Accounting For Informative Sampling When Learning to Forecast Treatment Outcomes Over Time

Toon Vanderschueren [*1 2] **Alicia Curth** [*3] **Wouter Verbeke** [1] **Mihaela van der Schaar** [3 4]

Abstract

Machine learning (ML) holds great potential for accurately forecasting treatment outcomes over time, which could ultimately enable the adoption of more individualized treatment strategies in many practical applications. However, a significant challenge that has been largely overlooked by the ML literature on this topic is the presence of *informative sampling* in observational data. When instances are observed irregularly over time, sampling times are typically not random, but rather informative–depending on the instance's characteristics, past outcomes, and administered treatments. In this work, we formalize informative sampling as a covariate shift problem and show that it can prohibit accurate estimation of treatment outcomes if not properly accounted for. To overcome this challenge, we present a general framework for learning treatment outcomes in the presence of informative sampling using inverse intensity-weighting, and propose a novel method, TESAR-CDE, that instantiates this framework using Neural CDEs. Using a simulation environment based on a clinical use case, we demonstrate the effectiveness of our approach in learning under informative sampling.

1. Introduction

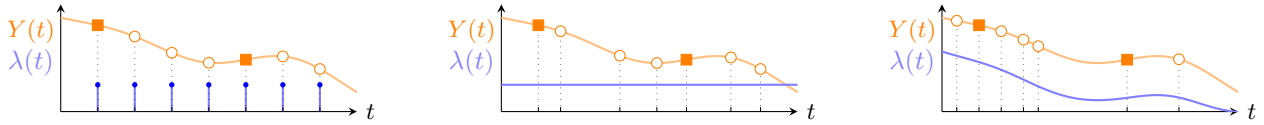
Due to its importance in applications ranging from economics to healthcare and marketing, the problem of estimating personalized causal effects of actions – e.g., treatments, interventions, or policies – has received wide attention in the recent machine learning (ML) literature (Curth & van der Schaar, 2021b). Effectively using real data for estimating such effects requires dealing with unique challenges arising from its observational nature. Therefore, the recent ML literature on treatment effect estimation has paid great attention to solving methodological issues arising due to treatment assignment biases in static (Shalit et al., 2017) and longitudinal settings (Bica et al., 2019). This paper focuses on another challenge that has been largely overlooked by the ML literature on treatment effect estimation, despite its relevance and prevalence in practice: the problem of *informative sampling*, sometimes also called *informed presence bias* (Goldstein et al., 2016). That is, in observational data, the timing at which an observation was made is often not random, but rather indicative of some underlying information relevant to the estimation problem of interest.

In electronic health records, for example, patients are typically not recorded *randomly* over time, but *informatively* (Lin et al., 2004): observations are only recorded at irregular visits to a health care provider, with visit times typically depending on the patient's past and present characteristics, evolving health state, and administered treatments. The resulting sampling mechanism is inherently intertwined with the patient's observed outcomes and treatments, with more check-ups being scheduled for patients in critical condition or to follow up after a treatment. Throughout this work, we will refer to examples from health care due to their societal relevance and intuitive appeal, but the problem of informative sampling appears in a wide variety of other domains, such as policy design (Lin et al., 2004), epidemiology (Del Moral & Murray, 2015), economics (Clithero, 2018), or maintenance (Vanderschueren et al., 2023).

Informative sampling poses an important challenge as it can bias estimates of causal effects when not accounted for (Robins et al., 1995; Lin et al., 2004; McCulloch et al., 2016). Intuitively, informative sampling leads to relatively more measurements of abnormal values and fewer measurements of normal values and, therefore, selection bias in the data (Liu et al., 2008; Gasparini et al., 2020). Standard statistical methods can estimate causal effects in the presence of informative sampling given a well-specified model of the sampling mechanism (Hernán et al., 2009). However, existing approaches for modeling the sampling mechanism from the (bio)statistics literature assume a certain parametric

[*]Equal contribution [1]Decision Sciences and Information Management, KU Leuven [2]Applied Mathematics, University of Antwerp [3]DAMTP, University of Cambridge [4]The Alan Turing Institute. Correspondence to: Toon Vanderschueren <toon.vanderschueren@kuleuven.be>, Alicia Curth <amc253@cam.ac.uk>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).



(a) Regular sampling (b) Sampling completely at random (SCAR) (c) Sampling at random (SAR)

Figure 1: **Problem illustration: sampling mechanisms.** We show an instance's latent trajectory ($Y(t)$) and sampling intensity ($\lambda(t)$) over time t, along with administered treatments (■) and observations (○) resulting from different sampling mechanisms. **(a) Regular.** Samples are obtained at regular intervals over time. **(b) SCAR.** Samples are irregular, drawn at completely random intervals over time. **(c) SAR.** Sampling times are *irregular, but not completely random*: e.g., there might be more samples when the outcome is large. We refer to the dependence of the sampling intensity on an instance's covariates, treatments, and/or outcomes as informative sampling. Whereas existing work in the ML literature assumes regular sampling or SCAR, this work is, to the best of our knowledge, the first to consider learning to forecast treatment outcomes given SAR.

form or latent variable(s), which might not match the actual data-generating process. Moreover, Farzanfar et al. (2017)'s survey on longitudinal healthcare research finds that these methods are rarely used in practice, leaving potential bias largely unaddressed. Therefore, this work examines the use of flexible ML methods for this task and investigates the unique methodological challenges arising therein.

Related work.[1] Since the initial seminal ML work on heterogeneous treatment effect estimation considering binary treatments and static data (Johansson et al., 2016; Shalit et al., 2017), this literature has grown rapidly both by making methodological refinements in the original setting (Hassanpour & Greiner, 2020; Curth & van der Schaar, 2021a) and by considering new settings, such as continuous treatments (Bica et al., 2019) or survival outcomes (Curth et al., 2021). Recent extensions have specifically explored using ML methods for estimating treatment effects *over time*, such as RNNs (Lim et al., 2018; Bica et al., 2019; Li et al., 2021; Berrevoets et al., 2021), transformers (Melnychuk et al., 2022), and Neural ODEs (Gwak et al., 2020; De Brouwer et al., 2022) or Neural CDEs (Seedat et al., 2022).

Most existing ML work on causal inference in a temporal setting has, to the best of our knowledge, implicitly relied on strict assumptions regarding the sampling mechanism. The majority assumes regular and uninformative sampling times (Fig. 1a). Only very recent work relying on neural differential equations to model the effects of treatment in continuous time (Gwak et al., 2020; Seedat et al., 2022; De Brouwer et al., 2022) *allows* for observations to be irregular (Fig. 1b), but *does not consider or account for potential bias* resulting from sampling times being informative rather than completely random, which is the focus of this work (Fig. 1c). This stands in stark contrast to the close attention paid in the treatment effect estimation literature to other sources of *covariate shift* arising in observational data, e.g., due to static treatment assignment (Johansson et al., 2016), treatment assignment over time (Bica et al., 2019), censoring (Curth et al., 2021) or competing events (Curth & van der Schaar, 2023). In this spirit, we find it important to study when and how the informativeness of sampling acts as an additional source of covariate shift in this setting.

Contributions. Despite the rapid recent expansion of the ML literature on estimating treatment effects, we believe that there is still a fundamental lack of understanding, or even formalization, of the challenges arising due to one of the most fundamental features of observational data: *sampling, the act of observation itself, can be inherently informative*. Therefore, we focus on understanding and analyzing the challenges that arise from informative sampling and propose strategies to alleviate bias arising in this context. In doing so, we make three contributions: **(1)** We formalize the problem of forecasting treatment outcomes under informative sampling as a machine learning problem and characterize the key challenges arising therein as a consequence of covariate shift induced by informative sampling. **(2)** We present a general strategy for tackling this challenge and propose a novel method for learning under informative sampling, TESAR-CDE, that instantiates this framework using Neural CDEs. **(3)** We design a simulation environment based on a clinical use case to study the effect and different drivers of informative sampling and use it to empirically demonstrate that our proposed method is able to correct for the resulting bias that existing methods can suffer from.

2. Problem Formalization: Data Structure and Informative Sampling Mechanism

This section describes and formalizes the problem of forecasting treatment outcomes in the presence of informative sampling. We investigate the assumptions required for and challenges inherent to tackling this problem in Section 3. We build on the exposition in Lin et al. (2004), who study longitudinal *outcome prediction* in the presence of informative sampling but do not explicitly consider estimating treatment effects, and build on ideas from Lok (2008); Seedat et al. (2022) who study forecasting treatment outcomes

[1] We discuss the related work more extensively in Appendix A.

in continuous time but do not consider informative sampling.

2.1. Problem structure: Complete versus observed data

Underlying complete data structure. We consider data collected over a period $[0, T]$ in which instances are characterized by a d-dimensional covariate path $X : [0, T] \to \mathbb{R}^d$ and a treatment path $A : [0, T] \to \{0, 1\}$ – which jumps to 1 only at time steps t when treatment is administered[2] – both of which possibly modulate an outcome process of interest $Y : [0, T] \to \mathbb{R}$. While we only observe the outcome Y associated with the treatment path A that was actually administered (sometimes also referred to as the *factual outcome*), we assume that any instance is characterised by a possibly infinite number of *potential outcomes* $Y_a : [0, T] \to \mathbb{R}$ associated with other feasible treatment paths a.

Observed data structure. Paths X, A and Y are only sampled (observed) at possibly irregular and discrete timepoints, such as scheduled check-ups or unscheduled appointments. Therefore, we additionally define a counting process $N : [0, T] \to \mathbb{N}_0$ recording the number of observations made by time t. This process jumps whenever a new observation is sampled[3], so that $dN(t) = 1$ if an instance is sampled at time t and $dN(t) = 0$ otherwise, where $dN(t) = N(t) - \lim_{s \uparrow t} N(s)$. As in Lin et al. (2004), for a variable $V(t)$, let $V^o(t) = V(\max s : 0 \leq t, dN(s) = 1)$ denote its most recent observation by time t, $\bar{V}^o(t) = \{V^o(s) : 0 \leq s \leq t\}$ its observed history by time t and $\bar{V}(t) = \{V(s) : 0 \leq s \leq t\}$ its full (yet possibly not observed) history by time t. Further, let $\bar{V}^o(t^-)$ and $\bar{V}(t^-)$ denote the same histories where the upper limit does *not* include t. Then, for a study following n instances until time T, we observe a dataset $\mathcal{D} = \{\mathcal{O}_i\}_{i=1}^n$ consisting of n i.i.d. copies of $\mathcal{O} = \mathcal{F}^o(T)$ where $\mathcal{F}^o(T) = (\bar{X}^o(T), \bar{N}(T), \bar{A}^o(T), \bar{Y}^o(T))$.

2.2. Distinguishing between different sampling patterns

In order to define what distinguishes the *informativeness* of different sampling patterns, we first need to introduce a conditional *intensity* $\lambda(t)$ which governs the observation process $N(t)$. In its most general form, adopting the notation of Lin et al. (2004), this can be defined through

$$\mathbb{P}(dN(t) = 1 | \bar{X}(T), \bar{N}(T), \bar{A}(T), \bar{Y}(T)) = \lambda(t)dt, \quad (1)$$

For notational convenience, we omit conditioning in $\lambda(t)$.

Using this definition, we can differentiate between different sampling mechanisms (Pullenayegum & Lim, 2016), giving rise to a classification similar to missingness mechanisms in static data (Rubin, 1976). This categorization is based on the causal role of the instance history in relation to the observation intensity (Fig. 1 shows a graphical overview):

- **Regular sampling (Fig. 1a).** Instances are observed at K regular (pre-determined) timesteps $\mathcal{T} = \{t_1, \ldots, t_K\}$, so that $\lambda(t)dt = 1$ if $t \in \mathcal{T}$ else $\lambda(t)dt = 0$. Most existing related work (e.g. Lim et al., 2018; Bica et al., 2019; Melnychuk et al., 2022) implicitly relies on this assumption.
- **Sampling completely at random (SCAR; Fig. 1b).** Instances are observed at completely random time-steps, with the intensity independent of all variables: $\lambda(t)dt = \mathbb{P}(dN(t) = 1 | \bar{X}(T), \bar{N}(T), \bar{A}(T), \bar{Y}(T)) = \mathbb{P}(dN(t) = 1)$. Recent work on treatment effect estimation from irregularly sampled data using neural differential equations (Gwak et al., 2020; Seedat et al., 2022; De Brouwer et al., 2022) is explicitly only equipped to handle this scenario.
- **Sampling at random (SAR; Fig. 1c).** Being observed at time t is independent of the (up until then unknown) outcome at time t given the observed history up to time t:

$$\begin{aligned}\lambda(t)dt &= \mathbb{P}(dN(t) = 1 | \bar{X}(T), \bar{N}(T), \bar{A}(T), \bar{Y}(T)) \\ &= \mathbb{P}(dN(t) = 1 | \bar{X}^o(t), \bar{N}(t^-), \bar{A}^o(t^-), \bar{Y}^o(t^-))\end{aligned}$$

  This work investigates the challenges of learning given SAR, which is considerably weaker than SCAR: it allows, for example, for patients to have more frequent visits due to past outcomes, administered treatments, or worsening symptoms (provided that these are recorded in X).

  We also consider a stricter variant, which we will refer to as the *strong SAR assumption*: here we assume that $\lambda(t)dt = \mathbb{P}(dN(t) = 1 | \bar{X}(T), \bar{N}(T), \bar{A}(T), \bar{Y}(T)) = \mathbb{P}(dN(t) = 1 | \bar{X}^o(t^-), \bar{N}(t^-), \bar{A}^o(t^-), \bar{Y}^o(t^-))$. This differs from the more general (weaker) SAR assumption above in that observing a patient at time t, i.e. $dN(t)$, cannot depend on the covariates $X^o(t)$ to be observed *at* time t. As we discuss in the next sections, the weaker SAR assumption already allows identification of treatment effects in our setting, while the strong SAR assumption can greatly simplify estimation of intensities.
- **Sampling not at random (SNAR).** The most general scenario is one where observing is *not* independent of future outcomes conditional on the observed history – i.e. $\lambda(t)dt \neq \mathbb{P}(dN(t) = 1 | \bar{X}^o(t), \bar{N}(t^-), \bar{A}^o(t^-), \bar{Y}^o(t^-))$. This would be the case, e.g., if patients chose to visit due to worsening symptoms that are *not recorded* in X and hence act as a latent cause of the intensity *and* outcome. In this scenario, outcomes cannot be consistently forecast unless further assumptions regarding the sampling or

[2] In this exposition, we assume that $A(t) = 1$ only at single time-steps where treatment is administered; for treatments that are administered over a time-period $[t_1, t_2]$ one could instead define a counting process that jumps whenever treatment status is *changed* as in Lok (2008); Seedat et al. (2022). Further, as noted in Seedat et al. (2022), this definition can be generalized to multiple treatments by assuming A to be a multivariate process.

[3] For ease of exposition, we assume that whenever an instance is observed at t, we record all of $X(t)$, $A(t)$ and $Y(t)$. Nevertheless, it would be possible to relax this by instead introducing separate counting processes for each variable or component thereof.

outcome-generating mechanism are made. Therefore, we rely on sampling at random in this work.

3. Forecasting Treatment Outcomes Under Informative Sampling: Goals, Assumptions and Inherent Challenges

3.1. Goal: Forecasting treatment outcomes

We aim to estimate **conditional average potential outcomes (CAPOs)** $\mu_{a,t}(\tau)$ at a future time $t+\tau, \tau \in (0, \tau_{\max}]$ (with $\tau_{\max} \leq T - t$):

$$\mu_{a,t}(\tau) = \mathbb{E}[Y_a(t+\tau)|\bar{X}^o(t), \bar{N}(t), \bar{A}^o(t), \bar{Y}^o(t)] \quad (2)$$

i.e., the instance's expected outcome under treatment plan a conditional on its full observed history $\mathcal{H}^o(t) = \{\bar{X}^o(t), \bar{N}(t), \bar{A}^o(t), \bar{Y}^o(t)\}$ up to time t. We only consider viable treatment plans a subject to $a(t^*) = A(t^*)$ for $t^* \leq t$ – i.e., those that do not modify the past, factual treatment history prior to the current time t. Such an estimate could be used in practice to decide between competing treatment plans based on expected outcome under either choice. In line with Gische et al. (2021), we purposefully use the term *forecasting* instead of *predicting* throughout to signify that we wish to give *causal interpretation* to the modeled effects of treatments. This is because, analogously to the standard static setting, unless we make further identifying assumptions, we can in general not assume that predictions based on expectations of the form $\mathbb{E}[Y^o(t+\tau)|A = a, \mathcal{H}^o(t)]$ are equal to forecasts based on expectations of the form $\mathbb{E}[Y_a(t+\tau)|\mathcal{H}^o(t)]$.

3.2. Identifying assumptions

To ensure *identification* of causal claims from observational data, we need to introduce additional assumptions. First, we make assumptions that correspond to adaptations of the standard *ignorability* assumptions (Rubin, 2005) from the standard static setting to our setting. To do so, we define *treatment propensities* for single time-steps $\pi(a(t)) = \mathbb{P}(A(t) = a(t)|\bar{X}(T), \bar{Y}(T), \bar{A}(T), \bar{N}(T))$ and entire trajectories $\pi_t(a) = \mathbb{P}(A = a|\mathcal{H}^o(t))$ given history until time t.

Assumption 3.1. Consistency. Given an observed treatment path A, we observe the outcome corresponding to the associated potential outcome: $Y^o(t) = Y_A(\max s : 0 \leq s \leq t, dN(t) = 1)$.

Assumption 3.2. Unconfoundedness. The treatment propensity $\pi(a, t)$ does not depend on future outcomes or unobserved information:

$$\begin{aligned}\pi(a(t)) &= \mathbb{P}(A(t) = a(t)|\bar{X}(T), \bar{Y}(T), \bar{A}(T), \bar{N}(T)) \\ &= \mathbb{P}(A(t) = a(t)|\bar{X}^o(t), \bar{N}(t), \bar{A}^o(t^-), \bar{Y}^o(t^-))\end{aligned}$$

Assumption 3.3. Overlap (Positivity for treatment). $0 < \mathbb{P}(A = a|\mathcal{H}^o(t)) < 1$, for all admissible treatment paths a and histories $\mathcal{H}^o(t)$ of interest for forecasting.

These assumptions are required regardless of the sampling mechanism. For regular sampling, these reduce to the sequential ignorability assumptions made in earlier work (e.g., Lim et al., 2018; Bica et al., 2019; Melnychuk et al., 2022).

On top of these ignorability assumptions, estimating causal effects under informative sampling requires making additional assumptions regarding the sampling mechansism (Robins et al., 1995). In contrast to existing work which *implicitly* assumed regular observations (e.g., Lim et al., 2018; Bica et al., 2019) or sampling completely at random (e.g., Seedat et al., 2022), we *explicitly* state our assumed observation process. Specifically, we rely on the weaker, previously introduced *sampling at random* (SAR) assumption:

Assumption 3.4. Sampling at random (SAR). The sampling intensity process does not depend on unobserved or future information, i.e., $\lambda(t) = \mathbb{P}(dN(t) = 1|\bar{X}^o(t), \bar{N}(t^-), \bar{A}^o(t^-), \bar{Y}^o(t^-))$.

Analogous to assumptions on treatment overlap, we assume the probability of observing at any point in time is bounded away from zero, for any history of interest for forecasting:

Assumption 3.5. Positivity of observation. $\mathbb{P}(dN(t+\tau) = 1|\mathcal{H}^o(t)) > 0$ for any $\tau \in (0, T-t]$ and history $\mathcal{H}^o(t)$ of interest.

Finally, we assume that all treatment events are observed. In most applications, this is a natural assumption (e.g., if treatments are administered at a hospital). It is generally not possible to estimate treatment effects from observed outcomes without knowing which treatments were administered, unless further assumptions are made (Kennedy, 2020).

Assumption 3.6. Observed treatments. All treatments are observed, i.e. $\mathbb{P}(A(t) = 1|dN(t) = 0) = 0$.

The identifying assumptions discussed above can equivalently be expressed as a generative model, determining the temporal ordering of realizations of the different observed variables[4]. In particular, at each time t, the visit decision $dN(t)$ is realized first, which can depend on observed histories $\mathcal{H}^o(t^-)$ and covariates to be observed $X^o(t)$ (SAR) or on $\mathcal{H}^o(t^-)$ only (strong SAR); the former implies a setting where e.g. patients present themselves for an appointment due to worsening symptoms while the latter allows only scheduling of future appointments due to symptoms already observed earlier. If $dN(t) = 1$, then covariates $X^o(t)$ are

[4] In principle, other generative models could be assumed as long as sufficient exclusion restrictions between observation-/treatment-generating processes and outcome-generating processes are made. For example, the visit process can depend on future treatments if such treatments are pre-scheduled.

first observed, treatment $A(t)$ is then determined based only on observed information ($\mathcal{H}^o(t^-)$, $N(t)$ & $X^o(t)$) and, finally, the outcome is realized and observed as $Y^o(t)$.

3.3. What makes learning CAPOs from observational data challenging?

If we had access to the complete data structure with all potential outcomes $Y_a(t)$, learning an estimate $\hat{\mu}_{a,t}(\tau; \mathcal{H}^o(t))$ for the CAPOs with fixed a would be a standard ML problem: we would search a hypothesis function in some hypothesis class $\mathcal{F}$ that minimizes the expected (oracle) risk R^*, i.e.,

$$\hat{\mu}_{a,t}(\tau; \mathcal{H}^o(t)) \in \arg\min_{f_{a,\tau} \in \mathcal{F}} R^*(f_{a,\tau}) \tag{3}$$

where, for some loss function ℓ, and using the shorthands $t' = t + \tau$ and $h_t = \mathcal{H}^o(t)$

$$R^*(f_{a,\tau}) = \mathbb{E}\left[\int_0^T \int_t^{\tau_{\max}} \ell\left(Y_a(t'), f_{a,\tau}(h_t)\right) d\tau dt\right] =$$

$$\int_0^T \int_0^{\tau_{\max}} \iint \ell\left(y_a(t'), f_{a,\tau}(h_t)\right) dP(y_a(t')|h_t) dP(h_t) d\tau dt$$

However, as previously discussed, in reality we only have access to observational data in which patients are (i) incompletely, irregularly, and informatively observed and (ii) characterized by only a single *factual* outcome corresponding to the treatment actually received. If we were to learn a standard ML predictor from this observed data, we would instead be optimizing the observational risk $R^{obs}(h_{a,\tau}) =$

$$\mathbb{E}\left[\int_0^T \int_0^{\tau_{\max}} \mathbb{1}\{A=a\} dN(t') \ell\left(Y_a(t'), f_{a,\tau}(h_t)\right) d\tau dt\right] =$$

$$\int_0^T \int_t^{\tau_{\max}} \iint \ell(y_a(t'), f_{a,\tau}(h_t)) \pi_t(a) \lambda_t(t')$$

$$dP(y_a(t')|h_t) dP(h_t) d\tau dt$$

Thus, unless the τ-step ahead intensity $\lambda_t(t')$, defined through $\lambda_t(t')d\tau = \mathbb{P}(dN(t+\tau)=1|\mathcal{H}^o(t) \cup \bar{A}(t+\tau^-))$, and treatment propensity $\pi_t(a)$ are constant across patient histories, the minimizers of R^* and R^{obs} will in general be different. Intuitively, this is because the distribution of patient characteristics in the observed data can differ from the distribution in the underlying unobserved complete distribution, both due to informative sampling and treatment selection. Thus, the challenge we are facing here is one of *covariate shift* between the training data and hypothetical test data.

Covariate shift and its potential remedies have been studied in much depth in the recent ML literature (see e.g. Redko et al. (2020); Farahani et al. (2021) for recent overviews). Here, we explore the use of one of the oldest and most well-established solutions: *importance weighting* (Shimodaira, 2000). That is, as further discussed in the next section, we propose to minimize a weighted observational risk

$$R^w(f_{a,\tau}) = \mathbb{E}\left[\int_0^T \int_0^{\tau_{\max}} w_{a,\tau} \ell(f_{a,\tau}) d\tau dt\right] \tag{4}$$

with $\ell(f_{a,\tau}) = \mathbb{1}\{A=a\} dN(t') \ell\left(Y_a(t'), f_{a,\tau}(h_t)\right)$. For oracle importance weights $w^*_{a,\tau} = \frac{1}{\pi(a)\lambda_t(t')}$ it is easy to see that $R^{obs,w^*}(f_{a,\tau}) = R^*(f_{a,\tau})$.

4. Learning to Forecast Treatment Outcomes Under Informative Sampling

This section presents a methodology for learning to forecast treatment outcomes under informative sampling. Section 4.1 presents a general framework that is compatible with any ML algorithm capable of predicting outcomes over time. Section 4.2 instantiates this framework using Neural CDEs.

4.1. Learning To Forecast Using Inverse Intensity Weights

The analysis presented in Section 3.3 allows straightforward construction of a framework for learning to forecast treatment outcomes from informatively sampled (SAR) data. Given an ML algorithm $\mathcal{A}$ that can issue continuous-time predictions using irregularly sampled data, one simply needs to fit $\mathcal{A}$ on the observed data while providing appropriate importance weights w. As the true weights will generally be unknown in practice, one might have to use $\mathcal{A}$ to also learn (i) observation intensities and (ii) treatment propensities to gain access to estimates of the true importance weights. As we discuss for a specific example in Section 4.2.2, one could learn such weights either in a pre-processing step or in an end-to-end fashion.

When learning intensity weights, it becomes important whether one makes the general SAR or the strong SAR assumption: under the strong SAR assumption, learning $\lambda_t(t')$ comes down to the easier task of estimating $\mathbb{P}(dN(t') = 1|\mathcal{H}^o(t))$ directly, where $t' = t + \tau$. Under the more general SAR assumption, one needs to model $dN(t')$ and $X^o(t')$ *jointly* as a *marked* point process to learn the distributions $P(dN(t'), X^o(t')|\mathcal{H}^o(t)) = P(dN(t')|X^o(t'), \mathcal{H}^o(t)) P(X^o(t')|\mathcal{H}^o(t))$, where $P(X^o(t')|\mathcal{H}^o(t))$ could be a high-dimensional continuous density. In the remainder, we therefore restrict ourselves to the strong SAR setting – allowing us to highlight the challenges arising when learning under some form of informative sampling. It would be an interesting next step to incorporate some of the recent work on Neural Temporal Point Processes (see e.g. Shchur et al., 2021) to enable learning under more complex dependencies.

In the following, we discuss two possible implementations

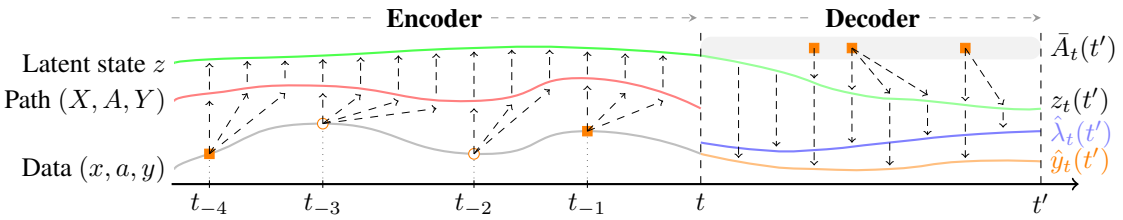


Figure 2: **TESAR-CDE: Adapting TE-CDE for learning given SAR.** The history of observations (○) and treatments (■) up to time t is first encoded as a continuous latent path $z(t)$. Based on a future treatment plan $\bar{A}_t(t')$, the decoder then forecasts a future latent path $z_t(t')$, with $t' = t + \tau$. In contrast to TE-CDE, TESAR-CDE (1) uses the latent path $z_t(t')$ to forecast both the outcome $\hat{y}_t(t')$ and intensity $\hat{\lambda}_t(t')$, and (2) uses the intensity to weight the outcome loss using $\mathcal{L}^{\text{WMSE}}$.

of this framework using Neural CDEs by extending the methodology for learning continuous time treatment effects of Seedat et al. (2022) (which originally did not correct for informative sampling). Nevertheless, the approach discussed above is more general and could be applied to any ML model that can predict $Y(t)$.

4.2. TESAR-CDE: Forecasting with Intensity-weighted Neural CDEs

This section presents TESAR-CDE, a specific implementation of the framework discussed above by extending TE-CDE (Seedat et al., 2022) to account for informative sampling (SAR), see Fig. 2 for a graphical overview. In the remainder of this work, we focus on the special case where complete treatment plans A are *randomly* assigned and *fixed* at time $t = 0$; such a situation commonly arises in practice, e.g., in a clinical trial with a dynamic observation plan (Lin et al., 2004; Bužková & Lumley, 2009). This allows to *single out* the challenges arising solely due to the presence of informative observations. Moreover, this allows us to highlight that the forecasting problem remains challenging even in the absence of all treatment selection bias (the main challenge addressed in related work). Nevertheless, if required, any existing method equipped to deal with outcome-treatment confounding – e.g., using importance weighting (Lim et al., 2018) or adversarial training (Seedat et al., 2022) – could simply be combined with the inverse intensity weighting approaches we discuss and test below.

4.2.1. Background: TE-CDE

Treatment Effect Neural Controlled Differential Equation (TE-CDE; Seedat et al., 2022) is a recently proposed model for forecasting treatment effects from irregularly sampled data. TE-CDE views observations as samples from an underlying continuous-time process and uses Neural CDEs (Kidger et al., 2020) to learn this latent trajectory. First, an encoder learns a latent path $z(t)$ as the solution of a CDE:

$$z(t_0) = g(X(t_0), A(t_0), Y(t_0)),$$

$$z(t) = z(t_0) + \int_{t_0}^{t} f_\theta(z(s)) \frac{d(X(s), A(s), Y(s))}{ds} ds$$

with g and f_θ neural networks. This is achieved by solving the above initial value problem (IVP), $\forall s \in [t_0, t]$:

$$z(t) = \texttt{ODESolve}(f_\theta, z(t_0), \bar{X}(t), \bar{A}(t), \bar{Y}(t))$$

using a numerical ODE solver (Kidger et al., 2020). The decoder forecasts the future latent path $z_t(t')$ by solving a second IVP given the future treatment plan $\bar{A}_t(t')$:

$$z_t(t') = \texttt{ODESolve}(f_\phi, z(t), \bar{A}_t(t')),$$

with decoder network f_ϕ and $t' = t + \tau$. A final network f_ψ maps the latent path $z_t(t')$ to the outcome $\hat{y}_t(t') = f_\psi(z_t(t'))$. Fig. 3a shows the complete architecture.

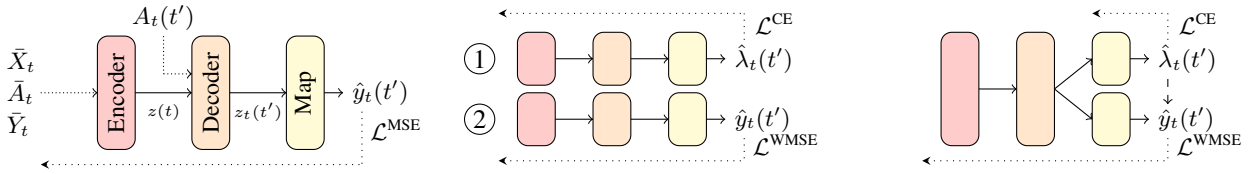
The entire model (g, f_θ, f_ϕ and f_ψ) is trained by optimizing the mean squared error (MSE) of the predicted outcome:

$$\mathcal{L}_i^{\text{MSE}} = \int_0^T \int_0^{\tau_{\max}} dN_i(t') \left(y_i(t') - \hat{y}_{i,t}(t')\right)^2 dt d\tau.$$

This way, the mean squared error is calculated using the observed outcomes in the considered forecasting horizon $(0, \tau_{\max}]$, for each timestep $t \in [0, T]$. To account for bias resulting from time-dependent confounding, TE-CDE also uses domain adversarial training to learn a treatment-invariant representation. However, as discussed above, we focus on unconfounded settings in the remainder of this work and therefore do not include this, though it is straightforward to add it in settings where needed.

4.2.2. TESAR-CDE: Learning to forecast with an inverse-intensity weighted loss

In this section, we instantiate our previously proposed framework using Neural CDEs, resulting in TESAR-CDE, Treatment Effects given Sampling At Random using Neural CDEs. Essentially, we extend TE-CDE for learning under informative sampling. Given (estimated) intensities $\hat{\lambda}_{i,t}(t')$, adapting TE-CDE's outcome loss to adjust for informative



(a) TE-CDE (b) TESAR-CDE (Two-step) (c) TESAR-CDE (Multitask)

Figure 3: **Comparing TESAR-CDE to TE-CDE.** We show TE-CDE and our proposed alternative, TESAR-CDE, in its two-step and multitask configuration. Arrows indicate the input (⇢), forward pass (⟶) and backpropagation (⇠). The multitask model uses the intensity loss only to train the intensity map, but not the shared encoder or decoder. The dashed arrow (- -›) indicates that the intensities are used as weights λ_t^{-1} in $\mathcal{L}^{\text{WMSE}}$, but not backpropagated as part of this loss.

observations is straightforward: the inverse of these estimated intensities can be used as importance weights to train TE-CDE for outcome prediction using a *weighted* MSE:

$$\mathcal{L}_i^{\text{WMSE}} = \int_0^T \int_0^{\tau_{\max}} dN_i(t') \, \frac{\left(y_i(t') - \hat{y}_{i,t}(t')\right)^2}{\hat{\lambda}_{i,t}(t')} \, dt d\tau.$$

We propose a multi-stage or end-to-end version of TESAR-CDE (Fig. 3 compares the proposed architectures with TE-CDE). Both predict the intensity from z_t as $\hat{\lambda}_{i,t}(t') = f_\psi^\lambda(z_t(t'))$. We assume there is a minimal sampling interval dt; e.g., doctors might not measure covariates more than once per hour. Let $dN_{i,t}(t') = 1$ if instance i was observed in interval $(t, t']$, 0 otherwise. The intensity $\lambda_{i,t}(t')$ can then be approximated by minimizing the cross-entropy

$$\mathcal{L}_i^{\text{CE}} = -\sum_{t=0}^{T} \sum_{\tau=0}^{\tau_{\max}} \Big[dN_{i,t}(t') \log\Big(\hat{\lambda}_{i,t}(t')\Big) + (1 - dN_{i,t}(t')) \log\Big(1 - \hat{\lambda}_{i,t}(t')\Big) \Big],$$

where $t' = t + \tau$. For applications where no minimal time step dt exists, neural point processes can be used to learn the intensity in continuous time (see e.g. Shchur et al., 2021).

The **two-step** procedure consists of two TE-CDE style models that are trained sequentially. A first model predicts the intensity $\lambda_{i,t}(t')$; a second model uses the inverse of these intensities $\lambda_{i,t}(t')^{-1}$ as weights in its weighted MSE loss. Alternatively, we can combine both tasks in a **multitask** framework to predict both intensities and outcomes:

$$\mathcal{L}^{\text{MT}} = (1 - \alpha)\mathcal{L}_i^{\text{WMSE}} + \alpha \mathcal{L}_i^{\text{CE}}, \tag{5}$$

with hyperparameter α balancing the importance of the two terms. The intensity loss only optimizes the intensity map f_ψ^λ; the weighted MSE is used to optimize the rest of the network $(g, f_\theta, f_\phi, f_\psi^y)$. Moreover, similar to Hassanpour & Greiner (2019)'s architecture for learning importance weights to correct for treatment-outcome confounding, we do not backpropagate with respect to the intensity weights in the weighted MSE for outcome prediction, as this could bias the network to predict small weights (i.e. large intensities) in order to minimize the weighted MSE.

The potential benefits of the multitask framework are threefold. First, learning a shared representation z_t to predict both outcome and intensity results in fewer parameters. Second, it requires only training one network and one call to the ODE solver per iteration, resulting in computational speedups. Third, to reduce the variability due to importance weighting, we only optimize the shared representation of the multitask model for outcome prediction. For bias correction using importance weighting, the shared representation does *not* need to be a sufficient statistic for predicting the intensity. This is because we only need to care about the non-uniformity in observation intensity insofar as it is related to the outcome. The reason for this is conceptually identical to why one should not include predictors of treatment only (a.k.a. instruments) in a propensity score (VanderWeele, 2019) and why sufficient dimensionality reduction before importance weighting is recommended in general applications with covariate shift (Maia Polo & Vicente, 2022): importance weighting generally only needs to adjust for shifts in variables that are themselves predictors of the outcome. Our multitask learner implicitly enforces this by optimizing the shared representation based on the outcome loss only. The two-step and multitask architectures are illustrated in Figs. 3b and 3c. Appendices C and D provide more details on the training procedure and implementation.

5. Results

To assess the impact of informative sampling, we propose a novel simulation environment that allows us to control the level of informativeness and assess its effect on the resulting model's performance. Our simulation is inspired by real-world randomized controlled trials that compared treatment regimes in the context of lung cancer (Furuse et al., 1999; Aupérin et al., 2010; Curran Jr et al., 2011). Given the patient's history, our goal is to forecast the patient's tumor size for a potential future treatment plan. Our code is available at https://github.com/toonvds/TESAR-CDE.

5.1. Simulation: lung cancer treatment

Following existing work (e.g., Melnychuk et al., 2022; Seedat et al., 2022), we simulate data based on the tumor growth model of Geng et al. (2017). To analyze the effect of informative sampling, we combine this tumor growth model with a sampling mechanism in which the degree of informativeness can be controlled. We refer to Appendix E for more detailed information and visualizations.

Tumor growth simulation. We simulate tumor growth based on a pharmacokinetic-pharmacodynamic model of Geng et al. (2017). This model simulates the outcome, tumor volume $Y(t)$, based on the historical tumor volume, tumor growth, chemotherapy, and radiotherapy:

$$\frac{dY(t)}{dt} = \Big[1 + \overbrace{\rho \log\left(\frac{K}{Y(t)}\right)}^{\text{Tumor growth}} - \overbrace{\beta_c C(t)}^{\text{Chemotherapy}} - \underbrace{\left(\alpha_r d(t) + \beta_r d(t)^2\right)}_{\text{Radiotherapy}} + \underbrace{\epsilon(t)}_{\text{Noise}}\Big] Y(t), \tag{6}$$

with $K, \rho, \beta_c, \alpha_r, \beta_r, \epsilon_t$ sampled following Geng et al. (2017); $C(t)$ and $d(t)$ are set following existing work (Lim et al., 2018; Bica et al., 2019; Seedat et al., 2022).

Treatment plans. We differentiate between a sequential and concurrent treatment regime (Curran Jr et al., 2011). In the sequential treatment arm, patients receive weekly chemotherapy for five weeks, followed by weekly radiotherapy for five weeks. In the concurrent treatment arm, patients biweekly receive both chemotherapy and radiotherapy for ten weeks. Patients are randomly divided among the two treatment arms based on a Bernoulli distribution with probability $p = 0.5$. This way, there is no confounding: treatment assignments are random and do not change during the trial.

Observation process. We observe patients based on a patient-specific, history-dependent intensity process. This is achieved by simulating each patient's observation process with intensity $\lambda_i(t)$ in which γ controls the informativeness:

$$\lambda_i(t) = \sigma\left[\gamma\left(\frac{\bar{D}_i(t-)}{D_{\text{max}}} - \frac{1}{2}\right)\right], \tag{7}$$

where σ denotes the sigmoid function. $D_{\text{max}} = 13$cm denotes the maximal tumor diameter and $\bar{D}(t-)$ is the average tumor diameter over the past 15 days. By simulating the observation process in this way, we can control the degree of informativeness: $\gamma = 0$ implies SCAR as $\lambda_i(t) = 0.5$, while $\gamma > 0$ implies SAR with a larger γ implying more informativeness or dependence between the tumor size and intensity. As γ increases, patients with larger tumors are more likely to be observed, those with smaller tumors less.

Experimental setup. We assume treatments are always observed, as these are planned in advance and administered in the hospital. Nevertheless, at treatment time, we do not necessarily observe the patient's tumor size, e.g., because observing tumor size requires an invasive procedure separate from the treatment. For the test set, we observe all information at daily intervals. This idealized setup allows us to assess whether the model is able to learn the underlying distribution, as opposed to fitting the observed samples. Similarly, the test data also contains the potential outcomes for both treatment arms. For each experiment, we generate a training set with 200 patients, validation set with 50 patients, and test set with 200 patients, all over a period of 120 days.

5.2. Empirical results

This section presents the empirical results using the experimental setup described above. More specifically, we aim to answer three questions: (1) What is the impact of informative sampling?; (2) What is the impact of observation scarcity?; and (3) When does informativeness matter? In Appendix F, we present additional experiments to evaluate TESAR-CDE's intensity prediction and analyze the sensitivity of the multitask configuration to hyperparameter α.

What is the impact of informative sampling? We compare the different models at varying levels of informativeness in Fig. 4. Increasing γ makes the observation process more informative by having patients with a large tumor visit more and patients with small tumors less (Eq. (7)). The left side of Fig. 4 shows the RMSE of the different models at increasing informativeness. As sampling becomes more informative, TE-CDE's performance deteriorates and is outperformed by the proposed TESAR-CDE. The multitask variant in particular is robust to high levels of informativeness, achieving the lowest RMSE overall. The two-step variant generally outperforms TE-CDE, but performs worse for very high informativeness. As a high γ results in very low intensities for some patients, the importance weights of these observations induce large variance and worse generalization properties in the weighted loss. This phenomenon is a well-known issue in importance weighting more generally (Cortes et al., 2010). The right of Fig. 4 shows the performance at different forecasting horizons τ ranging from one to five days at a fixed informativeness of $\gamma = 6$. Both TESAR-CDE variants outperform the standard TE-CDE over all horizons, with the multitask variant again performing the best overall.

What is the impact of observation scarcity? We analyze the influence of less frequent sampling across all patients. We simulate lower overall sampling by scaling all intensities as $\lambda'(t) = \frac{\lambda(t)}{S_\lambda}$ with $S_\lambda \in \{1, 2, 3, 4\}$. Fig. 5 shows the impact of increasing scarcity on the resulting RMSE. As expected, all models perform worse with less frequent sampling. The two-step TESAR-CDE performs worse than the baseline TE-CDE as scarcity increases, while the multitask TESAR-CDE is again the best performing model overall.

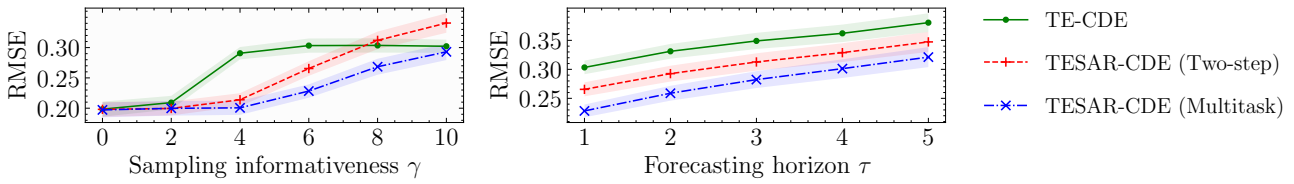


Figure 4: **Results for varying informativeness γ and different forecasting horizons τ.** We show the RMSE $\pm$ SE over ten runs. **(Left)** RMSE for increasing levels of informativeness γ, keeping the forecasting horizon fixed at $\tau = 1$. **(Right)** RMSE for an increasing forecasting horizon τ up to five days, keeping informativeness fixed at $\gamma = 6$.

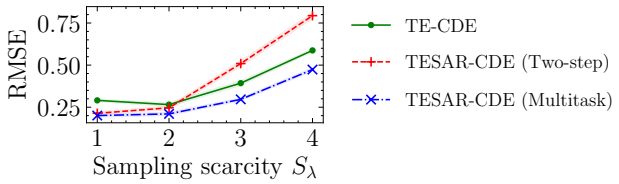


Figure 5: **Observation scarcity.** We show the RMSE $\pm$ SE over ten runs at increasing observation scarcity S_λ for fixed informativeness ($\gamma = 4$) and forecasting horizon ($\tau = 1$).

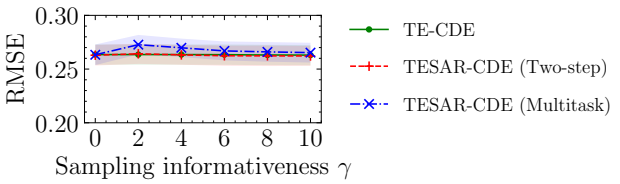


Figure 6: **Outcome-unrelated sampling.** We show the RMSE $\pm$ SE over ten runs for a sampling mechanism unrelated to the outcome $Y(t)$ in function of informativeness γ.

This result indicates that the parameter efficiency of the multitask model can be helpful when sampling is scarce.

When does informativeness matter? The previous experiments analyzed informative observation processes where the observation intensity $\lambda(t)$ was directly related to the outcome $Y(t)$. Next, we analyze a special case of SAR where the intensity depends on information that is *completely unrelated* to the outcome or treatments. This extreme scenario mimics a situation in which there are patients that visit often for reasons unrelated to underlying symptoms or outcome, e.g. when they suffer from hypochondria. We examine this using an intensity that depends on covariates x^λ: $\lambda_i(t) = \sigma\left(\gamma \sum_j^d (c_j x_j)\right)$, where we include $d = 10$ static variables $x^{(i)}$ for each patient, with each $x_j \sim \mathcal{N}(0, 1)$ influencing the intensity through a coefficient $c_j \sim \mathcal{U}(-1, 1)$, but not affecting the patient in any other way. Fig. 6 shows the prediction error of the different models for an increasing γ, averaged over $\tau \in \{1, \ldots, 5\}$. In this scenario, the sampling mechanism does *not* significantly affect performance, with all models having similar performance. If anything, TESAR-CDE (Multitask) performs slightly worse, possibly due to importance weighting being unnecessary and adding variance in this scenario. This result indicates that informative sampling may only matter when it depends on factors influencing *both* observation intensity and outcome.

6. Conclusion

This work analyzed and formalized an essential challenge in learning to forecast treatment outcomes over time from observational data: the presence of informative sampling. We differentiated between different sampling mechanisms depending on the causal role of the observation intensity. This categorization allowed us to identify an overlooked, yet common setting in which observations are sampled irregularly over time with the observation intensity depending on the history of the instance's covariates, outcome, and/or treatments. We formalized learning in this context and analyzed it as a problem of covariate shift. Based on this, we proposed a general framework for learning under informative sampling and a novel method, TESAR-CDE, that instatiates this framework using Neural CDEs. Empirical results demonstrate the improved performance of TESAR-CDE over the current state-of-the-art when sampling is informative.

Accounting for informative sampling when learning to forecast treatment outcomes relies on strong identification assumptions regarding both the treatment assignment and sampling mechanism. As these assumptions are untestable, we need to rely on domain expertise to judge their plausibility in practical applications. For example, the sampling at random assumption would be violated if the observation intensity is affected by an unobserved cause of the outcome. While beyond the scope of the current work, we believe that exploring learning under violations of these assumptions is an important and fruitful area for future research – similar to the rich lines of work exploring estimation of treatment effects with hidden confounders or missing treatment information.

Acknowledgements

We would like to thank Nabeel Seedat and the anonymous reviewers for insightful comments and discussions on earlier drafts of this paper. TV is supported by the Research Foundation – Flanders (FWO PhD Fellowship 11I7322N). AC gratefully acknowledges funding from AstraZeneca.

References

Alaa, A. M. and Van Der Schaar, M. Balancing suspense and surprise: Timely decision making with endogenous information acquisition. *Advances in Neural Information Processing Systems*, 29, 2016.

Alaa, A. M., Hu, S., and Schaar, M. Learning from clinical judgments: Semi-markov-modulated marked hawkes processes for risk prognosis. In *International Conference on Machine Learning*, pp. 60–69. PMLR, 2017.

Aupérin, A., Le Péchoux, C., Rolland, E., Curran, W. J., Furuse, K., Fournel, P., Belderbos, J., Clamon, G., Ulutin, H. C., Paulus, R., et al. Meta-analysis of concomitant versus sequential radiochemotherapy in locally advanced non-small-cell lung cancer. *Database of Abstracts of Reviews of Effects (DARE): Quality-assessed Reviews [Internet]*, 2010.

Bellot, A. and Van Der Schaar, M. Policy analysis using synthetic controls in continuous-time. In *International Conference on Machine Learning*, pp. 759–768. PMLR, 2021.

Berrevoets, J., Curth, A., Bica, I., McKinney, E., and van der Schaar, M. Disentangled counterfactual recurrent networks for treatment effect inference over time. *arXiv preprint arXiv:2112.03811*, 2021.

Berrevoets, J., Imrie, F., Kyono, T., Jordon, J., and van der Schaar, M. To impute or not to impute? missing data in treatment effect estimation. In *International Conference on Artificial Intelligence and Statistics*, pp. 3568–3590. PMLR, 2023.

Bica, I., Alaa, A. M., Jordon, J., and van der Schaar, M. Estimating counterfactual treatment outcomes over time through adversarially balanced representations. In *International Conference on Learning Representations*, 2019.

Biewald, L. Experiment tracking with weights and biases, 2020. URL https://www.wandb.com/. Software available from wandb.com.

Bužková, P. and Lumley, T. Semiparametric modeling of repeated measurements under outcome-dependent follow-up. *Statistics in Medicine*, 28(6):987–1003, 2009.

Che, Z., Purushotham, S., Cho, K., Sontag, D., and Liu, Y. Recurrent neural networks for multivariate time series with missing values. *Scientific reports*, 8(1):1–12, 2018.

Chen, R. T., Rubanova, Y., Bettencourt, J., and Duvenaud, D. K. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.

Clithero, J. A. Response times in economics: Looking through the lens of sequential sampling models. *Journal of Economic Psychology*, 69:61–86, 2018.

Cole, S. R. and Hernán, M. A. Constructing inverse probability weights for marginal structural models. *American journal of epidemiology*, 168(6):656–664, 2008.

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

Curran Jr, W. J., Paulus, R., Langer, C. J., Komaki, R., Lee, J. S., Hauser, S., Movsas, B., Wasserman, T., Rosenthal, S. A., Gore, E., et al. Sequential vs concurrent chemoradiation for stage iii non–small cell lung cancer: randomized phase iii trial rtog 9410. *Journal of the National Cancer Institute*, 103(19):1452–1460, 2011.

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

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

Curth, A. and van der Schaar, M. Understanding the impact of competing events on heterogeneous treatment effect estimation from time-to-event data. In *International Conference on Artificial Intelligence and Statistics*, pp. 7961–7980. PMLR, 2023.

Curth, A., Lee, C., and van der Schaar, M. Survite: Learning heterogeneous treatment effects from time-to-event data. *Advances in Neural Information Processing Systems*, 34: 26740–26753, 2021.

De Brouwer, E., Gonzalez, J., and Hyland, S. Predicting the impact of treatments over time with uncertainty aware neural differential equations. In *International Conference on Artificial Intelligence and Statistics*, pp. 4705–4722. PMLR, 2022.

Del Moral, P. and Murray, L. M. Sequential monte carlo with highly informative observations. *SIAM/ASA Journal on Uncertainty Quantification*, 3(1):969–997, 2015.

Farahani, A., Voghoei, S., Rasheed, K., and Arabnia, H. R. A brief review of domain adaptation. *Advances in data science and information engineering*, pp. 877–894, 2021.

Farzanfar, D., Abumuamar, A., Kim, J., Sirotich, E., Wang, Y., and Pullenayegum, E. Longitudinal studies that use data collected as part of usual care risk reporting biased results: a systematic review. *BMC Medical Research*

Methodology, 17(1):1–12, 2017.

Furuse, K., Fukuoka, M., Kawahara, M., Nishikawa, H., Takada, Y., Kudoh, S., Katagami, N., and Ariyoshi, Y. Phase iii study of concurrent versus sequential thoracic radiotherapy in combination with mitomycin, vindesine, and cisplatin in unresectable stage iii non–small-cell lung cancer. *Journal of Clinical Oncology*, 17(9):2692–2692, 1999.

Gasparini, A., Abrams, K. R., Barrett, J. K., Major, R. W., Sweeting, M. J., Brunskill, N. J., and Crowther, M. J. Mixed-effects models for health care longitudinal data with an informative visiting process: A monte carlo simulation study. *Statistica Neerlandica*, 74(1):5–23, 2020.

Geng, C., Paganetti, H., and Grassberger, C. Prediction of treatment response for combined chemo-and radiation therapy for non-small cell lung cancer patients using a bio-mathematical model. *Scientific reports*, 7(1):1–12, 2017.

Gische, C., West, S. G., and Voelkle, M. C. Forecasting causal effects of interventions versus predicting future outcomes. *Structural Equation Modeling: A Multidisciplinary Journal*, 28(3):475–492, 2021.

Goldstein, B. A., Bhavsar, N. A., Phelan, M., and Pencina, M. J. Controlling for informed presence bias due to the number of health encounters in an electronic health record. *American journal of epidemiology*, 184(11):847–855, 2016.

Gwak, D., Sim, G., Poli, M., Massaroli, S., Choo, J., and Choi, E. Neural ordinary differential equations for intervention modeling. *arXiv preprint arXiv:2010.08304*, 2020.

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

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

Hernán, M. A., McAdams, M., McGrath, N., Lanoy, E., and Costagliola, D. Observation plans in longitudinal studies with time-varying treatments. *Statistical methods in medical research*, 18(1):27–52, 2009.

Jeanselme, V., Martin, G., Peek, N., Sperrin, M., Tom, B., and Barrett, J. Deepjoint: Robust survival modelling under clinical presence shift. *arXiv preprint arXiv:2205.13481*, 2022.

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

Kennedy, E. H. Efficient nonparametric causal inference with missing exposure information. *The international journal of biostatistics*, 16(1), 2020.

Kidger, P., Morrill, J., Foster, J., and Lyons, T. Neural controlled differential equations for irregular time series. *Advances in Neural Information Processing Systems*, 33: 6696–6707, 2020.

Li, R., Shahn, Z., Li, J., Lu, M., Chakraborty, P., Sow, D., Ghalwash, M., and Lehman, L.-w. H. G-net: a deep learning approach to g-computation for counterfactual outcome prediction under dynamic treatment regimes. In *Machine Learning for Health*, 2021.

Lim, B., Alaa, A., and van der Schaar, M. Forecasting treatment responses over time using recurrent marginal structural networks. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/56e6a93212e4482d99c84a639d254b67-Paper.pdf.

Lin, H., Scharfstein, D. O., and Rosenheck, R. A. Analysis of longitudinal data with irregular, outcome-dependent follow-up. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(3):791–813, 2004.

Little, R. J. and Rubin, D. B. *Statistical analysis with missing data*, volume 793. John Wiley & Sons, 2019.

Liu, L., Huang, X., and O'Quigley, J. Analysis of longitudinal data in the presence of informative observational times and a dependent terminal event, with application to medical cost data. *Biometrics*, 64(3):950–958, 2008.

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

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

Mayer, I., Sverdrup, E., Gauss, T., Moyer, J.-D., Wager, S., and Josse, J. Doubly robust treatment effect estimation with missing attributes. *Annals of Applied Statistics*, 14 (3):1409–1431, 2020.

McCulloch, C. E., Neuhaus, J. M., and Olin, R. L. Biased and unbiased estimation in longitudinal studies with informative visit processes. *Biometrics*, 72(4):1315–1324, 2016.

Melnychuk, V., Frauen, D., and Feuerriegel, S. Causal transformer for estimating counterfactual outcomes. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S. (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 15293–15329. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/melnychuk22a.html.

Morrill, J., Kidger, P., Yang, L., and Lyons, T. Neural controlled differential equations for online prediction tasks. *arXiv preprint arXiv:2106.11028*, 2021.

Pullenayegum, E. M. and Lim, L. S. Longitudinal data subject to irregular observation: A review of methods with a focus on visit processes, assumptions, and study design. *Statistical methods in medical research*, 25(6): 2992–3014, 2016.

Qian, Z., Zhang, Y., Bica, I., Wood, A., and van der Schaar, M. Synctwin: Treatment effect estimation with longitudinal outcomes. *Advances in Neural Information Processing Systems*, 34:3178–3190, 2021.

Redko, I., Morvant, E., Habrard, A., Sebban, M., and Bennani, Y. A survey on domain adaptation theory: learning bounds and theoretical guarantees. *arXiv preprint arXiv:2004.11829*, 2020.

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

Robins, J. M. Causal inference from complex longitudinal data. In *Latent variable modeling and applications to causality*, pp. 69–117. Springer, 1997.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the american statistical association*, 90(429):106–121, 1995.

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

Roy, J., Lum, K. J., and Daniels, M. J. A bayesian nonparametric approach to marginal structural models for point treatments and a continuous or survival outcome. *Biostatistics*, 18(1):32–47, 2017.

Rubanova, Y., Chen, R. T., and Duvenaud, D. K. Latent ordinary differential equations for irregularly-sampled time series. *Advances in neural information processing systems*, 32, 2019.

Rubin, D. B. Inference and missing data. *Biometrika*, 63(3): 581–592, 1976.

Rubin, D. B. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

Schulam, P. and Saria, S. Reliable decision support using counterfactual models. *Advances in neural information processing systems*, 30, 2017.

Seedat, N., Imrie, F., Bellot, A., Qian, Z., and van der Schaar, M. Continuous-time modeling of counterfactual outcomes using neural controlled differential equations. In *International Conference on Machine Learning*, pp. 19497–19521. PMLR, 2022.

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

Shchur, O., Türkmen, A. C., Januschowski, T., and Günnemann, S. Neural temporal point processes: A review. *arXiv preprint arXiv:2104.03528*, 2021.

Shimodaira, H. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000.

Soleimani, H., Subbaswamy, A., and Saria, S. Treatment-response models for counterfactual reasoning with continuous-time, continuous-valued interventions. *arXiv preprint arXiv:1704.02038*, 2017.

Vanderschueren, T., Boute, R., Verdonck, T., Baesens, B., and Verbeke, W. Optimizing the preventive maintenance frequency with causal machine learning. *International Journal of Production Economics*, 258:108798, 2023.

VanderWeele, T. J. Principles of confounder selection. *European journal of epidemiology*, 34:211–219, 2019.

Xu, Y., Xu, Y., and Saria, S. A bayesian nonparametric approach for estimating individualized treatment-response curves. In *Machine learning for healthcare conference*, pp. 282–300. PMLR, 2016.

Yoon, J., Zame, W. R., and Van Der Schaar, M. Deep sensing: Active sensing using multi-directional recurrent neural networks. In *International Conference on Learning Representations*, 2018.

Yu, S., Krishnapuram, B., Rosales, R., and Rao, R. B. Active sensing. In *Artificial Intelligence and Statistics*, pp. 639–646. PMLR, 2009.

A. Extended Related Work

This section provides a more extensive discussion of several related areas of work.

A.1. Forecasting treatment effects over time

There is a growing interest in the ML literature in estimating personalized treatment effects over time. To this aim, different types of neural networks have been explored, including RNNs (Lim et al., 2018; Bica et al., 2019; Li et al., 2021; Berrevoets et al., 2021), transformers (Melnychuk et al., 2022), and Neural ODEs (Gwak et al., 2020; De Brouwer et al., 2022) or Neural CDEs (Seedat et al., 2022). All existing work in this area has implicitly relied upon assumptions of the observation process, assuming regular observations or completely random observation intervals, see Table 1. Conversely, our work relies on the less strict SAR assumption.

Reference	Sampling
Lim et al. (2018)	Regular
Bica et al. (2019)	Regular
Berrevoets et al. (2021)	Regular
Li et al. (2021)	Regular
Melnychuk et al. (2022)	Regular
Gwak et al. (2020)	SCAR
Seedat et al. (2022)	SCAR
De Brouwer et al. (2022)	SCAR
This work	SAR

Table 1: **An overview of existing work.** We categorize the related work according to the assumptions made regarding the observation process. Sampling is either assumed to be regular, completely at random (SCAR), or at random (SAR).

In addition to existing work leveraging neural networks, there is another line of work in the ML literature that uses Gaussian processes to estimate treatment effects over time in the presence of irregular samples (Xu et al., 2016; Schulam & Saria, 2017; Roy et al., 2017; Soleimani et al., 2017). Other work has looked at using synthetic controls to estimate the effect of a (single) intervention over time (Bellot & Van Der Schaar, 2021; Qian et al., 2021). Similar to the ML literature on estimating treatment effects over time using neural networks discussed above, these also rely on regular samples or SCAR. To the best of our knowledge, no existing work in the ML literature has studied the problem of estimating treatment effects given SAR, which is the focus of this work.

Outside the ML literature, several approaches have been proposed for estimating *average* treatment effects over time, most notably the seminal work using g-computation and marginal structural models (Robins, 1986; 1997; Robins et al., 2000). More specifically, several approaches have been proposed in the (bio)statistics literature to learn causal effects under informative sampling, see Gasparini et al. (2020) for an overview. Existing work on estimating causal effects under SAR can be categorized based on (1) whether they use inverse visiting weights or random effects, and (2) whether they assume discrete or continuous time. (1) The first group uses the *inverse probability of visiting* or its continuous-time equivalent, the inverse intensity of visiting, as weights in the objective function of the estimator (Robins et al., 1995; Lin et al., 2004; Pullenayegum & Lim, 2016). (2) The second uses *shared random effects* to jointly model the observation and outcome processes (Liu et al., 2008). However, these approaches assume a certain parametric form or latent variable(s), which might not reflect the actual (unknown) data generating procedure and usually focus on population average effects. Conversely, our approach is conceptually similar to inverse intensity of visit weighting, but leverages the use of flexible ML methods that do not require these assumptions.

A.2. Informative sampling in ML

Various other works in ML consider related problem settings. For example, informative sampling has been considered as a source of information for prognosis in health care (Alaa et al., 2017) and as a challenge to robustness of predictive models to distribution shifts (Jeanselme et al., 2022), though this line of work does not consider the estimation of treatment effects. Moreover, the literature on active sensing views takes observing as an active role in which the decision-maker controls the sampling mechanism (Yu et al., 2009). The key question addressed in this line of work is what and when to measure (Yoon et al., 2018) and, potentially, also when to stop measuring (Alaa & Van Der Schaar, 2016). This is in contrast to our setting, where data is not actively sampled, but *observed* over time and the observer has a passive role.

A.3. Neural ODEs

Neural ordinary differential equations (ODEs) have recently emerged as a novel class of machine learning models combining neural networks and differential equations (Chen et al., 2018). Due to their ability of handling irregular observations, Neural ODEs have been applied for time series, either directly or combined with a recurrent neural network (e.g., Rubanova et al., 2019). Neural controlled differential equations (CDEs) additionally allow for modeling covariates as a control, making them suitable for dealing with irregularly sampled time series (Kidger et al., 2020; Morrill et al., 2021), as in the setting considered in this work.

A.4. Missing data

Dealing with missing data is an important and established field in statistics and machine learning (Rubin, 1976; Little & Rubin, 2019). This literature is related to our setting, as we are interested in learning a continuous latent path based on irregular observations over time, which could also be seen as a form of missing data imputation. Moreover, the sampling mechanisms considered in this work are similar to missing data mechanisms. Several recent works explore dealing with missing data in the context of treatment effect estimation (Mayer et al., 2020; Berrevoets et al., 2023).

B. List of Mathematical Symbols

We compile a list of mathematical symbols and their explanation in Table 2. Moreover, we use a real-world example of a health care application to illustrate their meaning.

Symbol	Explanation	Cancer patient example
X	Covariate path	Blood pressure, heart rate, etc.
A	Treatment path	Chemotherapy, radiotherapy, etc.
Y	Outcome path	Tumor size
t	Time	
$N(t)$	Counting process over time	Five observations after two weeks
$\lambda(t)$	Observation intensity over time	Probability of observing tumor size at time t
$\mu_{a,t}(\tau)$	Expected treatment outcome at time $t+\tau$ given treatment path a	Tumor size next week absent any treatment
$\hat{y}_{i,t}(t')$	Instance i's predicted treatment outcome at time $t' = t+\tau$	Patient i's tumor size next week
$\hat{\lambda}_{i,t}(t')$	Instance i's predicted observation intensity at time $t' = t+\tau$	Patient i's observation intensity next week

Table 2: **List of symbols.** We compile a list of the main mathematical symbols used and their explanation. The final column illustrates each symbol for the case of a cancer patient.

C. TESAR-CDE: Multitask Training Procedure

We include a more detailed training procedure for the multitask configuration of TESAR-CDE in Algorithm 1.

Algorithm 1 Pseudo-code for the TESAR-CDE (Multitask) training procedure

Input: Observational data $\mathcal{D} = \{t_j^{(i)}, x_{t_j}^{(i)}, a_{t_j}^{(i)}, y_{t_j}^{(i)}\}$ for $i \in \{0, \ldots, n\}$ and $j \in \{0, \ldots, m_i\}$, weighted MSE loss $\mathcal{L}^{\text{WMSE}}$, cross-entropy loss $\mathcal{L}^{\text{CE}}$, total epochs E, learning rate η, and batch size b. TESAR-CDE architecture consisting of four networks: an embedding network g with weights W_g, an encoder CDE function f_θ with weights W_θ, a decoder CDE function f_ϕ with weights W_ϕ, a final intensity map f_ψ^λ with weights W_ψ^λ, and a final outcome map f_ψ^y with weights W_ψ^y.
for epochs = 1 to E **do**
  Sample batch $i_0, i_1, \ldots, i_b \subset \{0, \ldots, n\}$
  Encode the first observation $z(t_0)^{(i)} = g(t_0^{(i)}, x_{t_0}^{(i)}, a_{t_0}^{(i)}, y_{t_0}^{(i)})$ for each i in batch
  Encode the history up to time t: $z(t) = \texttt{ODESolve}(f_\theta, z(t_0), \bar{X}(t), \bar{A}(t), \bar{Y}(t))$
  Decode the history up to time $t+\tau$: $z_t(t+\tau) = \texttt{ODESolve}(f_\theta, z(t), \bar{A}_t(t+\tau))$
  Map to forecast the outcome at $t+\tau$: $\bar{y}_t(t+\tau) = f_\psi^y(z_t(t+\tau))$
  Map to forecast the intensity at $t+\tau$: $\bar{\lambda}_t(t+\tau) = f_\psi^\lambda(z_t(t+\tau))$
  Compute $\text{grad}_g = \nabla_{W_g} \frac{1}{n} \sum_i^n \mathcal{L}_i^{\text{WMSE}}$
  Compute $\text{grad}_\theta = \nabla_{W_\theta} \frac{1}{n} \sum_i^n \mathcal{L}_i^{\text{WMSE}}$
  Compute $\text{grad}_\phi = \nabla_{W_\phi} \frac{1}{n} \sum_i^n \mathcal{L}_i^{\text{WMSE}}$
  Compute $\text{grad}_{\psi^y} = \nabla_{W_{\psi^y}} \frac{1}{n} \sum_i^n \mathcal{L}_i^{\text{WMSE}}$
  Compute $\text{grad}_{\psi^\lambda} = \nabla_{W_{\psi^\lambda}} \frac{1}{n} \sum_i^n \mathcal{L}_i^{\text{CE}}$
  Update weights $W_g \leftarrow W_g - \eta\ \text{grad}_g$
  Update weights $W_\theta \leftarrow W_\theta - \eta\ \text{grad}_\theta$
  Update weights $W_\phi \leftarrow W_\phi - \eta\ \text{grad}_\phi$
  Update weights $W_\psi^y \leftarrow W_\psi^y - \eta\ \text{grad}_\phi$
  Update weights $W_\psi^\lambda \leftarrow W_\psi^\lambda - \eta\ \text{grad}_\psi^\lambda$
  if If $\frac{1}{n} \sum_i^n \mathcal{L}_i^{\text{MT}} = \frac{1}{n} \sum_i^n \left((1-\alpha)\mathcal{L}_i^{\text{WMSE}} + \alpha \mathcal{L}_i^{\text{CE}}\right)$ did not improve for 50 epochs **then**
    Break {Early stopping}
  end if
end for

D. TESAR-CDE: Implementation

This section provides more details on the implementation of TESAR-CDE.

D.1. Weight truncation

For more stable training, we truncate the estimated intensities at $c_{\text{min}} = 0.001$, such that the maximal importance weight is equal to 1000. This is similar to what is typically done with propensity scores when adjusting for confounding bias. The truncation constant c_{min} allows for trading off bias and variance, with $c_{\text{min}} = 1$ corresponding to the unweighted variant (Cole & Hernán, 2008). We did not tune the cutoff rate c_{min}.

D.2. Hyperparameter optimization

To allow for a fair comparison between the models, we do not tune hyperparameters for each model separately, but rather find the best configuration for the baseline TE-CDE model at a level of informativeness $\gamma = 0$ and use this for all models. We show the ranges and final values for all hyperparameters in Table 3. Hyperparameter optimization was done using wandb's Bayesian optimization (Biewald, 2020). For each network in the CDE (f_θ and f_ϕ), we use a final tanh activation layer, as recommended by Kidger et al. (2020). All models are trained with a batch size of 128 and learning rate of $5e-4$ for a maximum of 1000 epochs. Learning was terminated if the training loss did not improve for 50 epochs. For the multitask configuration, we use $\alpha = 0.8$ to balance the loss terms, though this is only used for early stopping as each part of the network has a different optimizer, see also Algorithm 1. For all models, we construct a control path for the Neural CDEs using a cubic interpolation.

Parameter	**Range**
Latent state z dimension	$\{8, 16, \mathbf{32}\}$
Encoder layers	$\{1, 2, \mathbf{3}\}$
Decoder layers	$\{1, \mathbf{2}, 3\}$
Map layers	$\{\mathbf{1}, 2\}$
Encoder hidden neurons	$\{4, \mathbf{8}, 16\}$
Decoder hidden neurons	$\{4, \mathbf{8}, 16\}$
Map hidden neurons	$\{4, \mathbf{8}, 16\}$

Table 3: **Hyperparameter optimization.** We show the range for each hyperparameter that was optimized. The optimal value is shown in bold.

D.3. A note on adding counts

A frequent practice in time series forecasting when observation times may be informative is to add observation counts to the data (Che et al., 2018; Kidger et al., 2020). However, in the context of estimating treatment effects, this is problematic. First, adding counts is complicated because estimating counterfactual treatments would then require counterfactual count data, which is not observed. Moreover, adding count data may itself introduce confounding or collider bias (Goldstein et al., 2016). Therefore, we do not add observation counts in this work.

E. Tumor Growth Simulation

We use the tumor growth simulation of Geng et al. (2017), which was also used in the previous ML literature on estamating treatment effects over time (e.g., Lim et al., 2018; Bica et al., 2019; Melnychuk et al., 2022; Seedat et al., 2022). We refer to these works for more details.

The tumor size is modelled as:

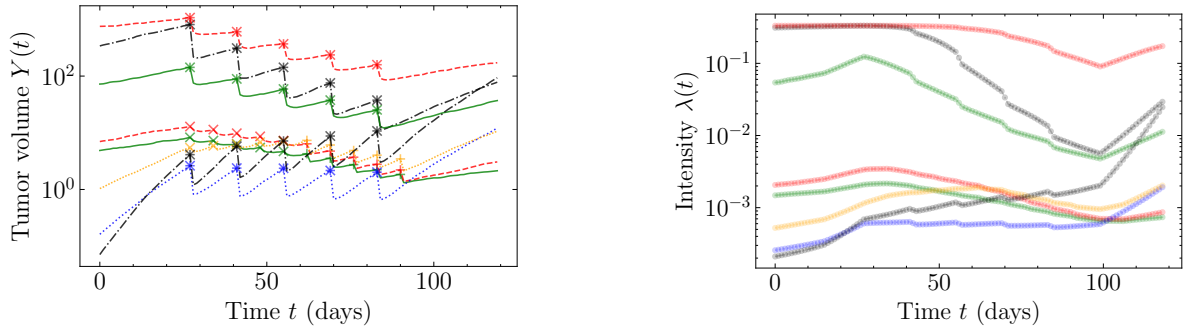
$$\frac{dY(t)}{dt} = \Big[1 + \underbrace{\rho \log\left(\frac{K}{Y(t)}\right)}_{\text{Tumor growth}} - \underbrace{\beta_c C_t}_{\text{Chemotherapy}} - \underbrace{\left(\alpha_r d(t) + \beta_r d(t)^2\right)}_{\text{Radiotherapy}} + \underbrace{\epsilon_t}_{\text{Noise}}\Big] Y(t).$$

Parameters are obtained as follows. Carrying capacity K is set equal to 30. Growth parameter ρ is sampled from a normal distribution $\rho \sim \mathcal{N}(7.00 \times 10^{-5}, 7.23 \times 10^{-3})$. β_c is also sampled from a normal distribution $\beta_c \sim \mathcal{N}(0.028, 0.0007)$. Finally, α_r and β_r are obtained as $\alpha_r \sim \mathcal{N}(0.0398, 0.168)$ and $\beta = \frac{\alpha}{10}$. Noise is added by sampling $\epsilon(t) \sim \mathcal{N}(0, 0.01)$.

Following earlier work (Lim et al., 2018; Bica et al., 2019; Seedat et al., 2022), we create heterogeniety in the treatment effects by creating three patient groups. Patient group 1 has a larger radiotherapy effect, achieved by multiplying $\mu(\alpha_r)$ with 1.1. Similarly, patient group 3 has a larger chemotherapy effect by increasing α_c with 10%.

We consider two types of treatment plans: a sequential and a concurrent plan (Furuse et al., 1999; Aupérin et al., 2010; Curran Jr et al., 2011). We show simulated tumor paths and intensities for several patients in Fig. 7. For the informative observation process Eq. (7), we visualize the intensity distribution at different levels of γ in Figs. 8 and 9. For the uninformative observation process, we show the intensity distributions in Fig. 10.

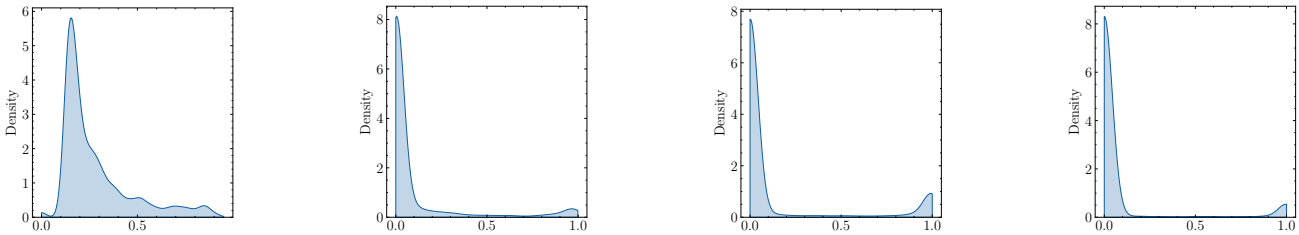
For all experiments, we generate patient trajectories over 120 days. For training, we split the data for forecasting to have a lookback window of seven days and a maximum forecasting horizon of five days.



(a) Tumor size $Y(t)$ over time

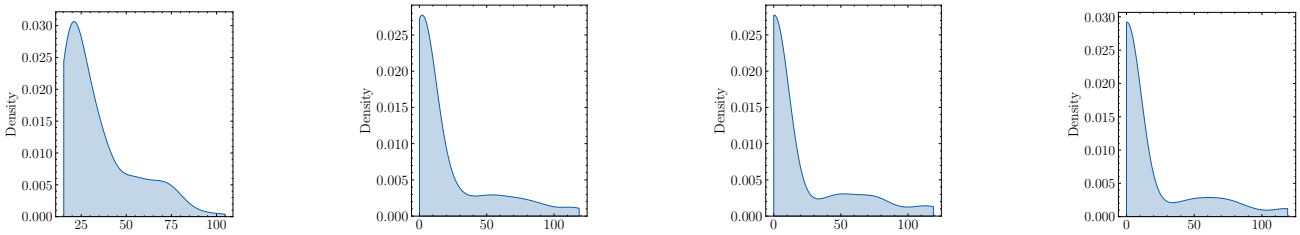
(b) Intensity $\lambda(t)$ over time

Figure 7: **Tumor and intensity evolution.** We show the simulated tumor size and corresponding intensity over time for several (randomly selected) patients. The intensity is simulated is based on Eq. (7) with an informativeness $\gamma = 4$.



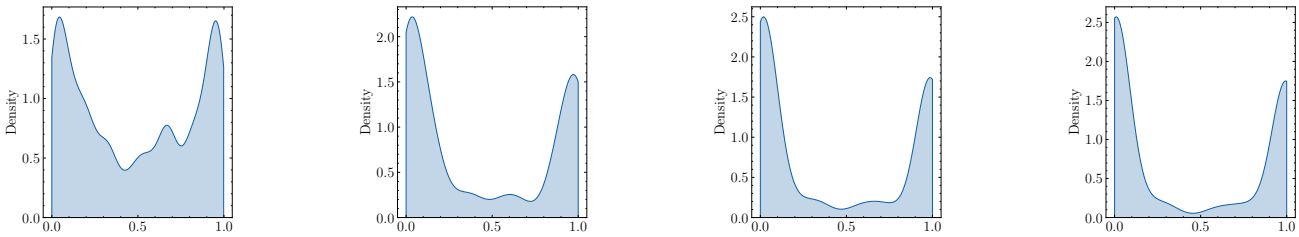
(a) $\gamma = 2$ (b) $\gamma = 4$ (c) $\gamma = 6$ (d) $\gamma = 8$

Figure 8: **Informative sampling – intensity distribution.** We show the distribution of intensities $\lambda(t)$ over all patients for different levels of informativeness γ. At $\gamma = 0$ (not shown), all intensities are equal $\lambda_i(t) = 0.5$.



(a) $\gamma = 2$ (b) $\gamma = 4$ (c) $\gamma = 6$ (d) $\gamma = 8$

Figure 9: **Informative sampling – expected observations.** We show the expected observations over the entire time period considered over all patients for different levels of informativeness γ. At $\gamma = 0$ (not shown), all intensities are equal $\lambda_i(t) = 0.5$ and all patients have 60 expected observations.



(a) $\gamma = 2$ (b) $\gamma = 4$ (c) $\gamma = 6$ (d) $\gamma = 8$

Figure 10: **Uninformative sampling – intensity distribution.** We show the distribution of intensities $\lambda(t)$ over all patients for different levels of "informativeness" γ. At $\gamma = 0$ (not shown), all intensities are equal $\lambda_i(t) = 0.5$.

F. Additional Results

In this section, we present additional results to further validate the proposed TESAR-CDE. First, we evaluate the predicted observation intensities. Second, we analyze the sensitivity of the multitask model to hyperparameter α.

We evaluate how accurate TESAR-CDE predicts the observation intensities in terms of the Brier Score: $\text{BS} = \sum_{t=0}^{T} \sum_{\tau=0}^{\tau_{\max}} \left(\lambda_i(t+\tau) - \hat{\lambda}_{i,t}(t+\tau) \right)^2$, see Fig. 11. Generally, both versions of our method can learn to accurately predict the observation intensities, with the two-step TESAR-CDE performing slightly better than the multitask configuration. These findings are consistent with our motivation of the multitask setup: while the two-step model learns a (generally better) model of the intensity itself using all available information, these more accurate intensities do not help with potential outcome prediction. Additionally, we find that the Brier score increases with informativeness for both models. This trend indicates that more informativeness makes it harder to learn to predict the observation intensity. We hypothesize that this is due to observing in general becoming more rare as increases.

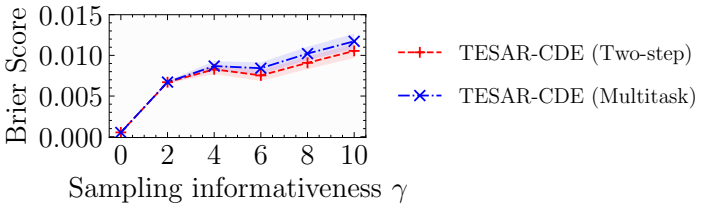


Figure 11: **Evaluating the intensity prediction at varying informativeness γ.** We show the Brier Score $\pm$ SE (lower is better) over ten runs at increasing levels of informativeness γ, keeping the forecasting horizon $\tau = 1$.

Next, we evaluate performance for the multitask TESAR-CDE for different values of its hyperparameter α, see Fig. 12. As the shared representation is only trained for outcome prediction, the only point of having the hyperparameter is to scale both terms such that they roughly influence the early stopping in the same way. We see that our method is robust to different values of this hyperparameter and that scaling them to approximately the same magnitude (using 0.8) results in good performance in practice. (Note that the loss terms are weighted by α and $(1-\alpha)$, which is why we restrict to be strictly between 0 and 1.)

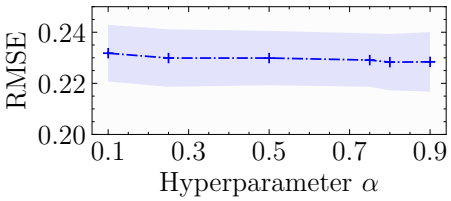


Figure 12: **Evaluating the multitask configuration's outcome prediction for different levels of hyperparameter α.** We show the RMSE $\pm$ SE over ten runs, while keeping the level of informativeness fixed at $\gamma = 6$ and averaging over $\tau \in \{1, \ldots, 5\}$.