: in this case, the total effect would show that treatment *reduces* the total risk of events due to cancer in this subgroup, but this is only true because no patients are *available* to experience cancer-related events. Note that this is not a problem with the identification of effects, but rather a feature when using *cause-specific* risks to estimate effects. Thus, when using cause-specific risks and associated total effects for individualized treatment decision making one should be aware that this results in entanglement of different treatment effect pathways.

Remark: Focusing on all-cause survival. A simple way to overcome the competing events problem could be to *combine* outcomes as $Y_k^{all} = Y_k \vee D_k$, i.e. letting go of the distinction of causes, and thus consider the *overall* effect of treatment on all-cause survival. We do not consider this approach further here as this changes the outcome of interest, does not give competing events a special status and can thus be regarded a simple TTE analysis problem to be solved with e.g. the strategies discussed in Chapfuwa et al. (2021); Curth et al. (2021a).

3.3 Direct Effects: *Using TTE models that treat competing events as censoring allows to estimate direct effects*

Another type of counterfactual one could therefore be interested in estimating is $Y_K^{a,\bar{d}=0}$, where an *additional intervention* is made and the competing event is eliminated: That is, an intervention that sets the entire history $\bar{D}_K$ to the deterministic value $\bar{d} = 0$. Considering differences $\tau(x) = \mathbb{P}(Y_K^{1,\bar{d}=0} = 1|X = x) - \mathbb{P}(Y_K^{0,\bar{d}=0} = 1|X = x)$ then corresponds to *direct* effects of treatment onto the event of interest (Young et al., 2020). Fig. 2(B) illustrates the treatment effect's path associated with a direct effect. This effect is identified under Assumption 1 and an additional strong assumption:

Assumption 2 (Ignorability w.r.t. competing event.) *For each $k \in \{1, \dots, K\}$, we have: (i) Exchangeability: w.r.t. D: $Y_k^a \perp\!\!\!\perp D_k^a | X, \bar{Y}_{k-1}{=}\bar{D}_{k-1}{=}0, A{=}a$, (ii) Positivity*



(A) Total Effect (B) Direct Effect (C) Separable Direct Effect

Figure 2: Illustration of the path of total, direct and separable direct effect of treatment A onto Y_2.

w.r.t. D: $\mathbb{P}(D_k = 0 | X{=}x, \bar{Y}_{k-1}{=}\bar{D}_{k-1}{=}0, A{=}a){>}0$ *whenever* $\mathbb{P}(X = x, \bar{Y}_{k-1} = \bar{D}_{k-1} = 0, A = a) > 0$*, (iii) Consistency w.r.t. elimination of D: For an observation with* $A = a$ *and* $\bar{D}_k = 0$ *we observe the corresponding counterfactual, i.e.* $\bar{Y}_k = \bar{Y}_k^{a,\bar{d}=0}$.

Then it is also possible to estimate direct effects from observational data, as $\mathbb{P}(Y_K^{a,\bar{d}=0}{=}1|X = x) = \mathbb{P}(Y_K{=}1|\bar{D}_K{=}0, A{=}a, X{=}x)$, using the formula

$$\mathbb{P}(Y_K^{a,\bar{d}=0}{=}1|X{=}x) = \sum_{l=1}^{K} h_Y(l,x,a) \prod_{q=1}^{l-1}(1{-}h_Y(q,x,a))$$

where, relative to the total risk, the dependence on $h_D(l,x,a)$ has been removed. That is, the direct risk treats competing events like a *source of independent censoring* (Young et al., 2020). This direct effect can therefore also be estimated using off-the-shelf ML methods using outcome modeling. In this case, however, one would no longer need to model the competing event as a separate cause, but instead treat it as a *censoring event* and use single survival models for Y_K only as e.g. Curth et al. (2021a) or simply use only the cause-specific hazard functions $h_Y(l,x,a)$ if they are available from a competing events model.

Note that direct effects not only require stronger identifying assumptions than total effects, but the presence of the intervention $do(\bar{D}_K = 0)$ also introduces a *conceptual challenge*: such a hypothetical intervention may not always be feasible (Stensrud et al., 2020) – e.g. an intervention eliminating *all* risk of cardiovascular events in cancer trial could be somewhat hard to conceptualize.

3.4 Separable Effects: *Estimating path-specific (separable) effects requires access to hazard estimators*

A final alternative effect definition, presenting fewer conceptual challenges than direct effects, was recently proposed in Stensrud et al. (2020): *separable* direct and indirect effects are path-specific effects that assume A conceptually consists of components A_Y and A_D, which affect only the primary event Y_k and the competing event D_k, respectively. While we have that $A{=}A_Y{=}A_D$ in the observed data, we could hypothesize an intervention that sets A_D and A_Y to separate values. This could be plausible if a treatment may consist of different active components with different biological functions that could be deactivated in the future (Stensrud et al., 2020) – allowing to define counterfactuals $Y_k^{a_Y,a_D}$ (and $D_k^{a_Y,a_D}$) which can be used to e.g. investigate separable direct effects on risk $\mathbb{E}[Y_K^{1,a_D} - Y_K^{0,a_D}|X{=}x]$ and separable indirect effects on risk $\mathbb{E}[Y_K^{a_Y,1} {-} Y_K^{a_Y,0}|X{=}x]$. Fig. 2(C) illustrates the treatment effect's path associated with a separable direct effect.

Risk under separable treatments can be estimated from observed data (where treatment was not separated) under Assumption 1 and additional Assumption 3, which, as discussed in Appendix C, has implications similar to Assumption 2 needed for direct effect estimation.

Assumption 3 (Identification w.r.t. separable treatment.) *For each* $k \in \{1, \dots, K\}$*, we have: (i) Dismissible components:*

$$\mathbb{P}(Y_k^{a_Y,a_D=1}{=}1|\bar{Y}_{k-1}^{a_Y,a_D=1}{=}0, \bar{D}_k^{a_Y,a_D=1}{=}0, X{=}x) = \mathbb{P}(Y_k^{a_Y,a_D=0}{=}1|\bar{Y}_{k-1}^{a_Y,a_D=0}{=}0, \bar{D}_k^{a_Y,a_D=0}{=}0, X{=}x)$$

and a similar condition equalizing the conditional hazards of $D_k^{a_y=0,a_D}$ *and* $D_k^{a_y=1,a_D}$ *(see Appendix C), (ii) positivity w.r.t. A (in surviving population):* $\mathbb{P}(A = a|\bar{D}_k = \bar{Y}_k = 0, X = x) > 0$ *whenever* $\mathbb{P}(\bar{D}_k = \bar{Y}_k = 0, X = x) > 0$ *for* $a \in \{0, 1\}$*, (iii) Consistency: For an observation with* $A = a$*, we observe the corresponding counterfactuals, i.e.* $Y_k = Y_k^{a,a}$ *and* $D_k = D_k^{a,a}$.

Then, risk under separable components can be estimated as $\mathbb{P}(Y_K^{a_Y,a_D} = 1|X = x) =$

$$\sum_{l=1}^{K} h_Y(l,x,a_Y) \prod_{q=1}^{l-1}(1{-}h_Y(q,x,a_Y))(1{-}h_D(q,x,a_D))$$

which differs from the identification formula for total risk in that it evaluates the hazard of the competing event under treatment $a_D \neq a_Y$. As no current ML prediction models target such separable treatment paths, most TTE models cannot directly be used by e.g. including treatment as a standard covariate and simply issuing predictions; yet any TTE model from which conditional hazard estimates $\hat{h}_D(k,x,a)$ and $\hat{h}_Y(k,x,a)$ can be extracted can be used to estimate separable risk through computation of the formula above instead of issuing its standard predictions.

Remark: Implications for the use of TTE prediction models for treatment decision making. Some existing work proposing ML TTE competing events prediction methods use the example of making treatment plans as motivation for their method (e.g. Alaa and van der Schaar (2017b)). We therefore wish to reemphasize that, through our discussion in this section, it becomes clear that the use of different types of TTE prediction methods implicitly means considering different types of effects in this context. Assuming that the necessary identifying assumptions hold, when TTE prediction models that explicitly model competing events (e.g. Alaa and van der Schaar (2017b); Lee et al. (2018)) are used to inform treatment plans, this implicitly corresponds to consideration of total effects, while approaches treating competing events as *censoring* events implicitly lead to consideration of direct effects. As discussed above, separable effects are generally not implicitly a by-product of generic TTE prediction methods.

4 UNDERSTANDING COVARIATE SHIFTS DUE TO COMPETING EVENTS

Assuming that all identifying assumptions described above[2] hold, all different types of treatment effects can be estimated from observed data – yet not without further challenges. Because the data available for training follows an observational distribution, while the target quantities are defined with respect to interventions, *covariate shift* arises. Covariate shift arising due to treatment selection on observables (confounding) has been studied in detail in the ML literature on HTE estimation with standard (binary/continuous) outcomes since Johansson et al. (2016); Shalit et al. (2017). More recently, Chapfuwa et al. (2021) tackled only confounding-induced covariate shift in the context of survival outcomes, and Curth et al. (2021a) showed that censoring acts as an additional source of covariate shift in the TTE setting.

In this section, we take a closer look at how covariate shift can arise in the TTE setting with competing events when learning treatment-specific hazard functions from observational data. We show that, because the different effects are defined with respect to different interventions, competing events *can* act as an additional source of covariate shift depending on the chosen treatment effect of interest.

4.1 Learning treatment-specific hazard functions

Here, we focus on learning hazard functions because they can be (a) used to compute all treatment effects defined in the previous section, and (b) easily estimated using off-the-shelf ML methods simply by restricting the training set – and can thus be analyzed like a standard supervised learning problem. The simplest and most flexible problem formulation, which we focus on here, imposes no assumption on how hazards evolve over time (e.g. no proportional hazards assumption) or on how treatment affects outcome, by fitting a separate model for each conditional hazard, giving an estimator for each time-step by treatment group by event type (i.e. $K\times 2\times 2$ estimators in total).

To do so, inspired by the approach described in e.g. Stitelman and van der Laan (2010); Curth et al. (2021a) in the standard TTE setting, we simply separate the observed data $\mathcal{D}_{obs} = \{(X_i, A_i, \bar{Y}_{K,i}, \bar{D}_{K,i}\}_{i=1}^n$ by treatment group, and then, for each time step k, first fit a classification model for outcome D_k and then fit a classification model for outcome Y_k, by using only the patients still at risk of the events: That is, for each time-step k, we estimate $\hat{h}_D(k,x,a) = \hat{\mathbb{P}}(D_k = 1|\bar{D}_{k-1}{=}\bar{Y}_{k-1}{=}0, X{=}x, A{=}a)$ by solving a standard classification problem with input-output tuples $\mathcal{D}^D_{a,k}{=}\{(X_i, D_{k,i})\}_{i\in\mathcal{I}_D(k,a)}$ where $\mathcal{I}_D(k,a){=}\{i \in [n] : \bar{D}_{k-1,i} = \bar{Y}_{k-1,i} = 0, A_i = a\}$ is the competing at-risk set at time-step k with $n^D_{k,a}{=}|\mathcal{I}_D(k,a)|$, and estimate $\hat{h}_Y(k,x,a){=}\hat{\mathbb{P}}(Y_k{=}1|\bar{D}_{k,}{=}\bar{Y}_{k-1,}{=}0, X{=}x, A{=}a)$ by fitting a classification model for input-output tuples $\mathcal{D}^Y_{a,k} = \{(X_i, Y_{k,i})\}_{i\in\mathcal{I}_Y(k,a)}$ using patients remaining in the main at-risk set $\mathcal{I}_Y(k,a){=}\{i \in [n] : \bar{Y}_{k-1,i}{=}\bar{D}_{k,i}{=}0, A_i{=}a\}$, with $n^Y_{k,a}{=}|\mathcal{I}_Y(k,a)|$, at time-step k.

4.2 How does covariate shift arise?

When fitting the conditional hazard for the main event[3] using empirical risk minimization (ERM) for the classification approach described above, the hazard estimator is

$$\hat{h}^Y(k,x,a) \in \arg\min_{h\in\mathcal{H}} \hat{R}^{obs}_{a,k}(h)$$

where $\hat{R}^{obs}_{a,k}(h){=}\sum_{i\in\mathcal{I}_Y(a,k)} \ell(Y_{k,i}, h(X_i))$ is the empirical version of the risk $R^{obs}_{a,k}(h) = \mathbb{E}_{X\sim p^{obs}_{a,k}, Y_k\sim h_Y(k,x,a)}[\ell(Y_k, h(X))]$, $\mathcal{H}$ denotes the hypothesis class under investigation, ℓ is some loss function and $\mathbb{P}^{obs}_{a,k}$ refers to the *observational* at-risk covariate distribution, i.e. $\mathbb{P}^{obs}_{a,k}(X = x) = \mathbb{P}(X{=}x|\bar{Y}_{k-1}{=}\bar{D}_k{=}0, A{=}a)$

$$\begin{aligned}&\propto \mathbb{P}(X{=}x)\mathbb{P}(A{=}a|X = x)(1{-}h_D(k,x,a))\\ &\quad\times \textstyle\prod_{l=1}^{k-1}(1{-}h_D(l,x,a))(1{-}h_Y(l,x,a))\end{aligned} \tag{3}$$

Our target distribution, however, is not the observational distribution: because the treatment effects under consideration are associated with different types of interventions, the target covariate distribution corresponds to an interventional distribution $\mathbb{P}^{int}_{a,k}$, giving rise to the (hypothetical) interventional risk $R^{int}_{a,k}(h) = \mathbb{E}_{X\sim p^{int}_{a,k}, Y_k\sim h_Y(k,x,a)}[\ell(Y_k, h(X))]$. This mismatch between observed and target distribution is known as *covariate shift* and has as a consequence that the learnt

[2] Appendix C contains extended discussions of assumptions.

[3] From here on, we focus on our discussion on the main event, but analogous derivations can be made when effects on the competing event are of interest, where the observational distribution is $\mathbb{P}(X = x|\bar{Y}_{k-1}{=}D_{k-1}{=}0, A{=}a)$.

function is not necessarily optimal as it is possible that $\arg\min_{h\in\mathcal{H}} R^{int}_{a,k}(h) \neq \arg\min_{h\in\mathcal{H}} R^{obs}_{a,k}(h)$. It is well-known that such a mismatch can be addressed by relying on so-called *importance weights* $w^*(x) \propto \frac{\mathbb{P}^{int}_{a,k}(X=x)}{\mathbb{P}^{obs}_{a,k}(X=x)}$ to give $\hat{R}^{w,obs}_{a,k}(h){=}\sum_{i\in\mathcal{I}_Y(a,k)} w^*(X_i)\ell(Y_{k,i}, h(X_i))$ which is unbiased for the interventional risk. Below, we discuss the type of covariate shift and associated importance weights arising when conceptualizing the interventions for the three effects under investigation:

- **Total effect:** The total effect requires only an intervention on treatment $do(A=a)$, thus, as can be read off from the causal graph in Fig. 1, the interventional at-risk distribution $\mathbb{P}^{do(A=a)}_{a,k}(X=x)$ is proportional to $\mathbb{P}(X{=}x)(1-h_D(k,x,a))\prod_{l=1}^{k-1}(1{-}h_D(l,x,a))(1{-}h_Y(l,x,a))$, which differs from the observational distribution only in the treatment assignment factor $\mathbb{P}(A{=}a|X{=}x)$. That is, if treatment was assigned completely at random as in a randomized trial, the two distributions $\mathbb{P}^{obs}_{a,k}$ and $\mathbb{P}^{int}_{a,k}$ would be the same. The covariate shift arising when estimating total effects is thus the same shift arising in the standard treatment effect setting considered in e.g. Shalit et al. (2017), leading to time-independent importance weights $w^{*,do(A=a)}_{a,k}(x) \propto P(A{=}a|X{=}x)^{-1}$.

- **Direct effect:** The direct effect requires intervention $do(A{=}a, \bar{D}_K{=}0)$, which, in addition to treatment assignment bias, also removes all competing events. Therefore, as can be read off from Fig. 1, the interventional at-risk distribution $\mathbb{P}^{do(A=a,\bar{D}_K=0)}_{a,k}(X = x) \propto \mathbb{P}(X{=}x)\prod_{l=1}^{k-1}(1{-}h_Y(l,x,a))$. This differs from the observational distribution in both the absence of the treatment assignment factor *and* removes all effects of competing event on the population composition through removal of the factor $\prod_{l=1}^{k}(1{-}h_D(l,x,a))$; the latter results in covariate shift only if the risk of the competing event is dependent on covariates. As discussed in Section 3.3, the competing event is effectively treated as a censoring event here and the shift is thus equivalent to the censoring-induced shift discussed in Curth et al. (2021a). The importance weights needed to correct for this shift would thus be $w^{*,do(A=a,\bar{D}_K=0)}_{a,k}(x) \propto \Big[P(A{=}a|X{=}x)\prod_{l=1}^{k}(1{-}h_D(l,x,a))\Big]^{-1}$.

- **Separable effects:** Finally, separable effects require separate interventions on both treatment components $do(A_Y{=}a_Y, A_D{=}a_D)$, thus the interventional distribution $\mathbb{P}^{do(A_Y=a_Y,A_D=a_D)}_{a,k}(X = x)$, which is identified due to the dismissible component condition (see Appendix C), is proportional to

$$\mathbb{P}(X{=}x)(1{-}h_D(k,x,a_D))$$
$$\times \prod_{l=1}^{k-1}(1{-}h_D(l,x,a_D))(1{-}h_Y(l,x,a_Y))$$

This differs from the observational distribution in the treatment assignment factor and in that it has $\prod_{l=1}^{k}(1{-}h_D(l,x,a_D))$ instead of $\prod_{l=1}^{k}(1{-}h_D(l,x,a_Y))$. Note that the latter results in covariate shift in the at-risk distribution only if *the effect of treatment on the competing event* is dependent on covariates. The importance weights needed to correct for this shift would thus be $w^{*,do(A_Y=a_Y,A_D=a_D)}_{a,k}(x) \propto \Big[P(A = a|X = x)\frac{\prod_{l=1}^{k} h_D(l,x,a_Y)}{\prod_{l=1}^{k} h_D(l,x,a_D)}\Big]^{-1}$.

Remark: How would the addition of censoring play a role? As we noted above, direct effect estimation essentially corresponds to TTE estimation with ignorable censoring – as one would usually ‘switch off’ censoring for a treatment effect analysis (Young et al., 2020). Thus, if we were to add an additional event C_k to our setting to allow censoring (patient drop-out), this would (a) require additional identifying assumptions similar to assumption 2, e.g. $Y^a_k \perp\!\!\!\perp C^a_k|X$, and (b) lead to the same covariate shift issues as a competing event in direct effect estimation setting – any observational distribution would thus gain an additional factor $\prod_{l=1}^{k}\mathbb{P}(C_l = 0|\bar{D}_{l-1}=\bar{Y}_{l-1}=\bar{C}_{l-1}=0, X=x, A=a)$, which should then be (inversely) multiplied with any importance weight.

4.3 How does this covariate shift affect estimation of hazards from observational data?

To analyze how the learning of hazard functions will be impacted by the multitude of sources of covariate shift we outlined above, we can apply well-known results from the literature on domain adaptation and importance weighting to our problem. Below, we adapt the bound of Cortes et al. (2010) to our setting; refer to Appendix C for proofs.

Proposition 1 *Given timestep k and treatment a, for a loss function $\ell_h \in [0,1]$ of any hypothesis $h \in \mathcal{H}$, such that $d = Pdim(\{\ell_h : h \in \mathcal{H}\})$ (where Pdim is the pseudo-dimension) and $\ell_h \in \mathcal{L}$, where $\mathcal{L}$ is a space of pointwise loss functions, and a weighting function $w(x)$ with $\mathbb{E}[w(X)] = 1$, with probability $1-\delta$ over at-risk sample $\mathcal{D}^Y_{a,k}$ with empirical distribution $\hat{p}^{obs}_{a,k}$, we have*

$$R^{int}_{a,k}(h) - \hat{R}^{w,obs}_{a,k}(h) \leq \Big|\mathbb{E}_{p^{obs}_{a,k}}[(w^*_{a,k}(x) - w(x))\ell_h(x)]\Big|$$
$$+\frac{\max(\sqrt{\mathbb{E}_{p^{obs}_{a,k}}[w^2(x)l^2_h(x)]}, \sqrt{\mathbb{E}_{\hat{p}^{obs}_{a,k}}[w^2(x)l^2_h(x)]})}{{n^Y_{a,k}}^{3/8}}\mathcal{C}^{\mathcal{H}}_{n^Y_{a,k}}$$

with $\mathcal{C}^{\mathcal{H}}_{n^Y_{a,k}} = 2^{5/4}(d\log\frac{e2n^Y_{a,k}}{d} + \log\frac{4}{\delta})^{3/8}$ due to Cortes et al. (2010) (Theorem 4).

The proposition above, broken down further in two corollaries below by incorporating ideas from Johansson et al. (2018) and Maia Polo and Vicente (2022), tells us multiple things about the difficulty of the hazard estimation problem and the consequences of the arising covariate shifts. A first observation from the general statement above is that, unsurprisingly, the at-risk sample size $n^Y_{a,k}$ determines the speed of learning. This sample size naturally decreases in k through the occurance of main events even in the absence of competing events, but decreases further as a function of the frequency of competing events, meaning that $n^{Y,obs}_{a,k} \leq n^{Y,int}_{a,k}$ for hypothetical datasets of size n created using the interventions corresponding to direct (or separable) effects, where the inequality is strict if competing events exist (or if treatment influences the rate of competing events occurring).

Corollary 1 (Perfect importance weights) *For $w(x){=}w^{*,int}_{a,k}(x)$, we have*

$$R^{int}_{a,k}(h) - \hat{R}^{w^*,obs}_{a,k}(h) \leq \frac{1}{\sqrt{ESS^*_{rel}(p^{int}_{a,k}, p^{obs}_{a,k}) n^Y_{a,k}}^{3/8}} \mathcal{C}^{\mathcal{H}}_{n^Y_{a,k}}$$

where $ESS^*_{rel} = \exp_2(D_2(p^{int}_{a,k}||p^{obs}_{a,k}))$ is the expected relative effective sample size, with $D_2(p||q) = log_2 \mathbb{E}_{x\sim p}\left[\frac{p(x)}{q(x)}\right]$ the Rényi divergence of order 2.

In Corollary 1, we see that for the special case of perfect importance weights the first term of the RHS in Proposition 1 is zero and the speed of learning is slowed down if the relative effective sample size $ESS^*_{rel} < 1$ – i.e. whenever interventional and observational covariate distributions differ. To gain further intuition, note that Maia Polo and Vicente (2022) show that for self-normalized importance weights $\bar{w}^*_i$, we can approximate ESS^*_{rel} from data as $\frac{1}{n \sum_{i \in \mathcal{I}} \bar{w}^{*,2}_i} \rightarrow ESS^*_{rel}$ a.s. as $|\mathcal{I}| \rightarrow \infty$, which is lowest when all weights are equal. Time-dependent importance weights $\frac{p^{int}_{a,k}}{p^{obs}_{a,k}}$ may generally be more variable as k increases, e.g. if $h_D(l, x, a)$ is constant in l, we have $(\prod^k_{l=1} h_D(l, x, a_D))^{-1} = (h_D(l, x, a_D))^{-k}$, meaning that the covariate shift problem could be exacerbated over time for the direct and separable effects. However, as the density ratio is integrated w.r.t. $p^{int}_{a,k}$ in the Rényi divergence, this effect can subside if the overall at-risk probability is small, as we also demonstrate empirically below.

Corollary 2 (Standard supervised learning (unweighted)) *For $w(x) = 1$, we have*

$$R^{int}_{a,k}(h) - \hat{R}^{obs}_{a,k}(h) \leq C_{\mathcal{L}} IPM_{\mathcal{L}}(p^{int}_{a,k}, p^{obs}_{a,k}) + {n^Y_{a,k}}^{-3/8} \mathcal{C}^{\mathcal{H}}_{n_{a,k}}$$

where IPM$_{\mathcal{G}}(p, q) = \sup_{g \in \mathcal{G}} \left| \int g(x)(p(x) - q(x))dx \right|$ is an integral probability metric and $C_{\mathcal{L}} > 0$ is s.t. $\frac{\ell}{C_{\mathcal{L}}} \in \mathcal{L}$.

Finally, we consider the special case of *no* weighting in Corollary 2. Here, $ESS^*_{rel} = 1$ due to constant weights and the first term of the RHS of 1 is bounded by an IPM-term – which does *not* decrease as the sample size grows[4], reflecting that standard ERM on the observational data may never recover the best interventional solution. As a consequence, $R^{int}_{a,k}(h) - \hat{R}^{obs}_{a,k}(h)$ may never vanish – depending also on how rich the underlying hypothesis class is. This is well-known to be a problem for misspecified parametric models[5] (Sugiyama et al., 2007), while, when using rich hypotheses classes through flexible nonparametric or deep methods, one generally does not have to trade off model performance in different regions of the covariate space, meaning that, given sufficient data, importance weighting would not be expected to make a difference (Byrd and Lipton, 2019).

5 EXPERIMENTS

Finally, we empirically investigate whether, when and how the different shifts play a role when learning hazard functions with the purpose of estimating the different HTEs. As is common practice in the HTE literature (Curth et al., 2021b), we have to rely on *simulated* data because counterfactuals are not available in real data, meaning that real datasets provide *no ground truth* for evaluating methods. While the standard HTE estimation problem in absence of competing events is only missing a single counterfactual (with respect to interventions on treatment assignment), the problem is exacerbated in our setting where additional (unobservable) interventions on competing events would be required to create ground truth targets for evaluation. In addition to the fully synthetic and highly stylized experiments considered below, we present additional results from a semi-synthetic setup using the real Twins dataset Louizos et al. (2017) in Appendix E, leading to similar insights as the results presented below.

An illustrative DGP. Because there are many different forces at play which we wish to disentangle, we focus on a simple setup here that allows us to highlight important problem features. We assume that individuals are characterized by $x{=}(x_1, x_2)$, two binary risk factors $X_1 \sim \mathcal{B}(0.5)$ and $X_2 \sim \mathcal{B}(0.5{-}\rho(1{-}2X_1))$ that may be correlated; unless indicated otherwise we set $\rho{=}0.35$. We assume a very simple hazard for both outcomes $E \in \{Y, D\}$:

$$h_E(k, x, a) = \begin{cases} p^E_{low} + ap^E_{low,\tau} \text{ if } x_{S_E} = 0 \\ p^E_{high} + ap^E_{high,\tau} \text{ if } x_{S_E} = 1 \end{cases} \quad (4)$$

with $0 < p^E_{\cdot} {+} p^E_{\cdot,\tau} \leq 1$ for both settings. This model is constant over time and depends only on the covariate x_{S_E}, where S_E is the index of the support covariate for

[4] A large proportion of the ML HTE literature has therefore focused on learning representations that minimize an empirical estimate of the IPM-term Shalit et al. (2017). This is easily possible in the standard setting as $\mathbb{P}^{int}$ is observed in the marginal covariate distribution $\mathbb{P}(X{=}x)$ and can hence be used to approximate the IPM term – however, in our setting, the interventional distribution is an unobserved at-risk distribution that differs per time-step (i.e. $\mathbb{P}^{int}_{a,k}$ is not the marginal $\mathbb{P}(X = x)$ except for at $k{=}1$), giving no straightforward analogue to this approach.

[5] For correctly specified parametric models and likelihood loss, we always have $\arg\min_{h \in \mathcal{H}} R^{obs}_{a,k}(h) = \arg\min_{h \in \mathcal{H}} R^{int}_{a,k}(h)$.

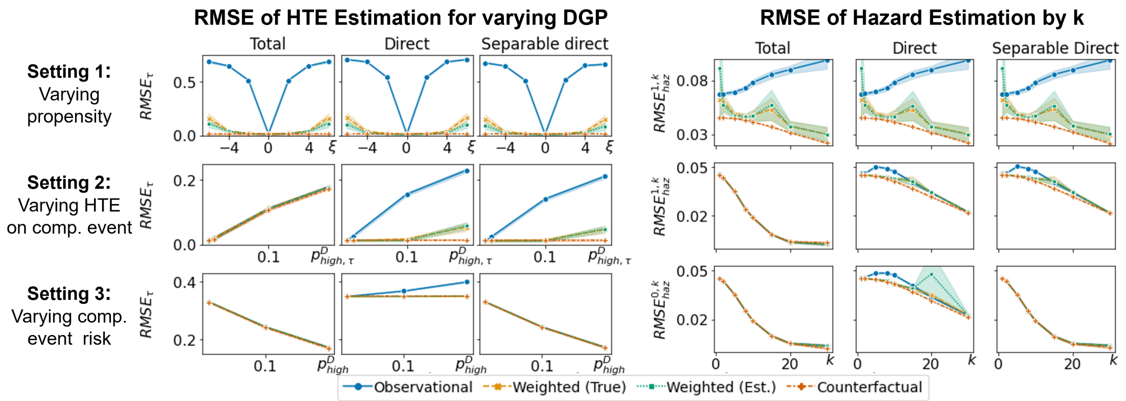


Figure 3: Estimation performance in $RMSE_\tau$ as parameters of the DGP vary (left) and $RMSE^{a,k}_{haz}$ over time k (right), for the 3 effects (columns) across 3 settings (rows). For $RMSE^{a,k}_{haz}$, each DGP's varying parameter is fixed its highest value.

the event model. We let $S_Y = S_D = 1$, and, unless otherwise indicated, set p^Y_{low}=p^D_{low}=p^D_{high} = 0.01, and create a high primary outcome risk group with p^Y_{high}=0.1 and assume no treatment effect $p^E_{high,\tau}$=$p^E_{low,\tau}$=0 for $E \in \{Y, D\}$. We assign treatment based on the propensity score $\pi(x)$=expit($\xi(x_{S_A}-0.5)$) where both $\xi \in [-6, 6]$ and whether S_A overlaps S_E determines the strength of confounding. We generate samples of size n=5000 for K=30 time-steps, and elaborate further on the experiments and data generating processes (DGP) in Appendix D[6]. Using this DGP, we consider three main settings:

1. **Setting 1:** There is confounding as $S_A = 1$ when $|\xi| > 0$, but treatment has no effect on either event. Varying ξ should lead to different levels of covariate shift due to confounding for all effects.
2. **Setting 2:** There is no confounding ($\xi = 0$), treatment has no effect on the main event but affects the competing event ($p^D_{low,\tau} = .01$). Varying $p^D_{high,\tau}$, the heterogeneous effect of treatment on competing events in the high-risk group, should lead to a covariate shift in the at-risk group when interventions associated with direct and separable effect are considered.
3. **Setting 3:** There is no confounding ($\xi = 0$), treatment has a heterogeneous effect on the main event (it equalizes main event risk between both groups, $p^Y_{high,\tau} = -.09$) but has no effect on the competing event. When the high risk group is also at higher (baseline) risk of the competing event (as p^D_{high} varies) this may mask the protective effect of treatment on the main event, and should lead to a covariate shift in the at-risk group for the intervention associated with the direct effect only.

Estimators. We focus on a setup where $\hat{h}_Y(k, x, a)$ is *misspecified*, illustrating the effects of covariate shift on the main outcome model. We assume that the competing event model can be correctly specified, which may be the case in reality if D_k is a well-studied comorbidity. Due to the simple DGP, both models could easily be estimated with an (unrestricted) logistic regression (LR) per time step using the classification framework discussed in the previous section. We use unrestricted (and hence correctly specified) LRs for $h_D(k, x, a)$, but induce misspecfication in $h_Y(k, x, a)$ by fitting *constant* $\hat{h}^Y_{a,k}$ for each $k \leq K$ (this is equivalent to fitting a simple Kaplan-Meier estimator (Kaplan and Meier, 1958) by treatment arm). We also demonstrate that similar conclusions apply when using a LR with L2-penalty that is set too aggressively. To highlight how different covariate shifts affect learning of the different effects, we compare the estimates obtained by *Observational* ERM to importance-weighted ERM with *Estimated* weights $\hat{w}^{*,int}_{a,k}$ and to two oracle solutions: weighted ERM with *true* importance weights $w^{*,int}_{a,k}$ and unweighted ERM on a *counterfactual* sample of size n from the (usually inaccessible) interventional distribution.

Evaluation Metrics. Using independent test sets of size n_{te}=10^4, we report the root-mean-squared-error $RMSE_\tau$=$RMSE(\tau(x))$ of estimating the three different types of risk differences of Section 3 (capturing a total, direct and separable direct HTE), which corresponds to an adaptation of Hill (2011)'s popular Precision in Estimating Heterogeneous Effects (PEHE) metric to our setting. To link back to our theoretical analysis, we also report the RMSE of estimating the hazard function, $RMSE^{a,k}_{haz}$=$\sqrt{\mathbb{E}_{X\sim\mathbb{P}^{int}_{a,k}}[(h^Y(l, X, a)-\hat{h}^Y(l, X, a))^2]}$ where $P^{int}_{a,k}$ is the interventional at-risk distribution corresponding to the effect of interest. We report mean and standard error across 10 replications of each experiment.

5.1 Empirical insights

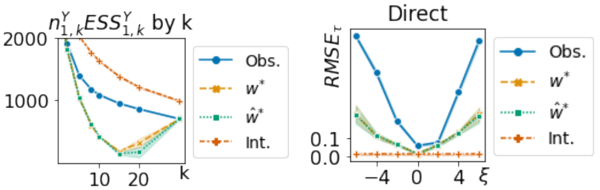
• ***Confounding-induced covariate shifts indeed impact estimation of all effects.*** In Fig. 3 we present results for all 3 settings, highlighting that some effect estimates are indeed impacted by *more* covariate shifts than others. In first

[6] Code to replicate all experiments can be found at `https://github.com/AliciaCurth/CompCATE` or `https://github.com/vanderschaarlab/CompCATE`.

setting, we observe that *all effect estimates* are impacted by increasing confounding strength $|\xi|$: standard ERM performs poorly while importance weighting performs almost identically to the counterfactual solution.

• ***Shifts induced by competing events indeed do not affect estimation of all effects.*** Next, we consider settings 2 and 3 in which treatment assignment is random, but covariate shift can arise due to a differential effect of treatment on the competing event (Setting 2), or due to the high risk group also being at higher risk of the competing event which may mask a protective treatment effect on the main event (Setting 3). In Fig. 3 we observe that, as expected, bias due to covariate shift in setting 2 arises only for separable and direct effect, both of which require elimination of the differential effect of treatment on the competing event – while the total effect does not. Setting 3 exhibits a covariate shift that induces bias only for the direct effect as expected, as treatment has no separable indirect effect here. We also find that, while the error in estimation of $h^Y_{a,k}$ appears small at each time step, it becomes substantial as all K separate hazard functions are cumulated and used to estimate the difference in risk. Note that the absolute error in estimation of the hazard appears to *decrease* over time in most settings – this may appear counterintuitive at first glance as $n^Y_{a,k}$ decreases over time, but is expected in our setting where the interventional at-risk population becomes more homogeneous over time as only low risk individuals are expected to survive, making the constant model approximately correct for later time steps.

• ***Effective sample size behaves and contributes as expected.*** In settings 2 & 3, when considering $RMSE^{a,k}_{haz}$ across k, we observe a tradeoff as expected from our theoretical analysis: as k increases, covariate shift becomes more extreme, initially widening the performance gap between weighted and unweighted ERM solution. As k grows larger, the increased variance (low ESS) of the weighted solution then starts to hinder its performance. Finally, for very large k all solutions converge as the target distribution becomes more homogeneous. This tradeoff is indeed also reflected in the (absolute) effective sample sizes as measured by $(\sum_{i \in \mathcal{I}^Y_{a,k}} \bar{w}^{*,2}_i)^{-1}$ in Fig. 4(a).



(a) ESS in setting 2.

(b) $RMSE_\tau$ in setting 4.

Figure 4: Additional results. Left: effective sample size (ESS) for direct effect estimation in setting 2 ($p^D_{high,\tau}$=.2). Right: $RMSE_\tau$ of direct effect estimation for varying confounding ξ in presence of side effects (additional setting 4)

• ***Multiple shifts can offset or exacerbate each other.*** Finally, we combine settings 1 and 2 (setting 4: $p^D_{high,\tau} = 0.1$, ξ varies); thus the high risk group experiences adverse reactions to treatment in the competing event, and setting $\xi > 0$ ($\xi < 0$) corresponds to assigning more (less) high risk individuals to treatment. In Fig. 4(b), we observe that having assigned more high risk individuals to treatment can offset the shift induced by competing events for (separable) direct effects (and, conversely, the more sensible practice of assigning less high risk individuals to treatment would exacerbate the covariate shift for (separable) direct effects).

• ***When do such covariate shifts truly matter?*** In Fig. 5 we finally investigate *when* shifts matter for ERM. We revisit setting 1, and, in panels A&B, highlight that when x_{S_A} is not a true confounder (does not affect Y_k), the resulting covariate shift biases ERM only when x_{S_Y} also shifts due to correlation with x_{S_A} (the same holds true for other shifts when $S_Y \neq S_D$). In panels C&D, we confirm the impact of misspecification by using a LR as outcome model: an unrestricted LR (C) can fit the DGP well and covariate shift thus has little effect, while the introduction of excessive regularization (D) leads to the need to prioritize regions of $\mathcal{X}$ (which observational ERM does incorrectly).

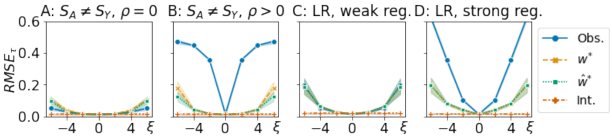


Figure 5: Variations on setting 1 with constant estimators (A & B) and logistic regressions (C & D)

6 CONCLUSION

We studied the challenges inherent to HTE estimation in the presence of competing events, and found that inclusion of competing events not only leads to multiple definitions of effects but also to multiple sources of covariate shifts when estimating them. Theoretically and empirically, we highlighted *that*, *when* and *how* different shifts bias estimation of different effects. Having gained understanding of its challenges, an interesting next step, further discussed in Appendix B, would be to consider how to adapt ideas from the ML literature on HTE & TTE estimation – such as the deep treatment-specific hazard estimator of Curth et al. (2021a) – to construct more sophisticated solutions for our problem setting.

Finally, note that we do *not* wish to argue here that one measure of effect is superior to others – instead, we only aim to raise awareness that using different types of outcome predictors can lead to different effect interpretations and that the challenges inherent to learning them differ. Ultimately, in applications the choice of estimand should be made by domain expert – who is also needed to verify that its *untestable* identifying assumptions hold – and correspond to their policy or research question of interest.

Acknowledgements

We would like to thank anonymous reviewers and members of the vanderschaar-lab for insightful comments on earlier drafts of this paper, and Mats Stensrud for a very interesting discussion on the topic. AC gratefully acknowledges funding from AstraZeneca.

References

Aalen, O. O. (1994). Effects of frailty in survival analysis. *Statistical Methods in Medical Research*, 3(3):227–243.

Alaa, A. and van der Schaar, M. (2018). Limits of estimating heterogeneous treatment effects: Guidelines for practical algorithm design. In *International Conference on Machine Learning*, pages 129–138.

Alaa, A. M. and van der Schaar, M. (2017a). Bayesian inference of individualized treatment effects using multitask gaussian processes. *Advances in Neural Information Processing Systems*, 30:3424–3432.

Alaa, A. M. and van der Schaar, M. (2017b). Deep multitask gaussian processes for survival analysis with competing risks. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 2326–2334.

Assaad, S., Zeng, S., Tao, C., Datta, S., Mehta, N., Henao, R., Li, F., and Duke, L. C. (2021). Counterfactual representation learning with balancing weights. In *International Conference on Artificial Intelligence and Statistics*, pages 1972–1980. PMLR.

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

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

Bellot, A. and van der Schaar, M. (2018a). Multitask boosting for survival analysis with competing risks. *Advances in Neural Information Processing Systems*, 31.

Bellot, A. and van der Schaar, M. (2018b). Tree-based bayesian mixture model for competing risks. In *International Conference on Artificial Intelligence and Statistics*, pages 910–918. PMLR.

Bica, I., Alaa, A. M., Jordon, J., and van der Schaar, M. (2020). Estimating counterfactual treatment outcomes over time through adversarially balanced representations. *International conference on learning representations*.

Byrd, J. and Lipton, Z. (2019). What is the effect of importance weighting in deep learning? In *International Conference on Machine Learning*, pages 872–881. PMLR.

Chapfuwa, P., Assaad, S., Zeng, S., Pencina, M. J., Carin, L., and Henao, R. (2021). Enabling counterfactual survival analysis with balanced representations. In *Proceedings of the Conference on Health, Inference, and Learning*, pages 133–145.

Cortes, C., Mansour, Y., and Mohri, M. (2010). Learning bounds for importance weighting. *Advances in neural information processing systems*, 23.

Curth, A., Lee, C., and van der Schaar, M. (2021a). Survite: Learning heterogeneous treatment effects from time-to-event data. *Proceedings of the Thirty-Fifth Conference on Neural Information Processing Systems*.

Curth, A., Svensson, D., Weatherall, J., and van der Schaar, M. (2021b). Really doing great at estimating cate? a critical look at ml benchmarking practices in treatment effect estimation. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Curth, A. and van der Schaar, M. (2021a). Nonparametric estimation of heterogeneous treatment effects: From theory to learning algorithms. In *International Conference on Artificial Intelligence and Statistics*, pages 1810–1818. PMLR.

Curth, A. and van der Schaar, M. (2021b). On inductive biases for heterogeneous treatment effect estimation. *Advances in Neural Information Processing Systems*, 34:15883–15894.

Fang, T., Lu, N., Niu, G., and Sugiyama, M. (2020). Rethinking importance weighting for deep learning under distribution shift. *Advances in Neural Information Processing Systems*, 33:11996–12007.

Fine, J. P. and Gray, R. J. (1999). A proportional hazards model for the subdistribution of a competing risk. *Journal of the American statistical association*, 94(446):496–509.

Gray, R. J. (1988). A class of k-sample tests for comparing the cumulative incidence of a competing risk. *The Annals of statistics*, pages 1141–1154.

Hassanpour, N. and Greiner, R. (2019). Counterfactual regression with importance sampling weights. In *IJCAI*, pages 5880–5887.

Hassanpour, N. and Greiner, R. (2020). Learning disentangled representations for counterfactual regression. In *International Conference on Learning Representations*.

Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240.

Johansson, F., Shalit, U., and Sontag, D. (2016). Learning representations for counterfactual inference. In *International conference on machine learning*, pages 3020–3029.

Johansson, F. D., Kallus, N., Shalit, U., and Sontag, D. (2018). Learning weighted representations for generalization across designs. *arXiv preprint arXiv:1802.08598*.

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

Kennedy, E. H. (2020). Optimal doubly robust estimation of heterogeneous causal effects. *arXiv preprint arXiv:2004.14497*.

Künzel, S. R., Sekhon, J. S., Bickel, P. J., and Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences*, 116(10):4156–4165.

Lambert, P. C., Dickman, P. W., Nelson, C. P., and Royston, P. (2010). Estimating the crude probability of death due to cancer and other causes using relative survival models. *Statistics in medicine*, 29(7-8):885–895.

Lee, C., Yoon, J., and Van Der Schaar, M. (2019). Dynamic-deephit: A deep learning approach for dynamic survival analysis with competing risks based on longitudinal data. *IEEE Transactions on Biomedical Engineering*, 67(1):122–133.

Lee, C., Zame, W. R., Yoon, J., and van der Schaar, M. (2018). Deephit: A deep learning approach to survival analysis with competing risks. In *Thirty-second AAAI conference on artificial intelligence*.

Li, Y., Jia, W., Kang, Y., Chen, T., Li, X., Du, X., Dong, J., Ma, C., Wang, F., and Xie, G. (2020). Deepcomp: Which competing event will hit the patient first? In *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 629–636. IEEE.

Lim, H. J., Zhang, X., Dyck, R., and Osgood, N. (2010). Methods of competing risks analysis of end-stage renal disease and mortality among people with diabetes. *BMC medical research methodology*, 10(1):1–10.

Louizos, C., Shalit, U., Mooij, J. M., Sontag, D., Zemel, R., and Welling, M. (2017). Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems*, pages 6446–6456.

Maia Polo, F. and Vicente, R. (2022). Effective sample size, dimensionality, and generalization in covariate shift adaptation. *Neural Computing and Applications*, pages 1–13.

Nie, X. and Wager, S. (2017). Quasi-oracle estimation of heterogeneous treatment effects. *arXiv preprint arXiv:1712.04912*.

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

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Prentice, R. L., Kalbfleisch, J. D., Peterson Jr, A. V., Flournoy, N., Farewell, V. T., and Breslow, N. E. (1978). The analysis of failure times in the presence of competing risks. *Biometrics*, pages 541–554.

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

Shalit, U., Johansson, F. D., and Sontag, D. (2017). Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, pages 3076–3085. PMLR.

Shi, C., Blei, D., and Veitch, V. (2019). Adapting neural networks for the estimation of treatment effects. In *Advances in Neural Information Processing Systems*, pages 2507–2517.

Stensrud, M. J., Young, J. G., Didelez, V., Robins, J. M., and Hernán, M. A. (2020). Separable effects for causal inference in the presence of competing events. *Journal of the American Statistical Association*, pages 1–9.

Stitelman, O. M. and van der Laan, M. J. (2010). Collaborative targeted maximum likelihood for time to event data. *The International Journal of Biostatistics*, 6(1).

Stojanov, P., Gong, M., Carbonell, J., and Zhang, K. (2019). Low-dimensional density ratio estimation for covariate shift correction. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3449–3458. PMLR.

Sugiyama, M., Krauledat, M., and Müller, K.-R. (2007). Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(5).

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.

Wang, Z. and Sun, J. (2022). Survtrace: transformers for survival analysis with competing events. In *Proceedings of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pages 1–9.

Wen, J., Yu, C.-N., and Greiner, R. (2014). Robust learning under uncertain test distributions: Relating covariate shift to model misspecification. In *International Conference on Machine Learning*, pages 631–639. PMLR.

Yoon, J., Jordon, J., and van der Schaar, M. (2018). Ganite: Estimation of individualized treatment effects using generative adversarial nets. In *International Conference on Learning Representations*.

Young, J. G., Stensrud, M. J., Tchetgen Tchetgen, E. J., and Hernán, M. A. (2020). A causal framework for classical statistical estimands in failure-time settings with competing events. *Statistics in medicine*, 39(8):1199–1236.

Zhang, Q. and Zhou, M. (2018). Nonparametric bayesian lomax delegate racing for survival analysis with competing risks. *Advances in Neural Information Processing Systems*, 31.

APPENDIX

This Appendix is structured as follows: In Appendix A, we present an additional literature review. In Appendix B, we discuss possible extensions incorporating more sophisticated solutions from the ML literatures on TTE prediction and HTE estimation. In Appendix C, we discuss identifying assumptions and proofs. In Appendix D, we give additional details of the simulation experiments presented in the main text. In Appendix E, we finally conduct additional experiments based on real data.

A ADDITIONAL LITERATURE REVIEW

Time-to-event prediction using ML in the presence of competing events. While modeling time-to-event data in the presence of competing events has been studied in the statistics literature for decades (Prentice et al., 1978; Gray, 1988; Fine and Gray, 1999), Lee et al. (2018) are the first paper we are aware of that consider time-to-event analysis with competing events in a (deep) machine learning context. Lee et al. (2018)'s proposed DeepHit assumes a discrete-time setup and uses fully connected neural networks with both shared and time-specific components for prediction of cause-specific incidence at every time-step. This work has since been extended further proposing improvements upon using simple neural networks in discrete time e.g. using RNNs in Li et al. (2020) or transformers in Wang and Sun (2022) and been complemented with work in continuous time relying on e.g. multi-task boosting (Bellot and van der Schaar, 2018a) or Bayesian nonparametric methods e.g. multi-task gaussian processes (Alaa and van der Schaar, 2017a), Lomax delegate racing (Zhang and Zhou, 2018) or tree-based mixture models (Bellot and van der Schaar, 2018b).

Heterogeneous treatment effect estimation using ML. The ML literature on HTE estimation has centered mainly on binary or continuous outcomes, and has expanded rapidly over the last years. One stream of work has provided model-agnostic strategies (also sometimes referred to as meta-learner strategies (Künzel et al., 2019)) to estimate HTEs using *any* ML method (Künzel et al., 2019; Nie and Wager, 2017; Kennedy, 2020; Curth and van der Schaar, 2021a). A majority of the work published in machine learning has, however, focussed on adapting *specific* ML methods for HTE estimation: early work relied mainly on tree-based methods (Hill, 2011; Athey and Imbens, 2016; Wager and Athey, 2018; Athey et al., 2019), but was followed by adaptations of e.g. Gaussian processes Alaa and van der Schaar (2017a, 2018) and GANs Yoon et al. (2018). At this point, the most popular solution seems to be to adapt neural networks for treatment effect estimation, see e.g. Johansson et al. (2016); Shalit et al. (2017); Johansson et al. (2018); Shi et al. (2019); Hassanpour and Greiner (2019, 2020); Assaad et al. (2021); Curth and van der Schaar (2021a,b). The work outlined above focusses exclusively on binary/continuous outcomes, thus closest to our setting are two recent papers that have investigated challenges inherent to HTE estimation for TTE data *without* competing events, focussing on covariate shift: Chapfuwa et al. (2021) used generative models for counterfactual TTE analysis in continuous time and Curth et al. (2021a) used neural networks for discrete time analyses but neither considers how to incorporate competing events.

Estimating treatment effects in the presence of competing events. The most likely reason for a lack of work on estimating *heterogeneous* treatment effects in the presence of competing events is that even the simpler *average* treatment effect setting has received rigorous characterisation within a causal framework only very recently: Young et al. (2020); Stensrud et al. (2020) are the first to formally characterize and define different types of causal effects and their identifying conditions within a counterfactual framework; their formalization allowed us to extrapolate their insights, combined with the literature on HTEs from the standard settings, to *heterogeneous* effects. Lacking such unified characterisation, prior work has considered estimation of average effects either in a model-dependent fashion by considering regression coefficients in cause-specific hazard models Prentice et al. (1978) or by testing for treatment-differences across e.g. cause-specific cumulative incidence functions Gray (1988), thus considering mainly total effects.

B POSSIBLE METHODOLOGICAL EXTENSIONS

Because we put our focus on understanding the unique challenges in adding competing events to the HTE estimation problem, we relied on simple ML methods to allow for clear empirical insights. Having gained understanding of the challenges inherent to the HTE competing events problem, an interesting next step would therefore be to consider how to adapt and incorporate ideas from the vast ML literature on HTE & TTE estimation to construct more sophisticated solutions. We discuss some possible avenues of interest for future research below.

A first approach would be to make the time-to-event (hazard) predictions – which we use to compute the effects – themselves better, for example by sharing information across time-steps. As described in section 3, any time-to-event model that allows to compute cause-specific hazard functions for every time-step could be used to estimate all three types of effects. Instead of fitting separate models per time-step as we do in our experiments, one could therefore flexibly share information across time-steps (and possibly across causes) as in discrete-time neural networks for TTE prediction (Lee et al., 2018; Li et al., 2020; Curth et al., 2021a; Wang and Sun, 2022). Complementing such a methodological extension, it would also be interesting to theoretically study whether sharing of information over time-steps might mitigate some covariate shift issues.

Further, the literature on domain adaptation and HTE estimation has proposed more sophisticated solutions to address covariate shift than classical importance weighting. As we allude to in footnote 4 in the main text, a large proportion of the ML HTE literature (e.g. Shalit et al. (2017); Johansson et al. (2018); Assaad et al. (2021)) has focused on learning balanced representations that minimize an empirical estimate of the IPM-term – which is not readily available in our setting as the interventional at-risk distribution is not equal to the marginal covariate distribution. Faced with a similar obstacle, Curth et al. (2021a) simply used the marginal distribution for balancing nonetheless – which they demonstrated to work well empirically. Improving upon this naive solution by investigating how to correctly balance representations may be an interesting next step both in their and our setting. A different avenue would be to improve upon using standard importance weights by removing uninformative dimensions: if we knew which features caused outcome, we would only need to compute importance weights taking into account distributional differences in dimensions that actually matter for prediction, which could significantly reduce variance in the importance weights and hence speed up learning (Stojanov et al., 2019; Maia Polo and Vicente, 2022). One possible way of doing so might be to jointly learning importance weights and representations for time-to-event prediction, adapting ideas from e.g. Hassanpour and Greiner (2019); Fang et al. (2020)

Finally, an interesting future direction may be to consider more complex data-types, e.g. allowing for (some) patient characteristics to be repeatedly measured over time, necessitating the incorporation of time-varying covariates. This is also an interesting scenario because it makes identifying assumptions relying on measuring all common causes of Y_k and D_k (Assumptions D1 and S1 in the following section) more likely to hold. Such covariates could easily be incorporated in our problem formulation and possible solutions could rely on recurrent networks such as Lee et al. (2019) in the TTE prediction setting and Bica et al. (2020) in the longitudinal treatment effect estimation setting.

C ASSUMPTIONS AND PROOFS

C.1 Formal Presentation of Identification Assumptions

Below we discuss assumptions for nonparametric *identification* of effects, which are based on those given in Young et al. (2020) for total and direct effect and Stensrud et al. (2020) for separable effects.

Some assumptions are shared by all causal parameters described in Section 3; they are analogues to the standard *ignorability* assumptions of Rosenbaum and Rubin (1983) from the standard treatment effect setting and are known as *unconfoundedness*, *overlap* and *consistency* assumptions: for each $k \in \{1, \dots, K\}$ we require
- **Assumption 1. Exchangeability w.r.t. treatment:** $Y^a_k, D^a_k \perp\!\!\!\perp A | X$
- **Assumption 2. Positivity w.r.t. treatment:** $\mathbb{P}(A = a | X = x) > 0$ for $\forall x : \mathbb{P}(X = x) > 0$ and $a \in \{0, 1\}$
- **Assumption 3. Consistency w.r.t. treatment assignment:** We observe the counterfactuals associated with the given treatment A, i.e. $Y_k = AY^1_k + (1 - A)Y^0_k$ and $D_k = AD^1_k + (1 - A)D^0_k$

While total effects require no additional assumptions, both direct and separable effects require further identification assumptions associated with the additional interventions performed in their definitions.

Direct effects. Direct effects require additional unconfoundedness, overlap and consistency assumptions to identify distributions under elimination of the competing event. For each $k \in \{1, \dots, K\}$ we require
- **Assumption D1. Exchangeability: w.r.t. competing event** $Y^a_k \perp\!\!\!\perp D^a_k | X, \bar{Y}_{k-1} = \bar{D}_{k-1} = 0, A = a$
- **Assumption D2: Positivity w.r.t. competing event:** $\mathbb{P}(D_k = 0 | X = x, \bar{Y}_{k-1} = \bar{D}_{k-1} = 0, A = a) > 0$ whenever $\mathbb{P}(X = x, \bar{Y}_{k-1} = \bar{D}_{k-1} = 0, A = a) > 0$
- **Assumption D3. Consistency w.r.t. elimination of competing event:** For an observation with $A = a$ and $\bar{D}_k = 0$ we observe the corresponding counterfactual, i.e. $\bar{Y}_k = \bar{Y}^{a,\bar{d}=0}_k$.

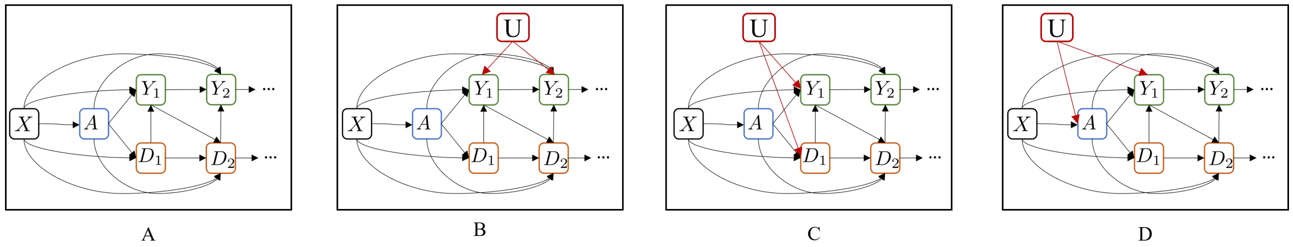


Figure 6: Figure illustrating causal graphs with hidden variables in which different effects are identified. (A): No hidden variables, all effects are identified. (B): Hidden variable causing all Y_k, all effects are identified. (C): Hidden variable causing Y_k and D_k; total effect is identified but separable and direct effect are not. (D): Hidden confounder of outcome treatment association, no effect is identified.

Separable effects. Stensrud et al. (2020) derived assumptions enabling identification of separable (in)direct effects; owing to the conceptual difference in intervention on only *parts* of the treatment, these are more involved to state than those above. In particular, for each $k \in \{1, \dots, K\}$ we require:

- **Assumption S0. Conceptual assumptions defining separable treatment:** A is separable into components A_Y and A_D. Setting $A = a$ is equivalent to setting both A_Y and A_D to a, so that $Y_k^{a_Y=a, a_D=a} = Y_k^a$ and$D_k^{a_Y=a,a_D=a} = D_k^a$. Further, A_Y exerts effects on D_k only through its effect on $\bar{Y}_{k-1}$ i.e.

$$Y_{k-1}^{a_Y=1,a_D} = D_{k-1}^{a_Y=1,a_D} = Y_{k-1}^{a_Y=0,a_D} = D_{k-1}^{a_Y=0,a_D} = 0 \implies D_k^{a_Y=1,a_D} = D_k^{a_Y=0,a_D} \text{ for } a_D \in \{0,1\} \tag{5}$$

and conversely, A_D exerts effects on Y_k only through its effect on $\bar{D}_k$ i.e.

$$Y_{k-1}^{a_Y,a_D=1} = D_k^{a_Y,a_D=1} = Y_{k-1}^{a_Y,a_D=0} = D_k^{a_Y,a_D=0} = 0 \implies Y_k^{a_Y,a_D=1} = Y_k^{a_Y,a_D=1} \text{ for } a_Y \in \{0,1\} \tag{6}$$

- **Assumption S1. Dismissible Component conditions:** W.r.t. primary event

$$\begin{aligned} \mathbb{P}(Y_k^{a_Y,a_D=1} = 1 | Y_{k-1}^{a_Y,a_D=1} = 0, D_k^{a_Y,a_D=1} = 0, X = x) = \\ \mathbb{P}(Y_k^{a_Y,a_D=0} = 1 | Y_{k-1}^{a_Y,a_D=0} = 0, D_k^{a_Y,a_D=0} = 0, X = x) \end{aligned} \tag{7}$$

and w.r.t. competing event

$$\begin{aligned} \mathbb{P}(D_k^{a_Y=1,a_D} = 1 | Y_{k-1}^{a_Y=1,a_D} = 0, D_k^{a_Y=1,a_D} = 0, X = x) = \\ \mathbb{P}(D_k^{a_Y=0,a_D} = 1 | Y_{k-1}^{a_Y=0,a_D} = 0, D_k^{a_Y=0,a_D} = 0, X = x) \end{aligned} \tag{8}$$

- **Assumption S3. Positivity w.r.t. treatment (in surviving population):** $\mathbb{P}(A = a | \bar{D}_k = \bar{Y}_k = 0, X = x) > 0$ whenever $\mathbb{P}(\bar{D}_k = \bar{Y}_k = 0, X = x) > 0$ for $a \in \{0, 1\}$
- **Assumption S3. Consistency:** For an observation with $A = a$, we observe the corresponding counterfactuals, i.e. $Y_k = Y_k^{a,a}$ and $D_k = D_k^{a,a}$.

C.2 Discussion of Assumptions.

Consistency assumptions are always needed to ensure that we observe *one* of the counterfactuals for each individual; these assumptions may not hold if e.g. the act of monitoring outcomes can change their value or if there is interference between individual units. Positivity assumptions ensure that there is some overlap between observational and interventional distributions; if these assumptions do not hold we could not (nonparametrically) extrapolate to the interventional distribution (and importance weights $w^*(x)$ would not be defined for all x).

Finally, exchangeability assumptions and the dismissible component assumptions ensure that there are no *hidden* variables (variables not included in X) causing treatment assignment *and* events (all effects) and Y_k *and* D_k (separable and direct effect only), which would make observed distributions inherently different from distributions under intervention. To illustrate when these assumptions do not hold, we highlight scenarios where different hidden variables do (not) violate assumptions in Fig. 6. In Panel A (the same as Fig. 1 in the main text), no hidden variables exist and all effects are identified. In Panel B, hidden variables causing both Y_1 and Y_2 exist – e.g. some form of underlying frailty (Aalen, 1994), inducing heterogeneity in risk of *only* the main event occuring – which are allowed under all effects. In Panel C, Y_k and D_k are caused by some shared hidden variable, which violates assumptions D1 and S1 – thus separable and direct effects are not identified, while the total effect is. Finally, in Panel D, there is a hidden confounder of the treatment-outcome association, which violates assumption 1 – therefore, no effect is identified.

C.3 Identification of interventional at-risk covariate distribution for separable effects (Sec. 4.2)

The interventional at-risk covariate distribution corresponding to the intervention $do(A_Y = a_Y, A_D = a_D)$ can be identified from observational data under the assumptions stated above and is proportional to $\mathbb{P}(X{=}x)(1{-}h_D(k,x,a_D)) \times \prod_{l=1}^{k-1}(1{-}h_D(l,x,a_D))(1{-}h_Y(l,x,a_Y))$ as stated in Sec. 4.2.

Proof: the interventional distribution is proportional to

$$\begin{aligned}&\mathbb{P}(X{=}x)\mathbb{P}(D_k^{a_Y,a_D}{=}0|D_{k-1}^{a_Y,a_D}{=}0, Y_{k-1}^{a_Y,a_D}{=}0, X{=}x)\\ \times \prod_{l=1}^{k-1} \mathbb{P}(Y_l^{a_Y,a_D}{=}0|D_l^{a_Y,a_D}{=}0, Y_{l-1}^{a_Y,a_D}{=}0, X{=}x)&\mathbb{P}(D_l^{a_Y,a_D}{=}0|D_{l-1}^{a_Y,a_D}{=}0, Y_{l-1}^{a_Y,a_D}{=}0, X = x)\end{aligned} \tag{9}$$

which is equal to

$$\begin{aligned}&\mathbb{P}(X{=}x)\mathbb{P}(D_k^{a_D,a_D}{=}0|D_{k-1}^{a_D,a_D}{=}0, Y_{k-1}^{a_D,a_D}{=}0, X{=}x)\\ \times \prod_{l=1}^{k-1} \mathbb{P}(Y_l^{a_Y,a_Y}{=}0|D_l^{a_Y,a_Y}{=}0, Y_{l-1}^{a_Y,a_Y}{=}0, X{=}x)&\mathbb{P}(D_l^{a_D,a_D}{=}0|D_{l-1}^{a_D,a_D}{=}0, Y_{l-1}^{a_D,a_D}{=}0, X = x)\end{aligned} \tag{10}$$

because of the dismissible component conditions if $a_Y \neq a_D$, and trivially if $a_Y = a_D$.

This is equal to

$$\begin{aligned}&\mathbb{P}(X{=}x)\mathbb{P}(D_k^{a_D}{=}0|D_{k-1}^{a_D}{=}0, Y_{k-1}^{a_D}{=}0, X{=}x)\\ \times \prod_{l=1}^{k-1} \mathbb{P}(Y_l^{a_Y}{=}0|D_l^{a_Y}{=}0, Y_{l-1}^{a_Y}{=}0, X{=}x)&\mathbb{P}(D_l^{a_D}{=}0|D_{l-1}^{a_D}{=}0, Y_{l-1}^{a_D}{=}0, X = x)\end{aligned} \tag{11}$$

$$\begin{aligned}&= \mathbb{P}(X{=}x)\mathbb{P}(D_k{=}0|D_{k-1}{=}0, Y_{k-1}{=}0, X{=}x, A{=}a_D)\\ \times \prod_{l=1}^{k-1} \mathbb{P}(Y_l{=}0|D_l{=}0, Y_{l-1}{=}0, X{=}x, A{=}a_Y)&\mathbb{P}(D_l{=}0|D_{l-1}{=}0, Y_{l-1}{=}0, X = x, A{=}a_D)\end{aligned} \tag{12}$$

$$= \mathbb{P}(X{=}x)(1{-}h_D(k,x,a_D)) \times \prod_{l=1}^{k-1}(1{-}h_D(l,x,a_D))(1{-}h_Y(l,x,a_Y) \tag{13}$$

by consistency (assumptions S0 and S3), exchangeability (assumption 1) and by definition of the hazard functions, respectively.

C.4 Proof of proposition 1 (Section 4.3)

Proposition 2 *(Restated) Given timestep k and treatment a, for a loss function $\ell_h \in [0,1]$ of any hypothesis $h \in \mathcal{H}$, such that $d = Pdim(\{\ell_h : h \in \mathcal{H}\})$ (where Pdim is the pseudo-dimension) and $\ell_h \in \mathcal{L}$, where $\mathcal{L}$ is a space of pointwise loss functions, and a weighting function $w(x)$ with $\mathbb{E}[w(X)] = 1$, with probability $1-\delta$ over at-risk sample $\mathcal{D}^Y_{a,k}$ with empirical distribution $\hat{p}^{obs}_{a,k}$, we have*

$$R^{int}_{a,k}(h) - \hat{R}^{w,obs}_{a,k}(h) \leq \left|\mathbb{E}_{p^{obs}_{a,k}}[(w^*_{a,k}(x) - w(x))\ell_h(x)]\right| + \frac{\max(\sqrt{\mathbb{E}_{p^{obs}_{a,k}}[w^2(x)l^2_h(x)]}, \sqrt{\mathbb{E}_{\hat{p}^{obs}_{a,k}}[w^2(x)l^2_h(x)]})}{{n^Y_{a,k}}^{3/8}}\mathcal{C}^{\mathcal{H}}_{n^Y_{a,k}} \tag{14}$$

with $\mathcal{C}^{\mathcal{H}}_{n^Y_{a,k}} = 2^{5/4}(d \log \frac{e2n^Y_{a,k}}{d} + \log \frac{4}{\delta})^{3/8}$ due to Cortes et al. (2010) (Theorem 4).

For $w(x){=}w^*_{a,k}(x)$ (exact importance weights, Lemma 1), we have

$$R^{int}_{a,k}(h){-}\hat{R}^{w^*,obs}_{a,k}(h) \leq \frac{1}{\sqrt{ESS^*_{rel}(p^{int}_{a,k}, p^{obs}_{a,k}) n^{Y}_{a,k}}^{3/8}} \mathcal{C}^{\mathcal{H}}_{n^Y_{a,k}} \tag{15}$$

where $ESS^*_{rel} = \exp_2(D_2(p^{int}_{a,k}||p^{obs}_{a,k}))$ is the expected relative effective sample size, with $D_2(p||q) = log_2\mathbb{E}_{x\sim p}\left[\frac{p(x)}{q(x)}\right]$ the Rényi divergence of order 2.

For $w(x) = 1$ (standard supervised learning, Lemma 2), we have

$$R^{int}_{a,k}(h){-}\hat{R}^{obs}_{a,k}(h) \leq C_{\mathcal{L}} IPM_{\mathcal{L}}(p^{int}_{a,k}, p^{obs}_{a,k}) + {n^{Y}_{a,k}}^{-3/8} \mathcal{C}^{\mathcal{H}}_{n_{a,k}} \tag{16}$$

where IPM$_{\mathcal{G}}(p, q) = \sup_{g\in\mathcal{G}} \left|\int g(x)(p(x) - q(x))dx\right|$ is an integral probability metric and $C_{\mathcal{L}}{>}0$ is s.t. $\frac{\ell}{C_{\mathcal{L}}} \in \mathcal{L}$.

Proof: Eq. (14) follows directly from Thm 4 in Cortes et al. (2010).

Further, when $w(x) = w^*_{a,k}(x)$, the statement in eq. (15, follows directly from Theorem 3 of Cortes et al. (2010), where we used the restatement in terms of ESS^*_{rel} of Maia Polo and Vicente (2022).

Finally, to prove equation (16), note that when $w(x) = 1$, we can bound the first term of eq. (14) as

$$\begin{aligned}\left|\mathbb{E}_{p^{obs}_{a,k}}[(w^*_{a,k}(x) - 1)\ell_h(x)]\right| &= \left|\mathbb{E}_{p^{obs}_{a,k}}[(\frac{p^{int}_{a,k}}{p^{obs}_{a,k}} - 1)\ell_h(x)]\right| = \left|\int \ell_h(x)(p^{int}_{a,k}(x) - p^{obs}_{a,k}(x))dx\right| \\ &\leq C_{\mathcal{L}} \sup_{f\in\mathcal{L}} \left|\int f_h(x)(p^{int}_{a,k}(x) - p^{obs}_{a,k}(x))dx\right| \leq C_{\mathcal{L}} IPM_{\mathcal{L}}(p^{int}_{a,k}, p^{obs}_{a,k})\end{aligned} \tag{17}$$

as in e.g. Shalit et al. (2017); Johansson et al. (2018)'s bounds for the standard treatment effect estimation setting. Note further that because $\ell_h \in [0, 1]$ by assumption, $\max(\sqrt{\mathbb{E}_{p^{obs}_{a,k}}[w^2(x)l^2_h(x)]}, \sqrt{\mathbb{E}_{\hat{p}^{obs}_{a,k}} w^2(x)l^2_h(x)]}) \leq 1$ in the second term. Putting the two together gives (16).

D EXPERIMENTAL DETAILS

Synthetic DGPs For clarity, we explicitly write out the DGPs used to create settings 1-4 in the main text below.

Setting 1 (ξ varies): $h_Y(k, x, a) = 0.01(1 - x_1) + 0.1x_1$ and $h_D(k, x, a) = 0.01$ and $\pi(x)$=expit($\xi(x_1{-}0.5)$)

Setting 2 ($p^D_{high,\tau}$ varies): $h_Y(k, x, a) = 0.01(1 - x_1) + 0.1x_1$ and $h_D(k, x, a) = 0.01 + p^D_{high,\tau} * x_1 * a$ and $\pi(x) = 0.5$

Setting 3 (p^D_{high} varies): $h_Y(k, x, a) = 0.01(1 - x_1) + (0.1 - 0.09a) * x_1$ and $h_D(k, x, a) = 0.01(1 - x_1) + p^D_{high} * x_1$ and $\pi(x) = 0.5$

Setting 4 (ξ varies): $h_Y(k, x, a) = 0.01(1{-}x_1){+}0.1x_1$ and $h_D(k, x, a) = 0.01{+}0.1{*}x_1{*}a$ and $\pi(x)$=expit($\xi(x_1{-}0.5)$)

with $X_1 \sim \mathcal{B}(0.5)$ and $X_2 \sim \mathcal{B}(0.5{-}\rho(1{-}2X_1))$ and $\rho = .35$ for all Figures except in Fig. 6 panel A&B where $\pi(x)$ depends on x_2 instead of x_1.

Implementations[7] Throughout, as unrestricted/correctly specified estimators for the hazard of the competing event $\hat{h}_D$ for each time step and for all propensity estimators $\hat{\pi}$, we use logistic regressions (LR), relying on the sklearn(Pedregosa et al., 2011) implementation with default parameters and l2-penalty $C = 100$. As described in the main text, we use constants to induce a misspecified hazard model for the main event at each time step . For the results using LRs for the main event hazard in Fig. 5C&D in the main text, we set l2-penalty $C = 1$ in Fig. 5C to reduce capacity of the model and create a misspecified model, and $C = 100$ for the unrestricted version in Fig. 5D.

[7] Code to replicate all experiments can be found at `https://github.com/AliciaCurth/CompCATE` or `https://github.com/vanderschaarlab/CompCATE`.

E ADDITIONAL EXPERIMENTS USING REAL DATA

E.1 Creating a semi-synthetic benchmark from the Twins dataset

Due to the usual absence of counterfactuals in practice, benchmarking treatment effect estimation methods on real data is a challenging problem (Curth et al., 2021b). In our setting, this problem is exacerbated relative to the standard HTE setting because there are more interventions than 'just' intervention on treatment, meaning that there are even more unobserved counterfactuals (namely those corresponding to interventions on competing events or separable interventions).

The Twins Dataset. The (real-world) Twins benchmark dataset used in Louizos et al. (2017); Yoon et al. (2018) for binary outcomes and in Curth et al. (2021a) for TTE outcomes *without* competing events is an interesting exception as Twins could be treated as their respective counterfactual under treatment – which has been exploited in the standard setting where we only wish to intervene on treatment: The dataset consists of 11400 pairs of twins, for whom one can create a binary treatment such that $a = 1$ denotes being the heaver twin at birth, and use this to emulate a hypothetical study measuring the HTE of birthweight on 1-year infant mortality (binary outcome) or survival times (in days, administratively censored at t=365). Note that, fortunately, the mortality rate is relatively low, giving an event rate of around 16% over the full horizon. The dataset as used in Yoon et al. (2018); Louizos et al. (2017); Curth et al. (2021a) contains 30 covariates for each twin relating to the parents, the pregnancy, and the birth (e.g., marital status, race, residence, number of previous births, pregnancy risk factors, quality of care during pregnancy, and number of gestation weeks prior to birth), of which we use 27 in our experiments (we dropped the three categorical features due to low variance).

Semi-synthetic benchmarking setup. We use this dataset as a basis for a semi-synthetic benchmarking setup in which we use the original covariates and event times to create an observational time-to-event dataset with competing events by (i) selectively observing only one twin during training and (ii) introducing a simulated competing event. As both are simulated, we can intervene on these processes to give oracle solutions based on interventional distributions to compare to the solutions obtained from observational distributions. Note that, because most events happen early on – 80% of events happen on days 0-10, with 60% of events occuring on day 0 – we consider only the first 10 days, so that here $\mathcal{K} = \{0, \ldots, 10\}$.

In training, we selectively observe only one of the two twins, and for (i) induce confounding by sampling $A|X \sim \mathcal{B}(\pi(X))$ with propensity score $\pi(X_i) = \text{expit}(\xi_A \mathcal{Z}_{train}(|\mathcal{S}|^{-1} \sum_{p \in \mathcal{S}} X_{i,p}))$ where $\mathcal{Z}_{train}(\cdot)$ denotes standardization over the training set, ξ_A determines the strength of selection and $\mathcal{S}$ is a feature subset which is chosen as discussed below. For (ii), using the observed trajectories $\bar{Y}^a_K$ from the Twins dataset, we introduce competing events by simulating $D^a_k \sim h_d(k, x, a)$ for units with $Y^a_{k-1} = 0$, and setting all future values of Y^a_k to zero whenever a competing event occurs. As a hazard we use

$$h_D(k, x, a) = \begin{cases} \text{expit}\left(\log(0.1) + \xi_D(1-a)\mathcal{Z}_{train}(|\mathcal{S}|^{-1} \sum_{p \in \mathcal{S}} x_p)\right) \text{ if } k = 1 \\ \frac{0.1}{k-1} \text{ otherwise} \end{cases} \tag{18}$$

which mimics the main outcome in that most events and heterogeneity occurs initially in period 0 and then levels of. As above $\mathcal{Z}_{train}(\cdot)$ denotes standardization over the training set and ξ_D determines the heterogeneity in the competing event process. Note that treatment here *equalizes* the odds of the competing event across individuals.

As before, we consider 3 settings to be able to disentangle the different forces at play here. In setting 1, we remove the competing event and consider only the effect of treatment selection by varying ξ_A. In setting 2, we assign treatment randomly and consider only the effect of the covariate shift induced by the competing event by varying ξ_D. In setting 3, we combine the two, set $\xi_A = 2$ and vary ξ_D.

For both propensity and competing hazard, to ensure that the variables determining treatment assignment and competing event are actually important for the main event, we choose set $\mathcal{S}$ by selecting the most important covariates from a (treatment-agnostic) random forest for predicting mortality at time step 0 on the Twins data (using sklearn's `SelectFromModel` class). In Fig. 8 in the experiments below, we show that choosing outcome-relevant features in such manner makes a difference in an experiment where we instead randomly sample a feature set of size $|\mathcal{S}|$.

Estimators. Because the event data is real and not simulated, unlike in the main text, we do not know what correctly specified model is in this case. Here, we therefore consider as base models a constant model as in the main text, as well as a LR with cross-validated l1-penalty in $\{10^{-3}, 10^{-2}, 10^{-1}\}$ and a random forest with 100 trees. The competing events and propensity models are once more fit using a (correctly specified) LR with cross-validated l1-penalty in $\{10^{-3}, 10^{-2}, 10^{-1}, 1\}$. The four learning strategies – observational, weighting (true and estimated) and counterfactual

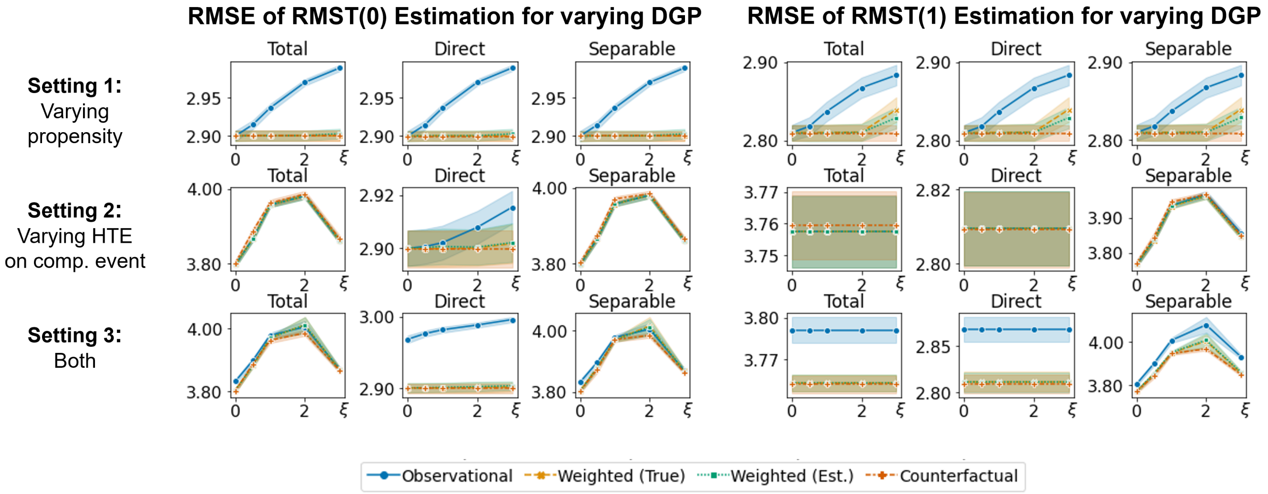


Figure 7: RMSE of estimating RMST under control (left) and under treatment (right) under the different interventions (columns) across different DGPs (rows) using (misspecified) constant estimators per time-step on Twins. Shaded area indicates one standard error.

– are as before. Note that because treatment assignment and competing event are simulated, ground truth weights and counterfactuals are accessible.

Evaluation. As the main event data is real and ground truth probabilities are unknown, we instead evaluate all models in terms of their predictions of event-free restricted mean survival time under intervention $RMST^{int}_K = min(T^{int}, K)$ which we can compute from the observed (twins) event times $T^{Y,twins}(a)$ and simulated competing event times $T^{D,sim}(a)$ (both are set to K+1 if event never occurs) as

$$RMST^{int}_K = \begin{cases} min(T^{Y,twins}(a), T^{D,sim}(a), K) \text{ if } int = do(A=a) \\ min(T^{Y,twins}(a), K) \text{ if } int = do(A=a, \bar{D}=0) \\ min(T^{Y,twins}(a_Y), T^{D,sim}(a_D), K) \text{ if } int = do(A_Y=a_y, A_D=a_D) \end{cases}$$

The expected value of the RMST is equal to the area under the event-free survival curve, which for the different interventions can be computed from the hazard functions as

$$\mathbb{E}[RMST^{int}_K] = \begin{cases} 1+\sum_{l=1}^{K-1}\left[\sum_{q=1}^{l}(1-h_D(q,x,a))(1-h_Y(q,x,a))\right] \text{ if } int = do(A=a) \\ 1+\sum_{l=1}^{K-1}\left[\sum_{q=1}^{l}(1-h_Y(q,x,a))\right] \text{ if } int = do(A=a, \bar{D}=0) \\ 1+\sum_{l=1}^{K-1}\left[\sum_{q=1}^{l}(1-h_D(q,x,a_D))(1-h_Y(q,x,a_Y))\right] \text{ if } int = do(A_Y=a_y, A_D=a_D) \end{cases}$$

Below we will sometimes refer to RMST(a); this refers to the different versions of RMST evaluated for $A = a$ at time $K = 10$ (only for the separable direct effect, RMST(0) corresponds to $a_Y = a_D = 0$ and RMST(1) corresponds to $a_Y = 1$, $a_D = 0$). We report the RMSE of estimating RMST using the hazards from the different models, $\sqrt{\frac{1}{n_{test}}\sum_{i=1}^{n_{test}}(RMST^{int}_{K,i} - \hat{\mathbb{E}}[RMST^{int}_K])^2}$; here we split the data 50/50 for training and testing (by twin pairs), and report means and standard errors of the RMSE across 5 replications of each experiment. Code to replicate the experiments will be released upon acceptance of the paper.

E.2 Results using the Twins data

In Fig. 7 we present results for all three settings and interventions corresponding to the three effects under consideration, using constant models at each time step to possibly induce misspecification. We observe that many of the conclusions from the stylized simulations from the main text carry over also to this more complex and realistic setup based on real data. We observe that varying confounding through ξ_A (top row) remains important for all effects. We observe that only the estimator of the control[8] RMST in the direct effect setting is impacted by varying the competing event intensity ξ_D *alone* (middle row), while when adding the two together we observe that varying ξ_D can impact also estimation of the other effects.

We also note that all differences between approaches overall appear much more salient for the RMST of the control group than for the treated group. While for settings 2 and 3 it is partially a consequence of how we designed the competing event hazard function, this is not the case for setting 1 – here it may give some evidence that the outcome model in the treated population is not substantially misspecified and may be near constant; therefore we now focus on the control RMST for the remainder of this section. Additionally, we focus on estimation under intervention $do(A = a, \bar{D} = 0)$, corresponding to direct effects, for which the results are most pronounced.

We next consider the effect of having selected outcome-relevant covariates for treatment assignment and competing event. In Fig. 8 we instead select random features and observe that indeed, compared to outcome-relevant features as in Fig. 7, the shift induced in the different settings no longer systematically plays a role.

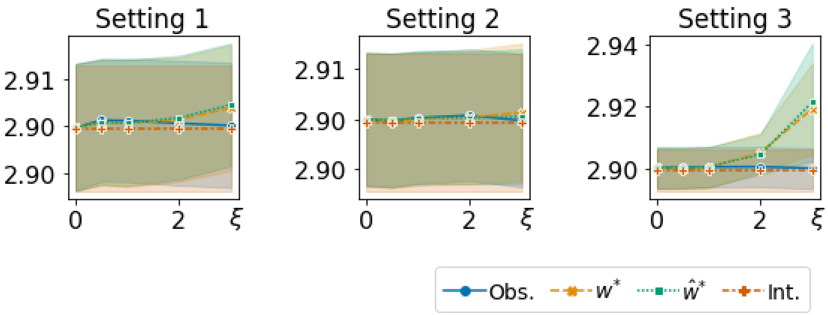
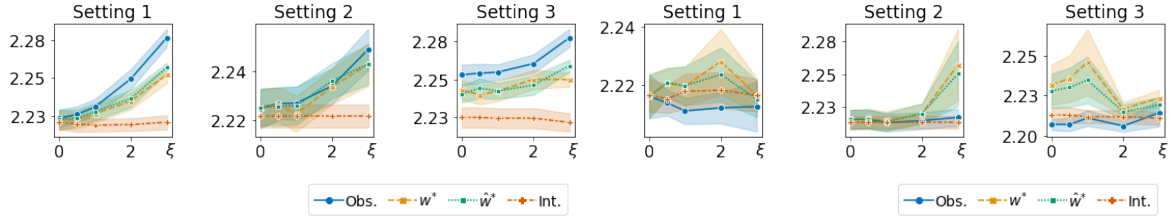


Figure 8: RMSE of estimating control RMST under intervention $do(A = 0, \bar{D} = 0)$ (corresponding to direct effects) using a constant model for the three settings when features in $\mathcal{S}$ are randomly chosen.

Finally, we use more flexible models – random forests (RFs) and logistic regressions (LRs) – instead of the simple constant models (returning to a setting where $\mathcal{S}$ is outcome-relevant). In Fig. 9 we present results for the three settings. We observe that for RFs, these covariate shifts appear to also play a role similarly to the constant estimator, albeit with smaller magnitude in effect on the estimation error. Interestingly, for LRs we do not observe any performance degradation of the observational solution relative to the counterfactual solution (and if anything, we observe that variance induced by weighting sometimes degrades performance). As robustness to arbitrary distribution shifts can indicate correct specification (Wen et al., 2014), this may provide some evidence that the LR-model is actually *correctly* specified to capture the complexity of the underlying Twins time-to-event outcome data.



(a) Random forest

(b) Logistic Regression

Figure 9: RMSE of estimating control RMST under intervention $do(A = 0, \bar{D} = 0)$ (corresponding to direct effects) using random forests (left) and logistic regressions (right) for the three settings.

[8] Note that, due to our hazard specification for the competing event, there is no covariate shift in the treatment group for the direct effect, thus it is expected that behaviour in estimation of RMST(1) is not impacted by ξ_D in rows 2 and 3